IBM SPSS Modeler 18.2.2 Source, Process, and Output Nodes

Note

Before you use this information and the product it supports, read the information in “Notices” on page 357.

Product Information

This edition applies to version 18, release 2, modification 2 of IBM® SPSS® Modeler and to all subsequent releases and modifications until otherwise indicated in new editions.

© Copyright International Business Machines Corporation .
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Preface ..... **xi**

Chapter 1. About IBM SPSS Modeler **1**
IBM SPSS Modeler Products 1
IBM SPSS Modeler 1
IBM SPSS Modeler Server 1
IBM SPSS Modeler Administration Console 2
IBM SPSS Modeler Batch 2
IBM SPSS Modeler Solution Publisher 2
IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services 2
IBM SPSS Modeler Editions 2
Documentation 3
SPSS Modeler Professional Documentation 3
SPSS Modeler Premium Documentation 4
Application examples 4
Demos Folder 4
License tracking 4

Chapter 2. Source Nodes **5**
Overview 5
Setting Field Storage and Formatting 6
List storage and associated measurement levels 8
Unsupported control characters 9
Analytic Server Source node 9
Selecting a data source 10
Amending credentials 10
Supported nodes 10
Database source node 14
Setting Database Node Options 15
Adding a database connection 15
Potential database issues 17
Specifying preset values for a database connection 18
Selecting a Database Table 20
Querying the database 21
Using a custom database configuration file 22
Variable File Node 24
Setting options for the Variable File Node 24
Importing geospatial data into the Variable File Node 26
Fixed File Node 27
Setting options for the Fixed File node 27
Statistics File Node 28
Data Collection Node 29
Data Collection Import File Options 30
Data Collection Import Metadata Properties 32
Database Connection String 32
Advanced Properties 32
Importing Multiple Response Sets 32
Data Collection Column Import Notes 33
IBM Cognos source node 33
Cognos object icons 34

Importing Cognos data....34
Importing Cognos reports....35
Cognos connections....36
Cognos location selection....36
Specifying parameters for data or reports....36
IBM Cognos TM1 Source Node....37
Importing IBM Cognos TM1 data....37
TWC source node....38
SAS Source Node....39
Setting Options for the SAS Source Node....39
Excel Source Node....39
XML Source Node....40
Selecting from Multiple Root Elements....41
Removing Unwanted Spaces from XML Source Data....41
User Input Node....42
Setting Options for the User Input Node....42
Simulation Generate Node....46
Setting Options for the Simulation Generate Node....47
Clone Field....52
Fit Details....52
Specify Parameters....53
Distributions....55
Extension Import node....58
Extension Import node - Syntax tab....58
Extension Import node - Console Output tab....58
Filtering or renaming fields....58
Geospatial Source Node....59
Setting Options for the Geospatial Source Node....59
JSON Source node....59
Common Source Node Tabs....60
Setting Measurement Levels in the Source Node....60
Filtering Fields from the Source Node....61

Chapter 3. Record Operations Nodes....63
Overview of Record Operations....63
Select Node....64
Sample Node....65
Sample node options....66
Cluster and Stratify Settings....67
Sample Sizes for Strata....68
Balance Node....69
Setting Options for the Balance Node....69
Aggregate Node....70
Setting options for the Aggregate node....71
Aggregate optimization settings....72
RFM Aggregate Node....73
Setting Options for the RFM Aggregate Node....73
Sort Node....74
Sort Optimization Settings....74
Merge Node....74
Types of Joins....75
Specifying a Merge Method and Keys....76
Selecting Data for Partial Joins....77
Specifying Conditions for a Merge....78
Specifying Ranked Conditions for a Merge....78
Filtering Fields from the Merge Node....79
Setting Input Order and Tagging....80

Merge Optimization Settings....80
Append Node....81
Setting Append Options....81
Distinct node....82
Distinct Optimization Settings....84
Distinct Composite Settings....84
Streaming Time Series node....85
Streaming Time Series node - field options....86
Streaming Time Series node - data specification options....86
Streaming Time Series node - build options....90
Streaming Time Series node - model options....93
SMOTE node....95
SMOTE node Settings....95
Extension Transform node....96
Extension Transform node - Syntax tab....96
Extension Transform node - Console Output tab....97
Space-Time-Boxes Node....97
Defining Space-Time-Box density....99
Streaming TCM Node....99
Streaming TCM Node - Time Series options....99
Streaming TCM Node - Observations options....100
Streaming TCM Node - Time Interval options....101
Streaming TCM Node - Aggregation and Distribution options....102
Streaming TCM Node - Missing Value options....102
Streaming TCM Node - General Data options....103
Streaming TCM Node - General Build options....103
Streaming TCM Node - Estimation Period options....103
Streaming TCM Node - Model options....104
CPLEX Optimization node....104
Setting options for the CPLEX Optimization node....105

Chapter 4. Field Operations Nodes....107
Field Operations Overview....107
Automated Data Preparation....109
Fields tab....110
Settings tab....111
Analysis tab....115
Generating a Derive node....121
Type Node....121
Measurement levels....122
Converting Continuous Data....126
What is instantiation?....126
Data Values....127
Defining Missing Values....130
Checking Type Values....131
Setting the field role....131
Copying Type Attributes....132
Field Format Settings Tab....132
Filtering or renaming fields....133
Setting filtering options....134
Derive node....136
Setting Basic Options for the Derive Node....137
Deriving Multiple Fields....137
Setting Derive Formula Options....138
Setting Derive Flag Options....140
Setting Derive Nominal Options....141
Setting Derive State Options....141

Setting Derive Count Options....141
Setting Derive Conditional Options....141
Recoding Values with the Derive Node....142
Filler node....142
Storage Conversion Using the Filler Node....143
Reclassify Node....143
Setting Options for the Reclassify Node....144
Reclassifying Multiple Fields....145
Storage and Measurement Level for Reclassified Fields....145
Anonymize Node....145
Setting Options for the Anonymize Node....146
Anonymizing Field Values....147
Binning Node....147
Setting Options for the Binning Node....148
Fixed-Width Bins....148
Tiles (Equal Count or Sum)....149
Rank Cases....150
Mean/Standard Deviation....151
Optimal Binning....151
Previewing the Generated Bins....152
RFM Analysis Node....152
RFM Analysis Node Settings....153
RFM Analysis Node Binning....154
Ensemble Node....154
Ensemble Node Settings....154
Partition Node....155
Partition Node Options....156
Set to Flag Node....157
Setting Options for the Set to Flag Node....157
Restructure Node....157
Setting Options for the Restructure Node....159
Transpose Node....159
Setting options for the Transpose node....159
History Node....161
Setting Options for the History Node....161
Field Reorder Node....161
Setting Field Reorder Options....162
Time Intervals Node....163
Time Interval - field options....163
Time Interval - build options....163
Reprojection Node....164
Setting options for the Reproject node....165

Chapter 5. Graph Nodes....167
Common Graph Nodes Features....167
Aesthetics, Overlays, Panels, and Animation....168
Using the Output Tab....169
Using the Annotations Tab....170
3-D Graphs....170
Graphboard node....171
Graphboard Basic Tab....171
Graphboard Detailed Tab....174
Available Built-in Graphboard Visualization Types....176
Creating Map Visualizations....183
Graphboard Examples....184
Graphboard appearance tab....194
Setting the Location of Templates, Stylesheets, and Maps....195

Managing Templates, Stylesheets, and Map Files ... 195
Converting and Distributing Map Shapefiles ... 196
Key Concepts for Maps ... 197
Using the Map Conversion Utility ... 197
Distributing map files ... 202
Plot Node ... 202
Plot Node Tab ... 205
Plot Options Tab ... 206
Plot Appearance Tab ... 207
Using a Plot Graph ... 207
Multiplot Node ... 208
Multiplot Plot Tab ... 208
Multiplot Appearance Tab ... 210
Using a Multiplot Graph ... 210
Time Plot Node ... 211
Time Plot Tab ... 211
Time Plot Appearance Tab ... 212
Using a Time Plot Graph ... 212
Distribution Node ... 213
Distribution Plot Tab ... 213
Distribution Appearance Tab ... 213
Using a Distribution Node ... 213
Histogram Node ... 215
Histogram Plot Tab ... 215
Histogram Options Tab ... 215
Histogram Appearance Tab ... 216
Using Histograms ... 216
Collection Node ... 217
Collection Plot Tab ... 217
Collection Options Tab ... 218
Collection Appearance Tab ... 218
Using a Collection Graph ... 219
Web Node ... 219
Web Plot Tab ... 220
Web Options Tab ... 221
Web Appearance Tab ... 222
Using a Web Graph ... 223
Evaluation node ... 226
Evaluation Plot Tab ... 230
Evaluation Options tab ... 231
Evaluation Appearance Tab ... 232
Reading the Results of a Model Evaluation ... 232
Using an Evaluation Chart ... 233
Map Visualization Node ... 234
Map Visualization Plot Tab ... 234
Map Visualization Appearance Tab ... 237
t-SNE node ... 238
t-SNE node Expert options ... 238
t-SNE node Output options ... 240
Accessing and plotting t-SNE data ... 240
t-SNE model nuggets ... 241
E-Plot (Beta) node ... 242
E-Plot (Beta) node Plot tab ... 242
E-Plot (Beta) node Options tab ... 242
E-Plot (Beta) Appearance tab ... 242
Using an e-plot graph ... 243
Exploring Graphs ... 245
Using Bands ... 246

Using Regions....249
Using Marked Elements....251
Generating Nodes from Graphs....252
Editing Visualizations....254
General Rules for Editing Visualizations....255
Editing and Formatting Text....256
Changing Colors, Patterns, Dashings, and Transparency....256
Rotating and Changing the Shape and Aspect Ratio of Point Elements....257
Changing the size of graphic elements....258
Specifying Margins and Padding....258
Formatting Numbers....258
Changing the Axis and Scale Settings....259
Editing Categories....260
Changing the Orientation Panels....261
Transforming the Coordinate System....261
Changing Statistics and Graphic Elements....262
Changing the Position of the Legend....263
Copying a Visualization and Visualization Data....263
Graphboard Editor Keyboard Shortcuts....264
Adding Titles and Footnotes....264
Using Graph Stylesheets....265
Printing, saving, copying, and exporting graphs....266

Chapter 6. Output Nodes....271
Overview of Output Nodes....271
Managing output....272
Viewing output....272
Publish to Web....273
Viewing Output in an HTML Browser....274
Exporting Output....274
Selecting cells and columns....275
Table node....275
Table Node Settings Tab....275
Table Node Format Tab....275
Output Node Output Tab....275
Table Browser....277
Matrix node....277
Matrix Node Settings Tab....277
Matrix Node Appearance Tab....278
Matrix node output browser....279
Analysis Node....280
Analysis Node Analysis Tab....280
Analysis Output Browser....281
Data Audit node....282
Data Audit Node Settings Tab....283
Data Audit Quality Tab....284
Data Audit Output Browser....284
Transform Node....289
Transform Node Options Tab....289
Transform Node Output Tab....289
Transform Node Output Viewer....290
Statistics Node....291
Statistics Node Settings Tab....291
Statistics Output Browser....292
Means Node....293
Comparing Means for Independent Groups....294
Comparing Means Between Paired Fields....294

Means Node Options....294
Means Node Output Browser....294
Report Node....295
Report Node Template Tab....296
Report Node Output Browser....297
Set Globals Node....297
Set Globals Node Settings Tab....297
Simulation Fitting Node....298
Distribution Fitting....298
Simulation Fitting Node Settings Tab....300
Simulation Evaluation Node....300
Simulation Evaluation Node Settings Tab....301
Simulation Evaluation node output....302
Extension Output node....307
Extension Output node - Syntax tab....307
Extension Output node - Console Output tab....308
Extension Output node - Output tab....308
Extension Output Browser....308
KDE nodes....309
KDE Modeling node and KDE Simulation node Fields....309
KDE nodes Build Options....309
KDE Modeling node and KDE Simulation node Model Options....311
IBM SPSS Statistics Helper Applications....311

Chapter 7. Export Nodes....313
Overview of Export Nodes....313
Database Export Node....314
Database Node Export Tab....314
Database Export Merge Options....315
Database export schema options....316
Database Export Index Options....318
Database export advanced options....319
Bulk loader programming....320
Flat File Export Node....326
Flat File Export Tab....326
Statistics Export Node....327
Statistics Export Node - Export Tab....327
Renaming or Filtering Fields for IBM SPSS Statistics....328
Data Collection Export Node....328
Analytic Server Export node....329
IBM Cognos export node....329
Cognos connection....330
ODBC connection....330
IBM Cognos TM1 Export Node....331
Connecting to an IBM Cognos TM1 cube to export data....332
Mapping IBM Cognos TM1 data for export....332
SAS Export Node....333
SAS Export Node Export Tab....333
Excel Export Node....333
Excel Node Export Tab....333
Extension Export node....334
Extension Export node - Syntax tab....334
Extension Export node - Console Output tab....334
XML Export Node....335
Writing XML Data....335
XML Mapping Records Options....336
XML Mapping Fields Options....336

XML Mapping Preview ... 336
JSON Export node ... 336
Common Export node tabs ... 337
Publishing streams ... 337

Chapter 8. IBM SPSS Statistics Nodes ... 339
IBM SPSS Statistics Nodes - Overview ... 339
Statistics File Node ... 340
Statistics Transform Node ... 341
Statistics Transform Node - Syntax Tab ... 341
Allowable Syntax ... 341
Statistics Model Node ... 343
Statistics Model Node - Model Tab ... 344
Statistics Model Node - Model Nugget Summary ... 344
Statistics Output Node ... 344
Statistics Output Node - Syntax Tab ... 344
Statistics Output Node - Output Tab ... 346
Statistics Export Node ... 346
Statistics Export Node - Export Tab ... 347
Renaming or Filtering Fields for IBM SPSS Statistics ... 347

Chapter 9. SuperNodes ... 349
Overview of SuperNodes ... 349
Types of SuperNodes ... 349
Source SuperNodes ... 349
Process SuperNodes ... 349
Terminal SuperNodes ... 350
Creating SuperNodes ... 350
Nesting SuperNodes ... 351
Locking SuperNodes ... 351
Locking and unlocking a SuperNode ... 351
Editing a locked SuperNode ... 351
Editing SuperNodes ... 352
Modifying SuperNode types ... 352
Annotating and renaming SuperNodes ... 352
SuperNode parameters ... 353
SuperNodes and caching ... 354
SuperNodes and scripting ... 355
Saving and loading SuperNodes ... 355

Notices ... 357
Trademarks ... 358
Terms and conditions for product documentation ... 358
C ... 361
K ... 361
M ... 361
R ... 361
S ... 362
U ... 362
V ... 363

Index ... 365

Preface

IBM SPSS Modeler is the IBM enterprise-strength data mining workbench. SPSS Modeler helps organizations to improve customer and citizen relationships through an in-depth understanding of data. Organizations use the insight gained from SPSS Modeler to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery.

SPSS Modeler's visual interface invites users to apply their specific business expertise, which leads to more powerful predictive models and shortens time-to-solution. SPSS Modeler offers many modeling techniques, such as prediction, classification, segmentation, and association detection algorithms. Once models are created, IBM SPSS Modeler Solution Publisher enables their delivery enterprise-wide to decision makers or to a database.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises - able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative, visit http://www.ibm.com/spss.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM web site at http://www.ibm.com/support. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Chapter 1. About IBM SPSS Modeler

IBM SPSS Modeler is a set of data mining tools that enable you to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making. Designed around the industry-standard CRISP-DM model, IBM SPSS Modeler supports the entire data mining process, from data to better business results.

IBM SPSS Modeler offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics. The methods available on the Modeling palette allow you to derive new information from your data and to develop predictive models. Each method has certain strengths and is best suited for particular types of problems.

SPSS Modeler can be purchased as a standalone product, or used as a client in combination with SPSS Modeler Server. A number of additional options are also available, as summarized in the following sections. For more information, see https://www.ibm.com/analytics/us/en/technology/spss/.

IBM SPSS Modeler Products

The IBM SPSS Modeler family of products and associated software comprises the following.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (included with IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adapters for IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler is a functionally complete version of the product that you install and run on your personal computer. You can run SPSS Modeler in local mode as a standalone product, or use it in distributed mode along with IBM SPSS Modeler Server for improved performance on large data sets.

With SPSS Modeler, you can build accurate predictive models quickly and intuitively, without programming. Using the unique visual interface, you can easily visualize the data mining process. With the support of the advanced analytics embedded in the product, you can discover previously hidden patterns and trends in your data. You can model outcomes and understand the factors that influence them, enabling you to take advantage of business opportunities and mitigate risks.

SPSS Modeler is available in two editions: SPSS Modeler Professional and SPSS Modeler Premium. See the topic “ IBM SPSS Modeler Editions” on page 2 for more information.

IBM SPSS Modeler Server

SPSS Modeler uses a client/server architecture to distribute requests for resource-intensive operations to powerful server software, resulting in faster performance on larger data sets.

SPSS Modeler Server is a separately-licensed product that runs continually in distributed analysis mode on a server host in conjunction with one or more IBM SPSS Modeler installations. In this way, SPSS Modeler Server provides superior performance on large data sets because memory-intensive operations can be done on the server without downloading data to the client computer. IBM SPSS Modeler Server also provides support for SQL optimization and in-database modeling capabilities, delivering further benefits in performance and automation.

IBM SPSS Modeler Administration Console

The Modeler Administration Console is a graphical user interface for managing many of the SPSS Modeler Server configuration options, which are also configurable by means of an options file. The console is included in IBM SPSS Deployment Manager, can be used to monitor and configure your SPSS Modeler Server installations, and is available free-of-charge to current SPSS Modeler Server customers. The application can be installed only on Windows computers; however, it can administer a server installed on any supported platform.

IBM SPSS Modeler Batch

While data mining is usually an interactive process, it is also possible to run SPSS Modeler from a command line, without the need for the graphical user interface. For example, you might have long-running or repetitive tasks that you want to perform with no user intervention. SPSS Modeler Batch is a special version of the product that provides support for the complete analytical capabilities of SPSS Modeler without access to the regular user interface. SPSS Modeler Server is required to use SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher is a tool that enables you to create a packaged version of an SPSS Modeler stream that can be run by an external runtime engine or embedded in an external application. In this way, you can publish and deploy complete SPSS Modeler streams for use in environments that do not have SPSS Modeler installed. SPSS Modeler Solution Publisher is distributed as part of the IBM SPSS Collaboration and Deployment Services - Scoring service, for which a separate license is required. With this license, you receive SPSS Modeler Solution Publisher Runtime, which enables you to execute the published streams.

For more information about SPSS Modeler Solution Publisher, see the IBM SPSS Collaboration and Deployment Services documentation. The IBM SPSS Collaboration and Deployment Services Knowledge Center contains sections called "IBM SPSS Modeler Solution Publisher" and "IBM SPSS Analytics Toolkit."

IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services

A number of adapters for IBM SPSS Collaboration and Deployment Services are available that enable SPSS Modeler and SPSS Modeler Server to interact with an IBM SPSS Collaboration and Deployment Services repository. In this way, an SPSS Modeler stream deployed to the repository can be shared by multiple users, or accessed from the thin-client application IBM SPSS Modeler Advantage. You install the adapter on the system that hosts the repository.

IBM SPSS Modeler Editions

SPSS Modeler is available in the following editions.

SPSS Modeler Professional

SPSS Modeler Professional provides all the tools you need to work with most types of structured data, such as behaviors and interactions tracked in CRM systems, demographics, purchasing behavior and sales data.

SPSS Modeler Premium

SPSS Modeler Premium is a separately-licensed product that extends SPSS Modeler Professional to work with specialized data and with unstructured text data. SPSS Modeler Premium includes IBM SPSS Modeler Text Analytics:

IBM SPSS Modeler Text Analytics uses advanced linguistic technologies and Natural Language Processing (NLP) to rapidly process a large variety of unstructured text data, extract and organize the key concepts, and group these concepts into categories. Extracted concepts and categories can be combined with existing structured data, such as demographics, and applied to modeling using the full suite of IBM SPSS Modeler data mining tools to yield better and more focused decisions.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription provides all the same predictive analytics capabilities as the traditional IBM SPSS Modeler client. With the Subscription edition, you can download product updates regularly.

Documentation

Documentation is available from the Help menu in SPSS Modeler. This opens the online Knowledge Center, which is always available outside the product.

Complete documentation for each product (including installation instructions) is also available in PDF format, in a separate compressed folder, as part of the product download. Or the latest PDF documents can be downloaded from the web at https://www.ibm.com/support/pages/spss-modeler-1822-documentation.

SPSS Modeler Professional Documentation

The SPSS Modeler Professional documentation suite (excluding installation instructions) is as follows.

- **IBM SPSS Modeler User's Guide.** General introduction to using SPSS Modeler, including how to build data streams, handle missing values, build CLEM expressions, work with projects and reports, and package streams for deployment to IBM SPSS Collaboration and Deployment Services or IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler Source, Process, and Output Nodes.** Descriptions of all the nodes used to read, process, and output data in different formats. Effectively this means all nodes other than modeling nodes.
- **IBM SPSS Modeler Modeling Nodes.** Descriptions of all the nodes used to create data mining models. IBM SPSS Modeler offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics.
- **IBM SPSS Modeler Applications Guide.** The examples in this guide provide brief, targeted introductions to specific modeling methods and techniques. An online version of this guide is also available from the Help menu. See the topic "Application examples" on page 4 for more information.
- **IBM SPSS Modeler Python Scripting and Automation.** Information on automating the system through Python scripting, including the properties that can be used to manipulate nodes and streams.
- **IBM SPSS Modeler Deployment Guide.** Information on running IBM SPSS Modeler streams as steps in processing jobs under IBM SPSS Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** CLEF provides the ability to integrate third-party programs such as data processing routines or modeling algorithms as nodes in IBM SPSS Modeler.
- **IBM SPSS Modeler In-Database Mining Guide.** Information on how to use the power of your database to improve performance and extend the range of analytical capabilities through third-party algorithms.
- **IBM SPSS Modeler Server Administration and Performance Guide.** Information on how to configure and administer IBM SPSS Modeler Server.
- **IBM SPSS Deployment Manager User Guide.** Information on using the administration console user interface included in the Deployment Manager application for monitoring and configuring IBM SPSS Modeler Server.
- **IBM SPSS Modeler CRISP-DM Guide.** Step-by-step guide to using the CRISP-DM methodology for data mining with SPSS Modeler.

- **IBM SPSS Modeler Batch User's Guide.** Complete guide to using IBM SPSS Modeler in batch mode, including details of batch mode execution and command-line arguments. This guide is available in PDF format only.

SPSS Modeler Premium Documentation

The SPSS Modeler Premium documentation suite (excluding installation instructions) is as follows.

- **SPSS Modeler Text Analytics User's Guide.** Information on using text analytics with SPSS Modeler, covering the text mining nodes, interactive workbench, templates, and other resources.

Application examples

While the data mining tools in SPSS Modeler can help solve a wide variety of business and organizational problems, the application examples provide brief, targeted introductions to specific modeling methods and techniques. The data sets used here are much smaller than the enormous data stores managed by some data miners, but the concepts and methods that are involved are scalable to real-world applications.

To access the examples, click **Application Examples** on the Help menu in SPSS Modeler.

The data files and sample streams are installed in the `Demos` folder under the product installation directory. For more information, see "Demos Folder" on page 4.

Database modeling examples. See the examples in the *IBM SPSS Modeler In-Database Mining Guide*.

Scripting examples. See the examples in the *IBM SPSS Modeler Scripting and Automation Guide*.

Demos Folder

The data files and sample streams that are used with the application examples are installed in the `Demos` folder under the product installation directory (for example: `C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`). This folder can also be accessed from the IBM SPSS Modeler program group on the Windows Start menu, or by clicking `Demos` on the list of recent directories in the **File** > **Open Stream** dialog box.

License tracking

When you use SPSS Modeler, license usage is tracked and logged at regular intervals. The license metrics that are logged are *AUTHORIZED_USER* and *CONCURRENT_USER*, and the type of metric that is logged depends on the type of license that you have for SPSS Modeler.

The log files that are produced can be processed by the IBM License Metric Tool, from which you can generate license usage reports.

The license log files are created in the same directory where SPSS Modeler Client log files are recorded (by default, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Chapter 2. Source Nodes

Overview

Source nodes enable you to import data stored in a number of formats, such as flat files, IBM SPSS Statistics (.sav), SAS, Microsoft Excel, and ODBC-compliant relational databases. You can also generate synthetic data using the User Input node.

The Sources palette contains the following nodes:

The Analytic Server source enables you to run a stream on Hadoop Distributed File System (HDFS). The information in an Analytic Server data source can come from a variety of places, such as text files and databases. See the topic "Analytic Server Source node" on page 9 for more information.

The Database node can be used to import data from a variety of other packages using ODBC (Open Database Connectivity), including Microsoft SQL Server, Db2, Oracle, and others. See the topic "Database source node" on page 14 for more information.

The Variable File node reads data from free-field text files—that is, files whose records contain a constant number of fields but a varied number of characters. This node is also useful for files with fixed-length header text and certain types of annotations. See the topic "Variable File Node" on page 24 for more information.

The Fixed File node imports data from fixed-field text files—that is, files whose fields are not delimited but start at the same position and are of a fixed length. Machine-generated or legacy data are frequently stored in fixed-field format. See the topic "Fixed File Node" on page 27 for more information.

The Statistics File node reads data from the *.sav* or *.zsav* file format used by IBM SPSS Statistics, as well as cache files saved in IBM SPSS Modeler, which also use the same format.

The Data Collection node imports survey data from various formats used by market research software conforming to the Data Collection Data Model. A Data Collection Developer Library must be installed to use this node. See the topic " Data Collection Node" on page 29 for more information.

The IBM Cognos source node imports data from Cognos Analytics databases.

The IBM Cognos TM1 source node imports data from Cognos TM1 databases.

The SAS File node imports SAS data into IBM SPSS Modeler. See the topic "SAS Source Node" on page 39 for more information.

The Excel node imports data from Microsoft Excel in the .xlsx file format. An ODBC data source is not required. See the topic "Excel Source Node" on page 39 for more information.

The XML source node imports data in XML format into the stream. You can import a single file, or all files in a directory. You can optionally specify a schema file from which to read the XML structure.

The User Input node provides an easy way to create synthetic data—either from scratch or by altering existing data. This is useful, for example, when you want to create a test dataset for modeling. See the topic "User Input Node" on page 42 for more information.

The Simulation Generate node provides an easy way to generate simulated data—either from scratch using user specified statistical distributions or automatically using the distributions obtained from running a Simulation Fitting node on existing historical data. This is useful when you want to evaluate the outcome of a predictive model in the presence of uncertainty in the model inputs.

Use the Geospatial source node to bring map or spatial data into your data mining session. See the topic "Geospatial Source Node" on page 59 for more information.

The JSON source node imports data from a JSON file.

To begin a stream, add a source node to the stream canvas. Next, double-click the node to open its dialog box. The various tabs in the dialog box allow you to read in data; view the fields and values; and set a variety of options, including filters, data types, field role, and missing-value checking.

Setting Field Storage and Formatting

Options on the Data tab for Fixed File, Variable File, XML Source, and User Input nodes allow you to specify the storage type for fields as they are imported or created in IBM SPSS Modeler. For Fixed File, Variable File and User Input nodes you can also specify the field formatting, and other metadata.

For data read from other sources, storage is determined automatically but can be changed using a conversion function, such as `to_integer`, in a Filler node or Derive node.

Field Use the **Field** column to view and select fields in the current dataset.

Override Select the check box in the **Override** column to activate options in the **Storage** and **Input Format** columns.

Data Storage

Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*. While storage must be determined at the source, measurement level can be changed using a Type node at any point in the stream. See the topic "Measurement levels" on page 122 for more information.

Available storage types are:

- **String** Used for fields that contain non-numeric data, also called alphanumeric data. A string can include any sequence of characters, such as *fred*, *Class 2*, or *1234*. Note that numbers in strings cannot be used in calculations.
- **Integer** A field whose values are integers.
- **Real** Values are numbers that may include decimals (not limited to integers). The display format is specified in the Stream Properties dialog box and can be overridden for individual fields in a Type node (Format tab).
- **Date** Date values specified in a standard format such as year, month, and day (for example, `2007-09-26`). The specific format is specified in the Stream Properties dialog box.
- **Time** Time measured as a duration. For example, a service call lasting 1 hour, 26 minutes, and 38 seconds might be represented as `01:26:38`, depending on the current time format as specified in the Stream Properties dialog box.
- **Timestamp** Values that include both a date and time component, for example `2007–09–26 09:04:00`, again depending on the current date and time formats in the Stream Properties dialog box. Note that timestamp values may need to be wrapped in double-quotes to ensure they are interpreted as a single value rather than separate date and time values. (This applies for example when entering values in a User Input node.)
- **List** Introduced in SPSS Modeler version 17, along with new measurement levels of Geospatial and Collection, a List storage field contains multiple values for a single record. There are list versions of all of the other storage types.

Table 1. List storage type icons

Icon	Storage type
[A]	List of string
[◇]	List of integer
[◈]	List of real
[🕒]	List of time
[▦]	List of date
[▣]	List of timestamp
[[]]	List with a depth greater than zero

In addition, for use with the Collection measurement level, there are list versions of the following measurement levels.

Table 2. List measurement level icons

Icon	Measurement level
[✎]	List of continuous
[●]	List of categorical
[●]	List of flags
[●]	List of nominal
[▮]	List of ordinal

Lists can be imported into SPSS Modeler in one of three source nodes (Analytic Server, Geospatial, or Variable File), or created within your streams through use of the Derive or Filler field operation nodes.

For more information on Lists and their interaction with the Collection and Geospatial measurement levels, see “List storage and associated measurement levels” on page 8

Storage conversions. You can convert storage for a field using a variety of conversion functions, such as `to_string` and `to_integer`, in a Filler node. See the topic "Storage Conversion Using the Filler Node" on page 143 for more information. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the Stream Properties dialog box. For example, if you want to convert a string field with values *Jan 2018*, *Feb 2018*, (and so forth) to date storage, select **MON YYYY** as the default date format for the stream. Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node to perform other manipulations, such as recoding string fields with categorical values. See the topic "Recoding Values with the Derive Node" on page 142 for more information.

Reading in mixed data. Note that when reading in fields with numeric storage (either integer, real, time, timestamp, or date), any non-numeric values are set to null or system missing. This is because unlike some applications, IBM SPSS Modeler does not allow mixed storage types within a field. To avoid this, any fields with mixed data should be read in as strings, either by changing the storage type in the source node or in the external application as necessary.

Field Input Format (Fixed File, Variable File, and User Input nodes only)

For all storage types except String and Integer, you can specify formatting options for the selected field using the drop-down list. For example, when merging data from various locales, you may need to specify a period (.) as the decimal separator for one field, while another will require a comma separator.

Input options specified in the source node override the formatting options specified in the stream properties dialog box; however, they do not persist later in the stream. They are intended to parse input correctly based on your knowledge of the data. The specified formats are used as a guide for parsing the data as they are read into IBM SPSS Modeler, not to determine how they should be formatted after being read into IBM SPSS Modeler. To specify formatting on a per-field basis elsewhere in the stream, use the Format tab of a Type node. See the topic "Field Format Settings Tab" on page 132 for more information.

Options vary depending on the storage type. For example, for the Real storage type, you can select **Period (.)** or **Comma (,)** as the decimal separator. For timestamp fields, a separate dialog box opens when you select **Specify** from the drop-down list. See the topic "Setting Field Format options" on page 133 for more information.

For all storage types, you can also select **Stream default** to use the stream default settings for import. Stream settings are specified in the stream properties dialog box.

Additional Options

Several other options can be specified using the Data tab:

- To view storage settings for data that are no longer connected through the current node (train data, for example), select **View unused field settings**. You can clear the legacy fields by clicking **Clear**.
- At any point while working in this dialog box, click **Refresh** to reload fields from the data source. This is useful when you are altering data connections to the source node or when you are working between tabs on the dialog box.

List storage and associated measurement levels

Introduced in SPSS Modeler version 17, to work with the new measurement levels of Geospatial and Collection, a List storage field contains multiple values for a single record. Lists are enclosed in square brackets ([]). Examples of lists are: `[1,2,4,16]` and `["abc", "def"]`.

Lists can be imported into SPSS Modeler in one of three source nodes (Analytic Server, Geospatial, or Variable File), created within your streams through use of the Derive or Filler field operation nodes, or generated by the Merge node when using the Ranked Condition merge method.

Lists are considered to have a depth; for example, a simple list with items that are contained within single square brackets, in the format `[1,3]`, is recorded in IBM SPSS Modeler with a depth of zero. In addition

to simple lists that have a depth of zero, you can use nested lists, where each value within a list is a list itself.

The depth of a nested list depends on the associated measurement level. For Typeless there is no set depth limit, for Collection the depth is zero, and for Geospatial, the depth must be between zero and two inclusive depending on the number of nested items.

For zero depth lists you can set the measurement level as either Geospatial or Collection. Both of these levels are parent measurement levels and you set the measurement sublevel information in the **Values** dialog box. The measurement sublevel of a Collection determines the measurement level of the elements within that list. All measurement levels (except typeless and geospatial) are available as sublevels for Collections. The Geospatial measurement level has six sublevels of Point, LineString, Polygon, MultiPoint, MultiLineString, and MultiPolygon; for more information, see "Geospatial measurement sublevels" on page 124.

Note: The Collection measurement level can only be used with lists of depth 0, the Geospatial measurement level can only be used with lists of a maximum depth of 2, and the Typeless measurement level can be used with any list depth.

The following example shows the difference between a zero depth list and a nested list by using the structure of the Geospatial measurement sublevels Point and LineString:

- The Geospatial measurement sublevel of Point has a field depth of zero:

  `[1,3]` two coordinates

  `[1,3,-1]` three coordinates
- The Geospatial measurement sublevel of LineString has a field depth of one:

  `[ [1,3], [5,0] ]` two coordinates

  `[ [1,3,-1], [5,0,8] ]` three coordinates

The Point field (with a depth of zero) is a normal list where each value is made up of two or three coordinates. The LineString field (with a depth of one) is a list of points, where each point is made up of a further series of list values.

For more information about list creation, see "Deriving a list or geospatial field" on page 139.

Unsupported control characters

Some of the processes in SPSS Modeler cannot handle data that includes various control characters. If your data uses these characters you may see an error message such as the following example:

```
Unsupported control characters found in values of field {0}
```

The unsupported characters are: from `0x0` to `0x3F` inclusive, and `0x7F`; however, the tab (`0x9(\t)`), new line (`0xA(\n)`), and carriage return (`0xD(\r)`) characters do not cause a problem.

If you see an error message that relates to unsupported characters, in your stream, after your Source node, use a Filler node and the CLEM expression **`stripctrlchars`** to replace the characters.

Analytic Server Source node

The Analytic Server source enables you to run a stream on Hadoop Distributed File System (HDFS). The information in an Analytic Server data source can come from a variety of places, including:

- Text files on HDFS
- Databases
- HCatalog

Typically, a stream with an Analytic Server Source will be executed on HDFS; however, if a stream contains a node that is not supported for execution on HDFS, then as much of the stream as possible will be "pushed back" to Analytic Server, and then SPSS Modeler Server will attempt to process the remainder of

the stream. You will need to subsample very large datasets; for example, by placing a Sample node within the stream.

If you want to use your own Analytic Server connection instead of the default connection defined by your administrator, deselect **Use default Analytic Server** and select your connection.

Data source. Assuming you or your SPSS Modeler Server administrator has established a connection, you select a data source containing the data you wish to use. A data source contains the files and metadata associated with that source. Click **Select** to display a list of available data sources. See the topic "Selecting a data source" on page 10 for more information.

If you need to create a new data source or edit an existing one, click **Launch Data Source Editor...**.

Note that using multiple Analytic Server connections can be useful in controlling the flow of data. For example, when using the Analytic Server Source and Export nodes, you may want to use different Analytic Server connections in different branches of a stream so that when each branch runs it uses its own Analytic Server and no data will be pulled to the IBM SPSS Modeler Server. Note that if a branch contains more than one Analytic Server connection, the data will be pulled from the Analytic Servers to the IBM SPSS Modeler Server.

Selecting a data source

The Data Sources table displays a list of the available data sources. Select the source you want to use and click **OK**.

Click **Show Owner** to display the data source owner.

Filter by enables you to filter the data source listing on **Keyword**, which checks the filter criteria against the data source name and data source description, or **Owner**. You can enter a combination of string, numeric, or wild card characters described below as filter criteria. The search string is case sensitive. Click **Refresh** to update the Data Sources table.

_
An underscore can be used to represent any single character in the search string.

%
A percent sign can be used to represent any sequence of zero or more characters in the search string.

Amending credentials

If your credentials for accessing Analytic Server are different from the credentials for accessing SPSS Modeler Server, you will need to enter the Analytic Server credentials when running a stream on Analytic Server. If you do not know your credentials, contact your server administrator.

Supported nodes

Many SPSS Modeler nodes are supported for execution on HDFS, but there may be some differences in the execution of certain nodes, and some are not currently supported. This topic details the current level of support.

General

- Some characters that are normally acceptable within a quoted Modeler field name will not be accepted by Analytic Server.
- For a Modeler stream to be run in Analytic Server, it must begin with one or more Analytic Server Source nodes and end in a single modeling node or Analytic Server Export node.
- It is recommended that you set the storage of continuous targets as real rather than integer. Scoring models always write real values to the output data files for continuous targets, while the output data model for the scores follows the storage of the target. Thus, if a continuous target has integer storage, there will be a mismatch in the written values and the data model for the scores, and this mismatch will cause errors when you attempt to read the scored data.
- If a field measurement is Geospatial, the function for @OFFSET is not supported.

Source

- A stream that begins with anything other than an Analytic Server source node will be run locally.

Record operations

All Record operations are supported, with the exception of the Streaming TS and Space-Time-Boxes nodes. Further notes on supported node functionality follow.

Select

- Supports the same set of functions supported by the Derive node.

Sample

- Block-level sampling is not supported.
- Complex Sampling methods are not supported.
- First n sampling with `"Discard sample"` is not supported.
- First n sampling with `N>20000` is not supported.
- 1-in-n sampling is not supported when `"Maximum sample size"` not set.
- 1-in-n sampling isnot supported when `N * "Maximum sample size" > 20000`.
- Random % block level sampling is not supported.
- Random % currently supports supplying a seed.

Aggregate

- Contiguous keys are not supported. If you are reusing an existing stream that is set up to sort the data and then use this setting in the Aggregate node, change the stream to remove the Sort node.
- Order statistics (Median, 1st Quartile, 3rd Quartile) are computed approximately, and supported through the Optimization tab.

Sort

- The Optimization tab is not supported.

In a distributed environment, there are a limited number of operations that preserve the record order established by the Sort node.

- A Sort followed by an Export node produces a sorted data source.
- A Sort followed by a Sample node with **First** record sampling returns the first *N* records.

In general, you should place a Sort node as close as possible to the operations that need the sorted records.

Merge

- Merge by Order is not supported.
- The Optimization tab is not supported.
- Merge operations are relatively slow. If you have available space in HDFS, it can be much faster to merge your data sources once and use the merged source in following streams than to merge the data sources in each stream.

R Transform

The R syntax in the node should consist of record-at-a-time operations.

Field operations

All Field operations are supported, with the exception of the Anonymize, Transpose, Time Intervals, and History nodes. Further notes on supported node functionality follow.

Auto Data Prep

- Training the node is not supported. Applying the transformations in a trained Auto Data Prep node to new data is supported.

Derive

- All Derive functions are supported, with the exception of sequence functions.
- Deriving a new field as a Count is essentially a sequence operation, and thus not supported.
- Split fields cannot be derived in the same stream that uses them as splits; you will need to create two streams; one that derives the split field and one that uses the field as splits.

Filler

- Supports the same set of functions supported by the Derive node.

Binning

The following functionality is not supported.

- Optimal binning
- Ranks
- Tiles -> Tiling: Sum of values
- Tiles -> Ties: Keep in current and Assign randomly
- Tiles ->Custom N: Values over 100, and any N value where 100 % N is not equal to zero.

RFM Analysis

- The Keep in current option for handling ties is not supported. RFM recency, frequency, and monetary scores will not always match those computed by Modeler from the same data. The score ranges will be the same but score assignments (bin numbers) may differ by one.

Graphs

All Graph nodes are supported.

Modeling

The following Modeling nodes are supported: Time Series, TCM, Isotonic-AS, Extension Model, Tree-AS, C&R Tree, Quest, CHAID, Linear, Linear-AS, Neural Net, GLE, LSVM, TwoStep-AS, Random Trees, STP, Association Rules, XGBoost-AS, Random Forest, and K-Means-AS. Further notes on those nodes' functionality follow.

Linear

When building models on big data, you will typically want to change the objective to Very large datasets, or specify splits.

- Continued training of existing PSM models is not supported.
- The Standard model building objective is only recommended if split fields are defined so that the number of records in each split is not too large, where the definition of "too large" is dependent upon the power of individual nodes in your Hadoop cluster. By contrast, you also need to be careful to ensure that splits are not defined so finely that there are too few records to build a model.
- The Boosting objective is not supported.
- The Bagging objective is not supported.
- The Very large datasets objective is not recommended when there are few records; it will often either not build a model or will build a degraded model.
- Automatic Data Preparation is not supported. This can cause problems when trying to build a model on data with many missing values; normally these would be imputed as part of automatic data preparation. A workaround would be to use a tree model or a neural network with the Advanced setting to impute missing values selected.
- The accuracy statistic is not computed for split models.

Neural Net

When building models on big data, you will typically want to change the objective to Very large datasets, or specify splits.

- Continued training of existing standard or PSM models is not supported.

- The Standard model building objective is only recommended if split fields are defined so that the number of records in each split is not too large, where the definition of "too large" is dependent upon the power of individual nodes in your Hadoop cluster. By contrast, you also need to be careful to ensure that splits are not defined so finely that there are too few records to build a model.
- The Boosting objective is not supported.
- The Bagging objective is not supported.
- The Very large datasets objective is not recommended when there are few records; it will often either not build a model or will build a degraded model.
- When there are many missing values in the data, use the Advanced setting to impute missing values.
- The accuracy statistic is not computed for split models.

C&R Tree, CHAID, and Quest

When building models on big data, you will typically want to change the objective to Very large datasets, or specify splits.

- Continued training of existing PSM models is not supported.
- The Standard model building objective is only recommended if split fields are defined so that the number of records in each split is not too large, where the definition of "too large" is dependent upon the power of individual nodes in your Hadoop cluster. By contrast, you also need to be careful to ensure that splits are not defined so finely that there are too few records to build a model.
- The Boosting objective is not supported.
- The Bagging objective is not supported.
- The Very large datasets objective is not recommended when there are few records; it will often either not build a model or will build a degraded model.
- Interactive sessions is not supported.
- The accuracy statistic is not computed for split models.
- When a split field is present, tree models built locally in Modeler are slightly different from tree models built by Analytic Server, and thus produce different scores. The algorithms in both cases are valid; the algorithms used by Analytic Server are simply newer. Given the fact that tree algorithms tend to have many heuristic rules, the difference between the two components is normal.

Model scoring

All models supported for modeling are also supported for scoring. In addition, locally-built model nuggets for the following nodes are supported for scoring: C&RT, Quest, CHAID, Linear, and Neural Net (regardless of whether the model is standard, boosted bagged, or for very large datasets), Regression, C5.0, Logistic, Genlin, GLMM, Cox, SVM, Bayes Net, TwoStep, KNN, Decision List, Discriminant, Self Learning, Anomaly Detection, Apriori, Carma, K-Means, Kohonen, R, and Text Mining.

- No raw or adjusted propensities will be scored. As a workaround you can get the same effect by manually computing the raw propensity using a Derive node with the following expression: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

R

The R syntax in the nugget should consist of record-at-a-time operations.

Output

The Matrix, Analysis, Data Audit, Transform, Set Globals, Statistics, Means, and Table nodes are supported. Further notes on supported node functionality follow.

Data Audit

The Data Audit node cannot produce the mode for continuous fields.

Means

The Means node cannot produce a standard error or 95% confidence interval.

Table
The Table node is supported by writing a temporary Analytic Server data source containing the results of upstream operations. The Table node then pages through the contents of that data source.

Export
A stream can begin with an Analytic Server source node and end with an export node other than the Analytic Server export node, but data will move from HDFS to SPSS Modeler Server, and finally to the export location.

Database source node

The Database source node can be used to import data from a variety of other packages using ODBC (Open Database Connectivity), including Microsoft SQL Server, Db2, Oracle, and others.

To read or write to a database, you must have an ODBC data source installed and configured for the relevant database, with read or write permissions as needed. The IBM SPSS Data Access Pack includes a set of ODBC drivers that can be used for this purpose, and these drivers are available from the download site. If you have questions about creating or setting permissions for ODBC data sources, contact your database administrator.

Supported ODBC drivers

For the latest information on which databases and ODBC drivers are supported and tested for use with IBM SPSS Modeler, see the product compatibility matrices on the corporate Support site (http://www.ibm.com/support).

Where to install drivers

Note: ODBC drivers must be installed and configured on each computer where processing may occur.

- If you are running IBM SPSS Modeler in local (standalone) mode, the drivers must be installed on the local computer.
- If you are running IBM SPSS Modeler in distributed mode against a remote IBM SPSS Modeler Server, the ODBC drivers need to be installed on the computer where IBM SPSS Modeler Server is installed. For IBM SPSS Modeler Server on UNIX systems, see also "Configuring ODBC drivers on UNIX systems" later in this section.
- If you need to access the same data sources from both IBM SPSS Modeler and IBM SPSS Modeler Server, the ODBC drivers must be installed on both computers.
- If you are running IBM SPSS Modeler over Terminal Services, the ODBC drivers need to be installed on the Terminal Services server on which you have IBM SPSS Modeler installed.

Access data from a database

To access data from a database, complete the following steps.

- Install an ODBC driver and configure a data source to the database you want to use.
- In the **Database** node dialog box, connect to a database using Table mode or SQL Query mode.
- Select a table from the database.
- Using the tabs in the **Database** node dialog box, you can alter usage types and filter data fields.

More details on the preceding steps are given in the related documentation topics.

Note: If you call database Stored Procedures (SPs) from SPSS Modeler you might see a single output field returned named `RowsAffected` rather than the expected output of the SP. This occurs when ODBC does not return sufficient information to be able to determine the output datamodel of the SP. SPSS Modeler has only limited support for SPs that return output and it is suggested that instead of using SPs you should extract the SELECT from the SP and use either of the following actions.

- Create a view based on the SELECT and choose the view in the Database source node
- Use the SELECT directly in the Database source node.

Setting Database Node Options

You can use the options on the Data tab of the Database source node dialog box to gain access to a database and read data from the selected table.

Mode. Select **Table** to connect to a table using the dialog box controls.

Select **SQL Query** to query the database selected below using SQL. See the topic "Querying the database" on page 21 for more information.

Data source. For both Table and SQL Query modes, you can enter a name in the Data Source field or select **Add new database connection** from the drop-down list.

The following options are used to connect to a database and select a table using the dialog box:

Table name. If you know the name of the table you would like to access, enter it in the Table Name field. Otherwise, click the **Select** button to open a dialog box listing the available tables.

Quote table and column names. Specify whether you want table and column names to be enclosed in quotation marks when queries are sent to the database (if, for example, they contain spaces or punctuation).

- The **As needed** option will quote table and field names *only* if they include nonstandard characters. Nonstandard characters include non-ASCII characters, space characters, and any non-alphanumeric character other than a full stop (.).
- Select **Always** if you want *all* table and field names quoted.
- Select **Never** if you *never* want table and field names quoted.

Strip lead and trail spaces. Select options for discarding leading and trailing spaces in strings.

Note: Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Reading empty strings from Oracle. When reading from or writing to an Oracle database, be aware that, unlike IBM SPSS Modeler and unlike most other databases, Oracle treats and stores empty string values as equivalent to null values. This means that the same data extracted from an Oracle database may behave differently than when extracted from a file or another database, and the data may return different results.

Adding a database connection

To open a database, first select the data source to which you want to connect. On the Data tab, select **Add new database connection** from the Data Source drop-down list.

This opens the Database Connections dialog box.

Note: For an alternative way to open this dialog, from the main menu, choose: **Tools** > **Databases...**

Data sources. Lists the available data sources. Scroll down if you do not see the desired database. Once you have selected a data source and entered any passwords, click **Connect**. Click **Refresh** to update the list.

Mode. Select one of the following modes:

- **Username and password.** If the data source is password protected, enter your user name and the associated password.
- **Stored credential.** If a credential has been configured in IBM SPSS Collaboration and Deployment Services, you can select this option to browse for it in the repository. The credential's user name and password must match the user name and password required to access the database.

Connections. Shows currently connected databases.

- **Default.** You can optionally choose one connection as the default. Doing so causes Database source or export nodes to have this connection predefined as their data source, though this can be edited if desired.
- **Save.** Optionally select one or more connections that you want to redisplay in subsequent sessions.
- **Data source.** The connection strings for the currently connected databases.
- **Preset.** Indicates (with a * character) whether preset values have been specified for the database connection. To specify preset values, click this column in the row corresponding to the database connection, and choose Specify from the list. See the topic "Specifying preset values for a database connection" on page 18 for more information.

To remove connections, select one from the list and click **Remove**.

Driver. If you select **Driver** for the mode instead of **Data Source**, choose the desired driver from the list.

- In the **Attributes** field, enter the database connection string. See your database documentation for the proper string to use.
- Enter any display name.
- Enter the database user name and password and click **Connect**.

To read or write to a database, you must have an ODBC data source installed and configured for the relevant database, with read or write permissions as needed. The IBM SPSS Data Access Pack includes a set of ODBC drivers that can be used for this purpose, and these drivers are available from the download site. If you have questions about creating or setting permissions for ODBC data sources, contact your database administrator.

Supported ODBC drivers

For the latest information on which databases and ODBC drivers are supported and tested for use with IBM SPSS Modeler, see the product compatibility matrices on the corporate Support site (http://www.ibm.com/support).

Where to install drivers

Note: ODBC drivers must be installed and configured on each computer where processing may occur.

- If you are running IBM SPSS Modeler in local (standalone) mode, the drivers must be installed on the local computer.
- If you are running IBM SPSS Modeler in distributed mode against a remote IBM SPSS Modeler Server, the ODBC drivers need to be installed on the computer where IBM SPSS Modeler Server is installed. For IBM SPSS Modeler Server on UNIX systems, see also "Configuring ODBC drivers on UNIX systems" later in this section.
- If you need to access the same data sources from both IBM SPSS Modeler and IBM SPSS Modeler Server, the ODBC drivers must be installed on both computers.
- If you are running IBM SPSS Modeler over Terminal Services, the ODBC drivers need to be installed on the Terminal Services server on which you have IBM SPSS Modeler installed.

Configuring ODBC drivers on UNIX systems

By default, the DataDirect Driver Manager is not configured for IBM SPSS Modeler Server on UNIX systems. To configure UNIX to load the DataDirect Driver Manager, enter the following commands:

```
cd <modeler_server_install_directory>/bin
rm -f libspssodbc.so
```

Then run this command if you want to use the UTF8 driver wrapper:

```
ln -s libspssodbc_datadirect.so libspssodbc.so
```

Or run this command instead if you want to use the UTF16 driver wrapper:

```
ln -s libspssodbc_datadirect_utf16.so libspssodbc.so
```

Doing so removes the default link and creates a link to the DataDirect Driver Manager.

Note: The UTF16 driver wrapper is required to use SAP HANA or IBM Db2 CLI drivers for some databases. DashDB requires the IBM Db2 CLI driver.

To configure SPSS Modeler Server:

1. Configure the SPSS Modeler Server start up script `modelersrv.sh` to source the IBM SPSS Data Access Pack `odbc.sh` environment file by adding the following line to `modelersrv.sh`:

   ```
   . /<pathtoSDAPinstall>/odbc.sh
   ```

   Where `<pathtoSDAPinstall>` is the full path to your IBM SPSS Data Access Pack installation.
2. Restart SPSS Modeler Server.

In addition, for SAP HANA and IBM Db2 only, add the following parameter definition to the DSN in your `odbc.ini` file to avoid buffer overflows during connection:

```
DriverUnicodeType=1
```

Note: The `libspssodbc_datadirect_utf16.so` wrapper is also compatible with the other SPSS Modeler Server supported ODBC drivers.

Potential database issues

Depending on which database you use; there some potential issues that you should be aware of.

IBM Db2

When attempting to cache a node in a stream which reads data from a Db2 database, you may see the following error message:

```
A default table space could not be found with a pagesize of at least 4096 that authorization
 ID TEST is authorized to use
```

To configure Db2 to enable in-database caching to work properly in SPSS Modeler, the database administrator should create a "user temporary" tablespace and grant access to this tablespace to the relevant Db2 accounts.

We recommend using a pagesize of 32768 in the new tablespace, as this will increase the limit on the number of fields that can be successfully cached.

IBM Db2 for z/OS

- Scoring a subset of algorithms, with confidences enabled, using generated SQL can return an error on execution. The issue is specific to Db2 for z/OS; to fix this, use the SPSS Modeler Server Scoring Adapter for Db2 on z/OS.
- When running streams against Db2 for z/OS, you may experience database errors if the timeout for idle database connections is enabled and set too low. In Db2 for z/OS version 8, the default changed from no timeout to 2 minutes. The solution is to increase the value of the Db2 system parameter `IDLE THREAD TIMEOUT (IDTHTOIN)`, or reset the value to 0.

Oracle

When you run a stream that contains an Aggregate node, the values returned for the 1st and 3rd Quartiles, when pushing back SQL to an Oracle database, may differ from those that are returned in native mode.

Specifying preset values for a database connection

For some databases, you can specify a number of default settings for the database connection. The settings all apply to database export.

The database types that support this feature are as follows.

- SQL Server Enterprise and Developer editions. See the topic "Settings for SQL Server" on page 18 for more information.
- Oracle Enterprise or Personal editions. See the topic "Settings for Oracle" on page 18 for more information.
- IBM Db2 for z/OS and Teradata both connect to a database or schema in a similar way. See the topic "Settings for IBM Db2 for z/OS, IBM Db2 LUW, and Teradata" on page 19 for more information.

If you are connected to a database or schema that does not support this feature, you see the message **No presets can be configured for this database connection**.

Settings for SQL Server

These settings are displayed for SQL Server Enterprise and Developer editions.

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Row.** Enables row-level compression (for example, the equivalent of `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW);` in SQL).
- **Page.** Enables page-level compression (for example, `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE);` in SQL).

Settings for Oracle

Oracle settings - Basic option

These settings are displayed for Oracle Enterprise or Personal editions using the Basic option.

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Default.** Enables default compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS;` in SQL). In this case, it has the same effect as the **Basic** option.
- **Basic.** Enables basic compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS BASIC;` in SQL).

Oracle settings - Advanced option

These settings are displayed for Oracle Enterprise or Personal editions using the Advanced option.

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Default.** Enables default compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS;` in SQL). In this case, it has the same effect as the **Basic** option.

- **Basic.** Enables basic compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS BASIC;` in SQL).
- **OLTP.** Enables OLTP compression (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP;` in SQL).
- **Query Low/High.** (Exadata servers only) Enables hybrid columnar compression for query (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW;` or `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH;` in SQL). Compression for query is useful in data warehousing environments; `HIGH` provides a higher compression ratio than `LOW`.
- **Archive Low/High.** (Exadata servers only) Enables hybrid columnar compression for archive (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW;` or `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH;` in SQL). Compression for archive is useful for compressing data that will be stored for long periods of time; `HIGH` provides a higher compression ratio than `LOW`.

Settings for IBM Db2 for z/OS, IBM Db2 LUW, and Teradata

When you specify presets for IBM Db2 for z/OS, IBM Db2 LUW, or Teradata, you are prompted to select the following items:

Use server scoring adapter database or **Use server scoring adapter shema**. If selected, enables the **Server scoring adapter database** or **Server scoring adapter schema** option.

Server scoring adapter database or **Server scoring adapter schema** From the drop-down list, select the connection you require.

In addition, for Teradata, you can also set query banding details to provide additional metadata to assist with items such as workload management; collating, identifying, and resolving queries; and tracking database usage.

Spelling query banding. Choose if the query banding will be set once for the entire time you are working with a Teradata database connection (**For Session**), or if it will be set every time you run a stream (**For Transaction**).

Note: If you set query banding on a stream, the banding is lost if you copy that stream to another machine. To prevent this, you can use scripting to run your stream and use the key word *querybanding* in the script to apply the settings you require.

Required database permissions

For SPSS Modeler database capabilities to function correctly, grant access to the following items to any user IDs used:

Db2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

Db2/z

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Teradata

DBC.Functions

DBC.USERS

Selecting a Database Table

After you have connected to a data source, you can choose to import fields from a specific table or view. From the Data tab of the Database dialog box, you can either enter the name of a table in the Table Name field or click **Select** to open the Select Table/View dialog box which lists the available tables and views.

Show table owner. Select if a data source requires that the owner of a table must be specified before you can access the table. Deselect this option for data sources that do not have this requirement.

Note: SAS and Oracle databases usually require you to show the table owner.

Tables/Views. Select the table or view to import.

Show. Lists the columns in the data source to which you are currently connected. Click one of the following options to customize your view of the available tables:

- Click **User Tables** to view ordinary database tables created by database users.
- Click **System Tables** to view database tables owned by the system (for example, tables that provide information about the database, such as details of indexes). This option can be used to view the tabs used in Excel databases. (Note that a separate Excel source node is also available. See the topic "Excel Source Node" on page 39 for more information.)
- Click **Views** to view virtual tables based on a query involving one or more ordinary tables.
- Click **Synonyms** to view synonyms created in the database for any existing tables.

Name/Owner filters. These fields allow you to filter the list of displayed tables by name or owner. For example, type SYS to list only tables with that owner. For wildcard searches, an underscore (_) can be used to represent any single character and a percent sign (%) can represent any sequence of zero or more characters.

Set As Default. Saves the current settings as the default for the current user. These settings will be restored in the future when a user opens a new table selector dialog box *for the same data source name and user login only*.

Querying the database

Once you've connected to a data source, you can choose to import fields using SQL queries. From the main dialog box, select **SQL Query** as the connection mode. This adds a query editor window in the dialog box. Using the query editor, you can create or load one or more SQL queries whose result set will be read into the data stream.

If you specify multiple SQL queries, separate them with semicolons (;) and ensure that there's no multiple SELECT statement.

To cancel and close the query editor window, select **Table** as the connection mode.

You can include SPSS Modeler stream parameters (a type of user-defined variable) in the SQL query. See "Using Stream Parameters in an SQL Query" on page 21 for more information.

Load Query. Click to open the file browser and load a previously saved query.

Save Query. Click to open the Save Query dialog box to save the current query.

Import Default. Click to import an example SQL SELECT statement constructed automatically using the table and columns selected in the dialog box.

Clear. Clear the contents of the work area. Use this option when you want to start over.

Split text. The default option **Never** means that the query will be sent to the database as a whole. Alternatively, you can choose **As needed**, which means that SPSS Modeler attempts to parse the query and identify if there are SQL statements that should be send to the database one after another.

Important: Depending on the database you're using, SPSS Modeler may attempt to run the custom SQL you entered to obtain the data model the SQL results in. For example, when using MySQL or Google BigQuery, by default SQLExecute() is called when getting the table schema; so SPSS Modeler will run the SQL and obtain the data model. This isn't the case for most database drivers.

If you want to avoid this, see "Using a custom database configuration file" on page 22 for details about customizing how SPSS Modeler processes SQL in such cases.

Using Stream Parameters in an SQL Query

When writing an SQL query to import fields, you can include SPSS Modeler stream parameters that have been previously defined. All types of stream parameter are supported.

The following table shows how some examples of stream parameters will be interpreted in the SQL query.

Table 3. Examples of stream parameters

Stream Parameter Name (example)	Storage	Stream Parameter Value	Interpreted as
PString	String	ss	'ss'
PInt	Integer	5	5
PReal	Real	5.5	5.5
PTime	Time	23:05:01	t{'23:05:01'}
PDate	Date	2011-03-02	d{'2011-03-02'}

Table 3. Examples of stream parameters (continued)			
Stream Parameter Name (example)	**Storage**	**Stream Parameter Value**	**Interpreted as**
PTimeStamp	TimeStamp	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Unknown	IntValue	IntValue

In the SQL query, you specify a stream parameter in the same way as in a CLEM expression, namely by `'$P-<parameter_name>'`, where `<parameter_name>` is the name that has been defined for the stream parameter.

When referencing a field, the storage type must be defined as Unknown, and the parameter value must be enclosed in quotation marks if needed. Thus, using the examples shown in the table, if you entered the SQL query:

```
select "IntValue" from Table1 where "IntValue" < '$P-PInt';
```

it would be evaluated as:

```
select "IntValue" from Table1 where "IntValue" < 5;
```

If you were to reference the `IntValue` field by means of the `PColumn` parameter, you would need to specify the query as follows to get the same result:

```
select "IntValue" from Table1 where "'$P-PColumn'" < '$P-PInt';
```

Using a custom database configuration file

If you need to customize the way SPSS Modeler processes SQL, you can use a custom database configuration file. This allows for a range of customizations you might need depending on your specific circumstances. Full documentation for this capability isn't available. Contact your database administrator or IBM Support for assistance with your specific needs regarding this capability, but note that IBM doesn't support customizations done via these configuration files that aren't provided by IBM.

To implement a custom database configuration file

1. Create a `.cfg` file for your specific database.
2. Add options as desired. Extensive customization is possible. Note that the file must be properly named and properly formatted according to your specific database.
3. Place the file in the `config` folder of your SPSS Modeler Server and/or SPSS Modeler Client installation directory.

The `.cfg` file name must follow the format `odbc-<db>-custom-properties.cfg`, where `<db>` is one of the following database names:

- `bigquery`
- `db2`
- `greenplum`
- `hana`
- `hive`
- `impala`
- `informix`
- `mssql`
- `mysql`

- neoview
- netezza
- oracle
- postgresql
- redshift
- soliddb
- sybase
- teradata
- vertica

Example: execute_while_getting_schema

When using MySQL database with custom SQL in a Database source node, by default the MySQL database driver calls SQLExecute() when getting the table schema. As a result, SPSS Modeler needs to run your SQL and obtain the data model. If you don't want SPSS Modeler to obtain the data model, follow these steps:

1. Create a file called odbc-mysql-custom-properties.cfg.
2. Add the following line and set it to N to override the default behavior for MySQL:

```
execute_while_getting_schema, N
```

3. Copy the file into the config directory of your SPSS Modeler Server and SPSS Modeler Client installations.

Example: sqlmx_sort_by

By default, order_by isn't permitted within nested SQL during SQL pushback. By enabling order_by, you can force SQL pushback to work for the Sample node when there's a Sort node ahead of it. For example, to enable order_by on Google BigQuery database, follow these steps:

1. Create a file called odbc-bigquery-custom-properties.cfg.
2. Add the following line and set it to Y to override the default behavior for Google BigQuery:

```
sqlmx_sort_by, Y
```

3. Copy the file into the config directory of your SPSS Modeler Server and SPSS Modeler Client installations.

Example: uda_list_sql_basic and uda_list_sql_parameter

You may want to use custom SQL to retrieve database functions and aggregate functions. For example, on Oracle database, follow these steps:

1. Create a file called odbc-oracle-custom-properties.cfg.
2. Add the following lines to the file:

```
#Define the UDA (database window aggregates) sqls

uda_list_sql_basic, "SELECT '<src_database_name>',OBJECT_NAME,OWNER,'', '', CASE WHEN
OWNER='SYS' THEN 1 ELSE 0 END BUILTIN FROM ALL_ARGUMENTS WHERE OBJECT_ID IN (SELECT
OBJECT_ID FROM ALL_PROCEDURES WHERE AGGREGATE = 'YES') AND OBJECT_ID NOT IN (SELECT DISTINCT
OBJECT_ID FROM ALL_ARGUMENTS WHERE PLS_TYPE IS NULL) AND ARGUMENT_NAME IS NULL ORDER BY
OBJECT_ID"

uda_list_sql_parameter, "SELECT POSITION,
DATA_PRECISION,DATA_SCALE,DATA_TYPE,'',DATA_TYPE,'',0 FROM ALL_ARGUMENTS WHERE OBJECT_ID IN
(SELECT OBJECT_ID FROM ALL_PROCEDURES WHERE AGGREGATE = 'YES') AND OBJECT_ID NOT IN (SELECT
DISTINCT OBJECT_ID FROM ALL_ARGUMENTS WHERE PLS_TYPE IS NULL) ORDER BY OBJECT_ID,POSITION"
```

3. Copy the file into the `config` directory of your SPSS Modeler Server and SPSS Modeler Client installations.

`uda_list_sql_basic` and `uda_list_sql_parameter` may use other custom SQL provided they conform to the following table schemas below (step 2 is an example).

```
#table schema for uda_list_sql_basic
databaseName,function,schema,catalog,description,isBuiltIn

#table schema for uda_list_sql_parameter
position,precision,scale,returnType,returnTypeName,parameterTypes,parameterTypeNames,isVarArg
```

Variable File Node

You can use Variable File nodes to read data from free-field text files (files whose records contain a constant number of fields but a varied number of characters), also known as delimited text files. This type of node is also useful for files with fixed-length header text and certain types of annotations. Records are read one at a time and passed through the stream until the entire file is read.

Note regarding reading in geospatial data

If the node contains geospatial data, and the node was created as an export from a flat file, you must follow some extra steps to set up the geospatial metadata. For more information, see "Importing geospatial data into the Variable File Node" on page 26.

Notes for reading in delimited text data

- Records must be delimited by a newline character at the end of each line. The newline character must not be used for any other purpose (such as within any field name or value). Leading and trailing spaces should ideally be stripped off to save space, although this is not critical. Optionally these spaces can be stripped out by the node.
- Fields must be delimited by a comma or other character that ideally is used only as a delimiter, meaning it does not appear within field names or values. If this is not possible, then all text fields can be wrapped in double quotation marks, if none of the field names or text values contains a double quotation mark. If field names or values do contain double quotation marks, then text fields can be wrapped in single quotation marks as an alternative, again if single quotation marks are not used elsewhere within values. If neither single or double quotation marks can be used, then text values need to be amended to remove or replace either the delimiter character, or single or double quotation marks.
- Each row, including the header row, should contain the same number of fields.
- The first line should contain the field names. If not, clear **Read field names from file** to give each field a generic name such as Field1, Field2, and so on.
- The second line must contain the first record of data. There must no blank lines or comments.
- Numeric values must not include the thousands separator or grouping symbol—without the comma in 3,000.00, for example. The decimal indicator (period or full-stop in the US or the UK) must be used only where appropriate.
- Date and time values should be in one of the formats that are recognized in the Stream Options dialog box, such as `DD/MM/YYYY` or `HH:MM:SS`. All dates and time fields in the file should ideally follow the same format, and any field that contains a date must use the same format for all values within that field.

Setting options for the Variable File Node

You set the options on the File tab of the Variable File node dialog box.

File Specify the name of the file. You can enter a filename or click the ellipsis button (**...**) to select a file. The file path is shown once you select a file, and its contents are displayed with delimiters in the panel below it.

The sample text that is displayed from your data source can be copied and pasted into the following controls: EOL comment characters and user-specified delimiters. Use Ctrl-C and Ctrl-V to copy and paste.

Read field names from file Selected by default, this option treats the first row in the data file as labels for the column. If your first row is not a header, deselect to automatically give each field a generic name, such as *Field1, Field2*, for the number of fields in the dataset.

Specify number of fields. Specify the number of fields in each record. The number of fields can be detected automatically as long as the records are new-line terminated. You can also set a number manually.

Skip header characters. Specify how many characters you want to ignore at the beginning of the first record.

EOL comment characters. Specify characters, such as # or !, to indicate annotations in the data. Wherever one of these characters appears in the data file, everything up to but not including the next new-line character will be ignored.

Strip lead and trail spaces. Select options for discarding leading and trailing spaces in strings on import.

Note: Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Invalid characters. Select **Discard** to remove invalid characters from the data source. Select **Replace with** to replace invalid characters with the specified symbol (one character only). Invalid characters are null characters or any character that does not exist in the encoding method specified.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol Select the type of decimal separator that is used in your data source. The **Stream default** is the character that is selected from the Options tab of the stream properties dialog box. Otherwise, select either **Period (.)** or **Comma (,)** to read all data in this dialog box using the chosen character as the decimal separator.

Line delimiter is newline character To use the newline character as the line delimiter, instead of a field delimiter, select this option. For example, this may be useful if there is an odd number of delimiters in a row that cause the line to wrap. Note that selecting this option means you cannot select **Newline** in the Delimiters list.

Note: If you select this option, any blank values at the end of data rows will be stripped out.

Delimiters. Using the check boxes listed for this control, you can specify which characters, such as the comma (,), define field boundaries in the file. You can also specify more than one delimiter, such as ", |" for records that use multiple delimiters. The default delimiter is the comma.

Note: If the comma is also defined as the **Decimal symbol**, the default settings here will not work. In cases where the comma is both the **Field delimiter** and the **Decimal symbol**, select **Other** in the **Field delimiters** list. Then manually specify a comma in the entry field.

Select **Allow multiple blank delimiters** to treat multiple adjacent blank delimiter characters as a single delimiter. For example, if one data value is followed by four spaces and then another data value, this group would be treated as two fields rather than five.

Lines to scan for column and type Specify how many lines and columns to scan for specified data types.

Automatically recognize dates and times To enable IBM SPSS Modeler to automatically attempt to recognize data entries as dates or times, select this check box. For example, this means that an entry such as 07-11-1965 will be identified as a date and 02:35:58 will be identified as a time; however, ambiguous entries such as 07111965 or 023558 will show up as integers since there are no delimiters between the numbers.

Note: To avoid potential data problems when you use data files from previous versions of IBM SPSS Modeler, this box is turned off by default for information that is saved in versions prior to 13.

Treat square brackets as lists If you select this check box, the data included between opening and closing square brackets is treated as a single value, even if the content includes delimiter characters such as commas and double quotes. For example, this might include two or three dimensional geospatial data, where the coordinates contained within square brackets are processed as a single list item. For more information, see "Importing geospatial data into the Variable File Node" on page 26

Quotes. Using the drop-down lists, you can specify how single and double quotation marks are treated on import. You can choose to **Discard** all quotation marks, **Include as text** by including them in the field value, or **Pair and discard** to match pairs of quotation marks and remove them. If a quotation mark is unmatched, you will receive an error message. Both **Discard** and **Pair and discard** store the field value (without quotation marks) as a string.

Note: When using **Pair and discard**, spaces are kept. When using **Discard**, trailing spaces inside and outside quotes are removed (for example, **`' " ab c" , "d ef " , " gh i " '`** will result in **`'ab c, d ef, gh i'`**). When using **Include as text**, quotes are treated as normal characters, so leading and trailing spaces will be stripped naturally.

At any point while you are working in this dialog box, click **Refresh** to reload fields from the data source. This is useful when you are altering data connections to the source node or when you are working between tabs in the dialog box.

Importing geospatial data into the Variable File Node

If the node contains geospatial data, was created as an export from a flat file, and is used in the same stream in which it was created, the node retains the geospatial metadata and no further configuration steps are needed.

However, if the node is exported and used in a different stream, the geospatial list data is automatically converted to a string format; you must follow some extra steps to restore the list storage type and the associated geospatial metadata.

For more information about lists, see "List storage and associated measurement levels" on page 8.

For more information about the details you can set as geospatial metadata, see "Geospatial measurement sublevels" on page 124.

To set up the geospatial metadata, use the following steps.

1. On the File tab of the Variable File node, select the **Treat square brackets as lists** check box. Selecting this check box means that the data included between opening and closing square brackets is treated as a single value, even if the content includes delimiter characters such as commas and double quotation marks. Failure to select this check box means that your data is read as a string storage type, any commas in the field are processed as delimiters, and your data structure is interpreted incorrectly.
2. If your data includes single or double quotation marks, select the **Pair and discard** option in the **Single quotes** and **Double quotes** fields, as appropriate.
3. On the Data tab of the Variable File node, for the geospatial data fields, select the **Override** check box, and change the **Storage** type from a string to a list.
4. By default, the list **Storage** type is set as a *List of real* and the underlying value storage type of the list field is set to *Real*. To change the underlying value storage type or the depth, click **Specify...** to display the **Storage** subdialog box.
5. In the **Storage** subdialog box you can modify the following settings:
   - **Storage** Specify the overall storage type of the data field. By default the storage type is set to List; however, the drop-down list contains all other storage types (String, Integer, Real, Date, Time, and Timestamp). If you select any storage type other than List, **Value storage** and **Depth** options are unavailable.

- **Value storage** Specify the storage types of the elements in the list, as opposed to the field as a whole. When you import geospatial fields, the only relevant storage types are Real and Integer; the default setting is Real.
- **Depth** Specify the depth of the list field. The required depth depends on the type of geospatial field and follows these criteria:
  - Point – 0
  - LineString – 1
  - Polygon – 1
  - MultiPoint – 1
  - MultiLineString – 2
  - MultiPolygon – 2

  Warning: You must know the type of geospatial field you are converting back into a list and the required depth for that kind of field. If this information is set incorrectly, it is not possible to use the field.

6. On the Types tab of the Variable File node, for the geospatial data field, ensure the **Measurement** cell contains the correct measurement level. To change the measurement level, in the **Measurement** cell, click **Specify...** to display the **Values** dialog box.
7. In the **Values** dialog box, the **Measurement**, **Storage**, and **Depth** for the list are displayed. Select the **Specify values and labels** option and, from the **Type** drop-down list, select the correct type for the **Measurement**. Depending on the **Type**, you might be prompted for more details such as if the data represents 2 or 3 dimensions and which coordinate system is used.

Fixed File Node

You can use Fixed File nodes to import data from fixed-field text files (files whose fields are not delimited but start at the same position and are of a fixed length). Machine-generated or legacy data are frequently stored in fixed-field format. Using the File tab of the Fixed File node, you can easily specify the position and length of columns in your data.

Setting options for the Fixed File node

The File tab of the Fixed File node enables you to bring data into IBM SPSS Modeler and to specify the position of columns and length of records. Using the data preview pane in the center of the dialog box, you can click to add arrows specifying the break points between fields.

File. Specify the name of the file. You can enter a filename or click the ellipsis button (**...**) to select a file. Once you have selected a file, the file path is shown and its contents are displayed with delimiters in the panel below.

The data preview pane can be used to specify column position and length. The ruler at the top of the preview window helps to measure the length of variables and to specify the break point between them. You can specify break point lines by clicking in the ruler area above the fields. Break points can be moved by dragging and can be discarded by dragging them outside of the data preview region. The ruler is designed to handle ASCII characters.

- Each break-point line automatically adds a new field to the fields table below.
- Start positions indicated by the arrows are automatically added to the Start column in the table below.

Line oriented. Select if you want to skip the new-line character at the end of each record.

Skip header lines. Specify how many lines you want to ignore at the beginning of the first record. This is useful for ignoring column headers.

Record length. Specify the number of characters in each record.

Field. All fields that you have defined for this data file are listed here. There are two ways to define fields:

- Specify fields interactively using the data preview pane above.
- Specify fields manually by adding empty field rows to the table below. Click the button to the right of the fields pane to add new fields. Then, in the empty field, enter a field name, a start position, and a length. These options will automatically add arrows to the data preview pane, which can be easily adjusted.

To remove a previously defined field, select the field in the list and click the red delete button.

Start. Specify the position of the first character in the field. For example, if the second field of a record begins on the sixteenth character, you would enter 16 as the starting point.

Length. Specify how many characters are in the longest value for each field. This determines the cutoff point for the next field.

Strip lead and trail spaces. Select to discard leading and trailing spaces in strings on import.

Note: Comparisons between strings that do and do not use SQL pushback may generate different results where trailing spaces exist.

Invalid characters. Select **Discard** to remove invalid characters from the data input. Select **Replace with** to replace invalid characters with the specified symbol (one character only). Invalid characters are null (0) characters or any character that does not exist in the current encoding.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol. Select the type of decimal separator used in your data source. **Stream default** is the character selected from the Options tab of the stream properties dialog box. Otherwise, select either **Period (.)** or **Comma (,)** to read all data in this dialog box using the chosen character as the decimal separator.

Automatically recognize dates and times. To enable IBM SPSS Modeler to automatically attempt to recognize data entries as dates or times, select this check box. For example, this means that an entry such as 07-11-1965 will be identified as a date and 02:35:58 will be identified as a time; however, ambiguous entries such as 07111965 or 023558 will show up as integers since there are no delimiters between the numbers.

Note: To avoid potential data problems when using data files from previous versions of IBM SPSS Modeler, this box is turned off by default for information saved in versions prior to 13.

Lines to scan for type. Specify how many lines to scan for specified data types.

At any point while working in this dialog box, click **Refresh** to reload fields from the data source. This is useful when altering data connections to the source node or when working between tabs on the dialog box.

Statistics File Node

You can use the Statistics File node to read data directly from a saved IBM SPSS Statistics file (*.sav* or *.zsav*). This format is now used to replace the cache file from earlier versions of IBM SPSS Modeler. If you would like to import a saved cache file, you should use the IBM SPSS Statistics File node.

Import file. Specify the name of the file. You can enter a filename or click the ellipsis button (**...**) to select a file. The file path is shown once you have selected a file.

File is password encrypted. Select this box if you know the file is password protected; you are the prompted to enter the **Password**. If the file is password protected, and you do not enter the password, a warning message is displayed if you attempt to change to another tab, refresh the data, preview the node contents, or try to execute a stream containing the node.

Note: Password protected files can only be opened by IBM SPSS Modeler version 16 or greater.

Variable names. Select a method of handling variable names and labels upon import from an IBM SPSS Statistics *.sav* or *.zsav* file. Metadata that you choose to include here persists throughout your work in IBM SPSS Modeler and may be exported again for use in IBM SPSS Statistics.

- **Read names and labels.** Select to read both variable names and labels into IBM SPSS Modeler. By default, this option is selected and variable names are displayed in the Type node. Labels may be displayed in charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box. By default, the display of labels in output is disabled.
- **Read labels as names.** Select to read the descriptive variable labels from the IBM SPSS Statistics *.sav* or *.zsav* file rather than the short field names, and use these labels as variable names in IBM SPSS Modeler.

Values. Select a method of handling values and labels upon import from an IBM SPSS Statistics *.sav* or *.zsav* file. Metadata that you choose to include here persists throughout your work in IBM SPSS Modeler and may be exported again for use in IBM SPSS Statistics.

- **Read data and labels.** Select to read both actual values and value labels into IBM SPSS Modeler. By default, this option is selected and values themselves are displayed in the Type node. Value labels may be displayed in the Expression Builder, charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box.
- **Read labels as data.** Select if you want to use the value labels from the *.sav* or *.zsav* file rather than the numerical or symbolic codes used to represent the values. For example, selecting this option for data with a gender field whose values of 1 and 2 actually represent *male* and *female*, respectively, will convert the field to a string and import *male* and *female* as the actual values.

  It is important to consider missing values in your IBM SPSS Statistics data before selecting this option. For example, if a numeric field uses labels only for missing values (0 = *No Answer*, –99 = *Unknown*), then selecting the option above will import only the value labels *No Answer* and *Unknown* and will convert the field to a string. In such cases, you should import the values themselves and set missing values in a Type node.

Use field format information to determine storage. If this box is cleared, field values that are formatted in the *.sav* file as integers (i.e., fields specified as F*n*.0 in the Variable View in IBM SPSS Statistics) are imported using integer storage. All other field values except strings are imported as real numbers.

If this box is selected (default), all field values except strings are imported as real numbers, whether formatted in the *.sav* file as integers or not.

Multiple response sets. Any multiple response sets defined in the IBM SPSS Statistics file will automatically be preserved when the file is imported. You can view and edit multiple response sets from any node with a Filter tab. See the topic "Editing Multiple Response Sets" on page 135 for more information.

Data Collection Node

Data Collection source nodes import survey data based on the Survey Reporter Developer Kit that is provided with your Data Collection product. This format distinguishes *case data*—the actual responses to questions gathered during a survey—from the *metadata* that describes how the case data is collected and organized. Metadata consists of information such as question texts, variable names and descriptions, multiple response variable definitions, translations of the text strings, and the definition of the structure of the case data.

Note: This node requires the Survey Reporter Developer Kit, which is distributed along with yourData Collection product. Aside from installing the Developer Kit, no additional configuration is required.

Comments

- Survey data is read from the flat, tabular VDATA format, or from sources in the hierarchical HDATA format if they include a metadata source.
- Types are instantiated automatically by using information from the metadata.

- When survey data is imported into SPSS Modeler, questions are rendered as fields, with a record for each respondent.

Data Collection Import File Options

The File tab in the Data Collection node enables you to specify options for the metadata and case data you want to import.

Metadata Settings

Note: To see the full list of available provider file types, you need to install the Survey Reporter Developer Kit, available with your Data Collection software product.

Metadata Provider. Survey data can be imported from a number of formats as supported by your Data Collection Survey Reporter Developer Kit. Available provider types include the following:

- **DataCollectionMDD.** Reads metadata from a questionnaire definition file (*.mdd*). This is the standard Data Collection Data Model format.
- **ADO Database.** Reads case data and metadata from ADO files. Specify the name and location of the *.adoinfo* file that contains the metadata. The internal name of this DSC is *mrADODsc*.
- **In2data Database.** Reads In2data case data and metadata. The internal name of this DSC is *mrI2dDsc*.
- **Data Collection Log File.** Reads metadata from a standard Data Collection log file. Typically, log files have a *.tmp* filename extension. However, some log files may have another filename extension. If necessary, you can rename the file so that it has a *.tmp* filename extension. The internal name of this DSC is *mrLogDsc*.
- **Quancept Definitions File.** Converts metadata to Quancept script. Specify the name of the Quancept *.qdi* file. The internal name of this DSC is *mrQdiDrsDsc*.
- **Quanvert Database.** Reads Quanvert case data and metadata. Specify the name and location of the *.qvinfo* or *.pkd* file. The internal name of this DSC is *mrQvDsc*.
- **Data Collection Participation Database.** Reads a project's Sample and History Table tables and creates derived categorical variables corresponding to the columns in those tables. The internal name of this DSC is *mrSampleReportingMDSC*.
- **Statistics File.** Reads case data and metadata from an IBM SPSS Statistics *.sav* file. Writes case data to an IBM SPSS Statistics *.sav* file for analysis in IBM SPSS Statistics. Writes metadata from an IBM SPSS Statistics *.sav* file to an *.mdd* file. The internal name of this DSC is *mrSavDsc*.
- **Surveycraft File.** Reads SurveyCraft case data and metadata. Specify the name of the SurveyCraft *.vq* file. The internal name of this DSC is *mrSCDsc*.
- **Data Collection Scripting File.** Reads from metadata in an mrScriptMetadata file. Typically, these files have an *.mdd* or *.dms* filename extension. The internal name of this DSC is *mrScriptMDSC*.
- **Triple-S XML File.** Reads metadata from a Triple-S file in XML format. The internal name of this DSC is *mrTripleSDsc*.

Metadata properties. Optionally, select **Properties** to specify the survey version to import as well as the language, context, and label type to use. See the topic " Data Collection Import Metadata Properties" on page 32 for more information.

Case Data Settings

Note: To see the full list of available provider file types, you need to install the Survey Reporter Developer Kit, available with your Data Collection software product.

Get Case Data Settings. When reading metadata from *.mdd* files only, click **Get Case Data Settings** to determine what case data sources are associated with the selected metadata, along with the specific settings needed to access a given source. This option is available only for *.mdd* files.

Case Data Provider. The following provider types are supported:

- **ADO Database.** Reads case data using the Microsoft ADO interface. Select OLE-DB UDL for the case data type, and specify a connection string in the Case Data UDL field. See the topic "Database Connection String" on page 32 for more information. The internal name of this component is *mrADODsc*.
- **Delimited Text File (Excel).** Reads case data from a comma-delimited (.CSV) file, such as can be output by Excel. The internal name is *mrCsvDsc*.
- **Data Collection Data File.** Reads case data from a native Data Collection Data Format file. The internal name is *mrDataFileDsc*.
- **In2data Database.** Reads case data and metadata from an In2data database (*.i2d*) file. The internal name is *mrI2dDsc*.
- **Data Collection Log File.** Reads case data from a standard Data Collection log file. Typically, log files have a *.tmp* filename extension. However, some log files may have another filename extension. If necessary, you can rename the file so that it has a *.tmp* filename extension. The internal name is *mrLogDsc*.
- **Quantum Data File.** Reads case data from any Quantum-format ASCII file (*.dat*). The internal name is *mrPunchDsc*.
- **Quancept Data File.** Reads case data from a Quancept *.drs*, *.drz*, or *.dru* file. The internal name is *mrQdiDrsDsc*.
- **Quanvert Database.** Reads case data from a Quanvert *qvinfo* or *.pkd* file. The internal name is *mrQvDsc*.
- **Data Collection Database (MS SQL Server).** Reads case data to a relational Microsoft SQL Server database. See the topic "Database Connection String" on page 32 for more information. The internal name is *mrRdbDsc2*.
- **Statistics File.** Reads case data from an IBM SPSS Statistics *.sav* file. The internal name is *mrSavDsc*.
- **Surveycraft File.** Reads case data from a SurveyCraft *.qdt* file. Both the *.vq* and *.qdt* files must be in the same directory, with read and write access for both files. This is not how they are created by default when using SurveyCraft, so one of the files needs to be moved to import SurveyCraft data. The internal name is *mrScDsc*.
- **Triple-S Data File.** Reads case data from a Triple-S data file, in either fixed-length or comma-delimited format. The internal name is *mr TripleDsc*.
- **Data Collection XML.** Reads case data from a Data Collection XML data file. Typically, this format may be used to transfer case data from one location to another. The internal name is *mrXmlDsc*.

Case Data Type. Specifies whether case data is read from a file, folder, OLE-DB UDL, or ODBC DSN, and updates the dialog box options accordingly. Valid options depend on the type of provider. For database providers, you can specify options for the OLE-DB or ODBC connection. See the topic "Database Connection String" on page 32 for more information.

Case Data Project. When reading case data from an Data Collection database, you can enter the name of the project. For all other case data types, this setting should be left blank.

Variable Import

Import System Variables. Specifies whether system variables are imported, including variables that indicate interview status (in progress, completed, finish date, and so on). You can choose **None**, **All**, or **Common**.

Import "Codes" Variables. Controls import of variables that represent codes used for open-ended "Other" responses for categorical variables.

Import "SourceFile" Variables. Controls import of variables that contain filenames of images of scanned responses.

Import multi-response variables as. Multiple response variables can be imported as multiple flag fields (a multiple dichotomy set), which is the default method for new streams. Streams created in releases of IBM SPSS Modeler prior to 12.0 imported multiple responses into a single field, with values separate by commas. The older method is still supported to allow existing streams to run as they did previously, but

updating older streams to use the new method is recommended. See the topic "Importing Multiple Response Sets" on page 32 for more information.

Data Collection Import Metadata Properties

When importing Data Collection survey data, in the Metadata Properties dialog box, you can specify the survey version to import as well as the language, context, and label type to use. Note that only one language, context, and label type can be imported at a time.

Version. Each survey version can be regarded as a snapshot of the metadata used to collect a particular set of case data. As a questionnaire undergoes changes, multiple versions may be created. You can import the latest version, all versions, or a specific version.

- **All versions.** Select this option if you want to use a combination (superset) of all of the available versions. (This is sometimes called a superversion). When there is a conflict between the versions, the most recent versions generally take precedence over the older versions. For example, if a category label differs in any of the versions, the text in the latest version will be used.
- **Latest version.** Select this option if you want to use the most recent version.
- **Specify version.** Select this option if you want to use a particular survey version.

Choosing all versions is useful when, for example, you want to export case data for more than one version and there have been changes to the variable and category definitions that mean that case data collected with one version is not valid in another version. Selecting all of the versions for which you want to export the case data means that generally you can export the case data collected with the different versions at the same time without encountering validity errors due to the differences between the versions. However, depending on the version changes, some validity errors may still be encountered.

Language. Questions and associated text can be stored in multiple languages in the metadata. You can use the default language for the survey or specify a particular language. If an item is unavailable in the specified language, the default is used.

Context. Select the user context you want to use. The user context controls which texts are displayed. For example, select **Question** to display question texts or **Analysis** to display shorter texts suitable for displaying when analyzing the data.

Label type. Lists the types of labels that have been defined. The default is **label**, which is used for question texts in the Question user context and variable descriptions in the Analysis user context. Other label types can be defined for instructions, descriptions, and so forth.

Database Connection String

When using the Data Collection node to import case data from a database via an OLE-DB or ODBC, select **Edit** from the File tab to access the Connection String dialog box, which enables you to customize the connection string passed to the provider in order to fine-tune the connection.

Advanced Properties

When using the Data Collection node to import case data from a database that requires an explicit login, select **Advanced** to provide a user ID and password to access the data source.

Importing Multiple Response Sets

Multiple response variables can be imported from Data Collection as multiple dichotomy sets, with a separate flag field for each possible value of the variable. For example, if respondents are asked to select which museums they have visited from a list, the set would include a separate flag field for each museum listed.

After importing the data, you can add or edit multiple response sets from any node that includes a Filter tab. See the topic "Editing Multiple Response Sets" on page 135 for more information.

Importing multiple responses into a single field (for streams created in previous releases)

In older releases of SPSS Modeler, rather than import multiple responses as described above, they were imported into a single field, with values separate by commas. This method is still supported in order to support for existing streams, but it is recommended that any such streams be updated to use the new method.

Data Collection Column Import Notes

Columns from the Data Collection data are read into SPSS Modeler as summarized in the following table.

Table 4. Data Collection column import summary

Data Collection Column Type	SPSS Modeler Storage	Measurement Level
Boolean flag (yes/no)	String	Flag (values 0 and 1)
Categorical	String	Nominal
Date or time stamp	Timestamp	Continuous
Double (floating point value within a specified range)	Real	Continuous
Long (integer value within a specified range)	Integer	Continuous
Text (free text description)	String	Typeless
Level (indicates grids or loops within a question)	Doesn't occur in VDATA and is not imported into SPSS Modeler	
Object (binary data such as a facsimile showing scribbled text or a voice recording)	Not imported into SPSS Modeler	
None (unknown type)	Not imported into SPSS Modeler	
Respondent.Serial column (associates a unique ID with each respondent)	Integer	Typeless

To avoid possible inconsistencies between value labels read from metadata and actual values, all metadata values are converted to lower case. For example, the value label *E1720_years* would be converted to *e1720_years*.

IBM Cognos source node

The IBM Cognos source node enables you to bring Cognos database data or single list reports into your data mining session. In this way, you can combine the business intelligence features of Cognos with the predictive analytics capabilities of IBM SPSS Modeler. You can import relational, dimensionally-modeled relational (DMR), and OLAP data.

From a Cognos server connection, you first select a location from which to import the data or reports. A location contains a Cognos model and all of the folders, queries, reports, views, shortcuts, URLs, and job definitions associated with that model. A Cognos model defines business rules, data descriptions, data relationships, business dimensions and hierarchies, and other administrative tasks.

If you are importing data, you then select the objects that you want to import from the selected package. Objects that you can import include query subjects (which represent database tables) or individual query items (which represent table columns). See “Cognos object icons” on page 34 for more information.

If the package has filters defined, you can import one or more of these. If a filter that you import is associated with imported data, that filter is applied before the data is imported. The data to be imported must be in UTF-8 format.

If you are importing a report, you select a package, or a folder within a package, containing one or more reports. You then select the individual report you want to import. Only single list reports can be imported; multiple lists are not supported.

If parameters have been defined, either for a data object or a report, you can specify values for these parameters before importing the object or report.

Note: The Cognos source node only supports Cognos CQM packages. DQM packages are not supported.

Cognos object icons

The various types of objects you can import from a Cognos Analytics database are represented by different icons, as the following table illustrates.

Table 5. Cognos object icons

Icon	Object
	Package
	Namespace
	Query subject
	Query item
	Measure dimension
	Measure
	Dimension
	Level hierarchy
	Level
	Filter
	Report
	Standalone calculation

Importing Cognos data

To import data from an IBM Cognos Analytics database, on the Data tab of the IBM Cognos dialog box ensure that **Mode** is set to **Data**.

Connection. Click **Edit** to display a dialog box where you can define the details of a new Cognos connection from which to import data or reports. If you are already logged in to a Cognos server through IBM SPSS Modeler, you can also edit the details of the current connection. See “Cognos connections” on page 36 for more information.

Location. When you have established the Cognos server connection, click **Edit** next to this field to display a list of available packages from which to import content. See “Cognos location selection” on page 36 for more information.

Content. Displays the name of the selected package, together with the namespaces associated with the package. Double-click a namespace to display the objects that you can import. The various object types are denoted by different icons. See "Cognos object icons" on page 34 for more information.

To choose an object to import, select the object and click the upper of the two right arrows to move the object into the **Fields to import** pane. Selecting a query subject imports all of its query items. Double-clicking a query subject expands it so that you can choose one or more of its individual query items. You can perform multiple selections with Ctrl-click (select individual items), Shift-click (select a block of items) and Ctrl-A (select all items).

To choose a filter to apply (if the package has filters defined), navigate to the filter in the Content pane, select the filter and click the lower of the two right arrows to move the filter into the **Filters to apply** pane. You can perform multiple selections with Ctrl-click (select individual filters) and Shift-click (select a block of filters).

Fields to import. Lists the database objects that you have chosen to import into IBM SPSS Modeler for processing. If you no longer require a particular object, select it and click the left arrow to return it to the **Content** pane. You can perform multiple selections in the same way as for **Content**.

Filters to apply. Lists the filters that you have chosen to apply to the data before it is imported. If you no longer require a particular filter, select it and click the left arrow to return it to the **Content** pane. You can perform multiple selections in the same way as for **Content**.

Parameters. If this button is enabled, the selected object has parameters defined. You can use parameters to make adjustments (for example, perform a parameterized calculation) before importing the data. If parameters are defined but no default is provided, the button displays a warning triangle. Click the button to display the parameters and optionally edit them. If the button is disabled, the report has no parameters defined.

Aggregate data before importing. Check this box if you want to import aggregated data rather than raw data.

Importing Cognos reports

To import a predefined report from an IBM Cognos database, on the Data tab of the IBM Cognos dialog box ensure that **Mode** is set to **Report**. Only single list reports can be imported; multiple lists are not supported.

Connection. Click **Edit** to display a dialog box where you can define the details of a new Cognos connection from which to import data or reports. If you are already logged in to a Cognos server through IBM SPSS Modeler, you can also edit the details of the current connection. See "Cognos connections" on page 36 for more information.

Location. When you have established the Cognos server connection, click **Edit** next to this field to display a list of available packages from which to import content. See "Cognos location selection" on page 36 for more information.

Content. Displays the name of the selected package or folder that contains reports. Navigate to a specific report, select it, and click the right arrow to bring the report into the **Report to import** field.

Report to import. Indicates the report that you have chosen to import into IBM SPSS Modeler. If you no longer require the report, select it and click the left arrow to return it to the **Content** pane, or bring a different report into this field.

Parameters. If this button is enabled, the selected report has parameters defined. You can use parameters to make adjustments before importing the report (for example, specifying a start and end date for report data). If parameters are defined but no default is provided, the button displays a warning triangle. Click the button to display the parameters and optionally edit them. If the button is disabled, the report has no parameters defined.

Cognos connections

In the Cognos Connections dialog box, you can select the Cognos Analytics server from which to import or export database objects.

Cognos server URL Type the URL of the Cognos Analytics server from which to import or export. This is the value of the "External dispatcher URI" environment property of IBM Cognos Configuration on the Cognos server. Contact your Cognos system administrator if you are not sure which URL to use.

Mode Select **Set Credentials** if you want to log in with a specific Cognos namespace, username, and password (for example, as an administrator). Select **Use Anonymous connection** to log in with no user credentials, in which case you do not complete the other fields.

Alternatively, if you have an IBM Cognos credential that is stored in the IBM SPSS Collaboration and Deployment Services Repository, you can use this credential instead of entering user name and password information, or creating an anonymous connection. To use an existing credential, select **Stored Credentials** and either enter the **Credential Name** or browse for it.

The Cognos Namespace is modeled by a domain in IBM SPSS Collaboration and Deployment Services.

Namespace ID Specify the Cognos security authentication provider that is used to log on to the server. The authentication provider is used to define and maintain users, groups, and roles, and to control the authentication process. Note this is the Namespace ID, not the Namespace Name (the ID is not always the same as the Name).

User name Enter the Cognos user name with which to log on to the server.

Password Enter the password that is associated with the specified user name.

Save as Default Click this button to store these settings as your default, to avoid having to reenter them every time you open the node.

Cognos location selection

The Specify Location dialog box enables you to select a Cognos package from which to import data, or a package or folder from which to import reports.

Public Folders. If you are importing data, this lists the packages and folders available from the chosen server. Select the package you want to use and click **OK**. You can select only one package per Cognos source node.

If you are importing reports, this lists the folders and packages containing reports that are available from the chosen server. Select a package or report folder and click **OK**. You can select only one package or report folder per Cognos source node, though report folders can contain other report folders as well as individual reports.

Specifying parameters for data or reports

If parameters have been defined in Cognos Analytics, either for a data object or a report, you can specify values for these parameters before importing the object or report. An example of parameters for a report would be start and end dates for the report contents.

Name. The parameter name as it is specified in the Cognos database.

Type. A description of the parameter.

Value. The value to assign to the parameter. To enter or edit a value, double-click its cell in the table. Values are not validated here, so any invalid values are detected at run time.

Automatically remove invalid parameters from table. This option is selected by default and will remove any invalid parameters found within the data object or report.

IBM Cognos TM1 Source Node

The IBM Cognos TM1 source node enables you to bring Cognos TM1 data into your data mining session. In this way, you can combine the enterprise planning features of Cognos with the predictive analytics capabilities of IBM SPSS Modeler. You can import a flattened version of the multidimensional OLAP cube data.

Note: The TM1 user needs the following permissions: Write privilege of cubes, Read privilege of dimensions, and Write privilege of dimension elements. In addition, IBM Cognos TM1 10.2 Fix Pack 3, or later, is required before SPSS Modeler can import and export Cognos TM1 data. Existing streams that were based on previous versions will still function.

Administrator credentials are not required for this node. But if you are still using the old legacy pre-17.1 TM1 node, administrator credentials are still required.

SPSS Modeler only supports working with Cognos TM1 via IntegratedSecurityMode 1, 4, and 5.

You need to modify the data in TM1 before the data is imported. The data to be imported must be in UTF-8 format.

From an IBM Cognos TM1 administration host connection, you first select a TM1 server from which to import the data; a server contains one or more TM1 cubes. You then select the required cube and within the cube you select the columns and rows you want to import.

Note: Before you can use the TM1 Source or Export nodes in SPSS Modeler, you must verify some settings in the `tm1s.cfg` file; this is the TM1 server configuration file in the root directory of the TM1 server.

- HTTPPortNumber - set a valid port number; typically 1-65535. Note that this is not the port number you subsequently specify in the connection in the node; it is an internal port used by TM1 that is disabled by default. If necessary, contact your TM1 administrator to confirm the valid setting for this port.
- UseSSL - if you set this to *True*, HTTPS is used as the transport protocol. In this case you must import the TM1 certification to the SPSS Modeler Server JRE.

Importing IBM Cognos TM1 data

To import data from an IBM Cognos TM1 database, on the Data tab of the IBM Cognos TM1 dialog box select the relevant TM1 administration host and associated server, cube, and data details.

Note: Before importing data you must carry out some preprocessing within TM1 to ensure the data are in a format that is recognizable to IBM SPSS Modeler. This involves filtering your data using the Subset Editor to get the view into the correct size and shape for import.

Note that zero (0) values imported from TM1 will be treated as "null" values (TM1 does not distinguish between blank and zero values). Also, note that non-numeric data (or metadata) from *regular dimensions* can be imported into IBM SPSS Modeler. But importing non-numeric *measures* is not currently supported.

Admin Host. Type the URL of the administration host where the TM1 server you want to connect to is installed. The administration host is defined as a single URL for all TM1 servers. From this URL, all IBM Cognos TM1 servers installed and running on your environment can be discovered and accessed.

TM1 Server. When you have established the connection to the administration host, select the server that contains the data you want to import and click **Login**. If you have not previously connected to this server you are prompted to enter your **User name** and **Password**; alternatively, you can search for previously entered login details that you have saved as a **Stored Credential**.

Select a TM1 cube view to import. Displays the name of the cubes within the TM1 server from which you can import data. Double-click a cube to display the view data you can import.

Note:

Only cubes with a dimension can be imported into IBM SPSS Modeler.

If an alias has been defined for an element in your TM1 cube (for example, if a value of 23277 has an alias of `Sales`), the value will be imported – not the alias.

To choose data to import, select the view and click the right arrow to move it into the **View to import** pane. If the view you require isn't visible, double-click a cube to expand its list of views. You can select a public or a private view.

Row dimension(s). Lists the name of the row dimension in the data that you have chosen to import. Scroll through the list of levels and select the one you require.

Column dimension. Lists the name of the column dimension in the data that you have chosen to import. Scroll through the list of levels and select the one you require.

Context dimension(s). Display only. Shows the context dimensions that relate to the selected columns and rows.

TWC source node

The TWC source node imports weather data from The Weather Company, an IBM Business. You can use it to obtain historical or forecast weather data for a location. This can help you develop weather-driven business solutions for better decision-making using the most accurate and precise weather data available.

With this node, you can input weather-related data such as `latitude`, `longitude`, `time`, `day_ind` (indicates night or day), `temp`, `dewpt` (dew point), `rh` (relative humidity), `feels_like` temperature, `heat_index`, `wc` (wind chill), `wx_phrase` (mostly cloudy, partly cloudy, etc), `pressure`, `clds` (clouds), `vis` (visibility), `wspd` (wind speed), `gust`, `wdir` (wind direction), `wdir_cardinal` (NW, NNW, N, etc), `uv_index` (ultraviolet index), and `uv_desc` (low, high, etc).

The TWC source node uses the following APIs:

- TWC Historical Observations Airport (http://goo.gl/DplOKj) for historical weather data
- TWC Hourly Forecast (http://goo.gl/IJhhvZ) for forecast weather data

Location

Latitude. Enter the latitude value of the location you wish to obtain weather data for, in the format `[-90.0~90.0]`.

Longitude. Enter the longitude value of the location you wish to obtain weather data for, in the format `[-180.0~180.0]`.

Misc

License Key. A license key is required. Enter the license key you obtained from The Weather Company. If you don't have a key, contact your administrator or your IBM representative.

Instead of issuing the key to all users, your administrator may have instead specified the key in a new `config.cfg` file on the IBM SPSS Modeler Server, in which case you can leave this field blank. If specified in both locations, the key in this dialog takes precedence. Note to administrators: To add the license key on the server, create a new file called `config.cfg` with the contents `LicenseKey=<LICENCEKEY>` (where `<LICENSEKEY>` is the license key) at the location `<ModelerServerInstallation>\ext\bin\pasw.twcdata`.

Units. Select the measurement unit to use: **English**, **Metric**, or **Hybrid**. The default is **Metric**.

Time Format

UTC. Select UTC time format if you're importing historical weather data and you don't want SPSS Modeler to access the TWC Hourly Forecast API. Your license key only needs privileges to access the TWC Historical Observations Airport API if you select this option.

Local. Select local time format if you need SPSS Modeler to access the TWC Hourly Forecast API to convert the time from UTC time to local time. Your license key needs privileges to access both TWC APIs if you select this option.

Data Type

Historical. If you want to import historical weather data, select **Historical** and then specify a start and end date in the format YYYYMMDD (for example, 20120101 for January 1, 2012).

Forecast. If you want to import forecast weather data, select **Forecast** and then specify the hours to forecast.

SAS Source Node

This feature is available in SPSS Modeler Professional and SPSS Modeler Premium.

The SAS source node enables you to bring SAS data into your data mining session. You can import four types of files:

- SAS for Windows/OS2 (*.sd2*)
- SAS for UNIX (*.ssd*)
- SAS Transport File (*.tpt*)
- SAS version 7/8/9 (*.sas7bdat*)

When the data is imported, all variables are kept and no variable types are changed. All cases are selected.

Setting Options for the SAS Source Node

Import. Select which type of SAS file to transport. You can choose **SAS for Windows/OS2 (.sd2)**, **SAS for UNIX (.SSD)**, **SAS Transport File (.tpt)**, or **SAS Version 7/8/9 (.sas7bdat)**.

Import file. Specify the name of the file. You can enter a filename or click the ellipsis button (**...**) to browse to the file's location.

Member. Select a member to import from the SAS transport file selected above. You can enter a member name or click **Select** to browse through all members in the file.

Read user formats from a SAS data file. Select to read user formats. SAS files store data and data formats (such as variable labels) in different files. Most often, you will want to import the formats as well. If you have a large dataset, however, you may want to deselect this option to save memory.

Format file. If a format file is required, this text box is activated. You can enter a filename or click the ellipsis button (**...**) to browse to the file's location.

Variable names. Select a method of handling variable names and labels upon import from a SAS file. Metadata that you choose to include here persists throughout your work in IBM SPSS Modeler and may be exported again for use in SAS.

- **Read names and labels.** Select to read both variable names and labels into IBM SPSS Modeler. By default, this option is selected and variable names are displayed in the Type node. Labels may be displayed in the Expression Builder, charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box.
- **Read labels as names.** Select to read the descriptive variable labels from the SAS file rather than the short field names and use these labels as variable names in IBM SPSS Modeler.

Excel Source Node

The Excel source node enables you to import data from Microsoft Excel in the .xlsx file format.

File type. Select the Excel file type that you are importing.

Import file. Specifies the name and location of the spreadsheet file to import.

Use Named Range. Enables you to specify a named range of cells as defined in the Excel worksheet. Click the ellipses button (**...**) to choose from the list of available ranges. If a named range is used, other worksheet and data range settings are no longer applicable and are disabled as a result.

Choose worksheet. Specifies the worksheet to import, either by index or by name.

- **By index.** Specify the index value for the worksheet you want to import, beginning with 0 for the first worksheet, 1 for the second worksheet, and so on.
- **By name.** Specify the name of the worksheet you want to import. Click the ellipses button (**...**) to choose from the list of available worksheets.

Range on worksheet. You can import data beginning with the first non-blank row or with an explicit range of cells.

- **Range starts on first non-blank row.** Locates the first non-blank cell and uses this as the upper left corner of the data range.
- **Explicit range of cells.** Enables you to specify an explicit range by row and column. For example, to specify the Excel range `A1:D5`, you can enter `A1` in the first field and `D5` in the second (or alternatively, `R1C1` and `R5C4`). All rows in the specified range are returned, including blank rows.

On blank rows. If more than one blank row is encountered, you can choose whether to **Stop reading**, or choose **Return blank rows** to continue reading all data to the end of the worksheet, including blank rows.

First row has column names. Indicates that the first row in the specified range should be used as field (column) names. If not selected, field names are generated automatically.

Field Storage and Measurement Level

When reading values from Excel, fields with numeric storage are read in with a measurement level of *Continuous* by default, and string fields are read in as *Nominal*. You can manually change the measurement level (continuous versus nominal) on the Type tab, but the storage is determined automatically (although it can be changed using a conversion function, such as `to_integer`, in a Filler node or Derive node if necessary). See the topic "Setting Field Storage and Formatting" on page 6 for more information.

By default, fields with a mix of numeric and string values read in as numbers, which means that any string values will be set to null (system missing) values in IBM SPSS Modeler. This happens because, unlike Excel, IBM SPSS Modeler does not allow mixed storage types within a field. To avoid this, you can manually set the cell format to `Text` in the Excel spreadsheet, which causes all values (including numbers) to read in as strings.

XML Source Node

This feature is available in SPSS Modeler Professional and SPSS Modeler Premium.

Use the XML source node to import data from a file in XML format into an IBM SPSS Modeler stream. XML is a standard language for data exchange, and for many organizations it is the format of choice for this purpose. For example, a government tax agency might want to analyze data from tax returns that have been submitted online and which have their data in XML format (see http://www.w3.org/standards/xml/).

Importing XML data into an IBM SPSS Modeler stream enables you to perform a wide range of predictive analytics functions on the source. The XML data is parsed into a tabular format in which the columns correspond to the different levels of nesting of the XML elements and attributes. The XML items are displayed in XPath format (see http://www.w3.org/TR/xpath20/).

Important: The XML source node does not consider namespace declaration. So, for example, your XML files cannot contain a colon (`:`) character in the `name` tag. If they do, you'll receive errors during execution about invalid characters.

Read a single file. By default, SPSS Modeler reads a single file, which you specify in the **XML data source** field.

Read all XML files in a directory. Choose this option if you want to read all the XML files in a particular directory. Specify the location in the **Directory** field that appears. Select the **Include subdirectories** check box to additionally read XML files from all the subdirectories of the specified directory.

XML data source. Type the full path and file name of the XML source file you want to import, or use the Browse button to find the file.

XML schema. (Optional) Specify the full path and file name of an XSD or DTD file from which to read the XML structure, or use the Browse button to find this file. If you leave this field blank, the structure is read from the XML source file. An XSD or DTD file can have more than one root element. In this case, when you change the focus to a different field, a dialog is displayed where you choose the root element you want to use. See the topic "Selecting from Multiple Root Elements" on page 41 for more information.

Note: XSD Indicators are ignored by SPSS Modeler

XML structure. A hierarchical tree showing the structure of the XML source file (or the schema, if you specified one in the **XML schema** field). To define a record boundary, select an element and click the right-arrow button to copy the item to the **Records** field.

Display attributes. Displays or hides the attributes of the XML elements in the **XML structure** field.

Records (XPath expression). Shows the XPath syntax for an element copied from the XML structure field. This element is then highlighted in the XML structure, and defines the record boundary. Each time this element is encountered in the source file, a new record is created. If this field is empty, the first child element under the root is used as the record boundary.

Read all data. By default, all data in the source file is read into the stream.

Specify data to read. Choose this option if you want to import individual elements, attributes or both. Choosing this option enables the Fields table where you can specify the data you want to import.

Fields. This table lists the elements and attributes selected for import, if you have selected the **Specify data to read** option. You can either type the XPath syntax of an element or attribute directly into the XPath column, or select an element or attribute in the XML structure and click the right-arrow button to copy the item into the table. To copy all the child elements and attributes of an element, select the element in the XML structure and click the double-arrow button.

- **XPath.** The XPath syntax of the items to be imported.
- **Location.** The location in the XML structure of the items to be imported. **Fixed path** shows the path of the item relative to the element highlighted in the XML structure (or the first child element under the root, if no element is highlighted). **Any location** denotes an item of the given name at any location in the XML structure. **Custom** is displayed if you type a location directly into the XPath column.

Selecting from Multiple Root Elements

While a properly formed XML file can only have a single root element, an XSD or DTD file can contain multiple roots. If one of the roots matches that in the XML source file, that root element is used, otherwise you need to select one to use.

Choose the root to display. Select the root element you want to use. The default is the first root element in the XSD or DTD structure.

Removing Unwanted Spaces from XML Source Data

Line breaks in the XML source data may be implemented by a `[CR][LF]` character combination. In some cases these line breaks can occur in the middle of a text string, for example:

```
<description>An in-depth look at creating applications[CR][LF]
with XML.</description>
```

These line breaks may not be visible when the file is opened in some applications, for example a Web browser. However, when the data are read into the stream through the XML source node, the line breaks are converted to a series of space characters.

You can correct this by using a Filler node to remove these unwanted spaces:

Here is an example of how you can achieve this:

1. Attach a Filler node to the XML source node.
2. Open the Filler node and use the field chooser to select the field with the unwanted spaces.
3. Set **Replace** to **Based on condition** and set **Condition** to **true**.
4. In the **Replace with** field, enter `replace(" ","",@FIELD)` and click OK.
5. Attach a Table node to the Filler node and run the stream.

In the Table node output, the text now appears without the additional spaces.

User Input Node

The User Input node provides an easy way for you to create synthetic data--either from scratch or by altering existing data. This is useful, for example, when you want to create a test dataset for modeling.

Creating Data from Scratch

The User Input node is available from the Sources palette and can be added directly to the stream canvas.

1. Click the **Sources** tab of the nodes palette.
2. Drag and drop or double-click to add the User Input node to the stream canvas.
3. Double-click to open its dialog box and specify fields and values.

Note: User Input nodes that are selected from the Sources palette will be completely blank, with no fields and no data information. This enables you to create synthetic data entirely from scratch.

Generating Data from an Existing Data Source

You can also generate a User Input node from any nonterminal node in the stream:

1. Decide at which point in the stream you want to replace a node.
2. Right-click on the node that will feed its data into the User Input node and choose **Generate User Input Node** from the menu.
3. The User Input node appears with all downstream processes attached to it, replacing the existing node at that point in your data stream. When generated, the node inherits all of the data structure and field type information (if available) from the metadata.

Note: If data have not been run through all nodes in the stream, then the nodes are not fully instantiated, meaning that storage and data values may not be available when replacing with a User Input node.

Setting Options for the User Input Node

The dialog box for a User Input node contains several tools you can use to enter values and define the data structure for synthetic data. For a generated node, the table on the Data tab contains field names from the original data source. For a node added from the Sources palette, the table is blank. Using the table options, you can perform the following tasks:

- Add new fields using the Add a New Field button at the right in the table.
- Rename existing fields.
- Specify data storage for each field.
- Specify values.
- Change the order of fields on the display.

Entering Data

For each field, you can specify values or insert values from the original dataset using the value picker button to the right of the table. See the rules described below for more information on specifying values. You can also choose to leave the field blank—fields left blank are filled with the system null (`$null$`).

To specify string values, simply type them in the Values column, separated by spaces:

```
Fred Ethel Martin
```

Strings that include spaces can be wrapped in double-quotes:

```
"Bill Smith" "Fred Martin" "Jack Jones"
```

For numeric fields, you can either enter multiple values in the same manner (listed with spaces between):

```
10 12 14 16 18 20
```

Or you can specify the same series of numbers by setting its limits (10, 20) and the steps in between (2). Using this method, you would type:

```
10,20,2
```

These two methods can be combined by embedding one within the other, such as:

```
1 5 7 10,20,2 21 23
```

This entry will produce the following values:

```
1 5 7 10 12 14 16 18 20 21 23
```

Date and time values can be entered using the current default format selected in the Stream Properties dialog box, for example:

```
11:04:00 11:05:00 11:06:00
```

```
2007-03-14 2007-03-15 2007-03-16
```

For timestamp values, which have both a date and time component, double-quotes must be used:

```
"2007-03-14 11:04:00" "2007-03-14 11:05:00" "2007-03-14 11:06:00"
```

For additional details see comments on data storage below.

Generate data. Enables you to specify how the records are generated when you run the stream.

- **All combinations**. Generates records containing every possible combination of the field values, so each field value will appear in several records. This can sometimes generate more data than is wanted, so often you might follow this node with a sample node.
- **In order**. Generates records in the order in which the data field values are specified. Each field value only appears in one record. The total number of records is equal to the largest number of values for a single field. Where fields have fewer than the largest number, undefined ($null$) values are inserted.

Show example

For example, the following entries will generate the records listed in the two following table examples.

- **Age.** `30,60,10`
- **BP.** `LOW`
- **Cholesterol.** `NORMAL HIGH`
- **Drug.** (left blank)

Table 6. Generate data field set to All combinations

Age	BP	Cholesterol	Drug
30	LOW	NORMAL	$null$
30	LOW	HIGH	$null$
40	LOW	NORMAL	$null$

Table 6. Generate data field set to All combinations (continued)			
Age	**BP**	**Cholesterol**	**Drug**
40	LOW	HIGH	$null$
50	LOW	NORMAL	$null$
50	LOW	HIGH	$null$
60	LOW	NORMAL	$null$
60	LOW	HIGH	$null$

Table 7. Generate data field set to In order			
Age	**BP**	**Cholesterol**	**Drug**
30	LOW	NORMAL	$null$
40	$null$	HIGH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Data Storage

Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*. While storage must be determined at the source, measurement level can be changed using a Type node at any point in the stream. See the topic "Measurement levels" on page 122 for more information.

Available storage types are:

- **String** Used for fields that contain non-numeric data, also called alphanumeric data. A string can include any sequence of characters, such as *fred*, *Class 2*, or *1234*. Note that numbers in strings cannot be used in calculations.
- **Integer** A field whose values are integers.
- **Real** Values are numbers that may include decimals (not limited to integers). The display format is specified in the Stream Properties dialog box and can be overridden for individual fields in a Type node (Format tab).
- **Date** Date values specified in a standard format such as year, month, and day (for example, `2007-09-26`). The specific format is specified in the Stream Properties dialog box.
- **Time** Time measured as a duration. For example, a service call lasting 1 hour, 26 minutes, and 38 seconds might be represented as `01:26:38`, depending on the current time format as specified in the Stream Properties dialog box.
- **Timestamp** Values that include both a date and time component, for example `2007–09–26 09:04:00`, again depending on the current date and time formats in the Stream Properties dialog box. Note that timestamp values may need to be wrapped in double-quotes to ensure they are interpreted as a single value rather than separate date and time values. (This applies for example when entering values in a User Input node.)
- **List** Introduced in SPSS Modeler version 17, along with new measurement levels of Geospatial and Collection, a List storage field contains multiple values for a single record. There are list versions of all of the other storage types.

Table 8. List storage type icons	
Icon	**Storage type**
[A]	List of string
[1]	List of integer
[#]	List of real
[clock]	List of time
[cal]	List of date
[cal]	List of timestamp
[[]]	List with a depth greater than zero

In addition, for use with the Collection measurement level, there are list versions of the following measurement levels.

Table 9. List measurement level icons	
Icon	**Measurement level**
[ruler]	List of continuous
[cat]	List of categorical
[flag]	List of flags
[nom]	List of nominal
[ord]	List of ordinal

Lists can be imported into SPSS Modeler in one of three source nodes (Analytic Server, Geospatial, or Variable File), or created within your streams through use of the Derive or Filler field operation nodes.

For more information on Lists and their interaction with the Collection and Geospatial measurement levels, see "List storage and associated measurement levels" on page 8

Storage conversions. You can convert storage for a field using a variety of conversion functions, such as `to_string` and `to_integer`, in a Filler node. See the topic "Storage Conversion Using the Filler Node" on page 143 for more information. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the Stream Properties dialog box. For example, if you want to convert a string field with values *Jan 2018*, *Feb 2018*, (and so forth) to date storage, select **MON YYYY** as the default date format for the stream. Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node to perform other manipulations, such as recoding string fields with categorical values. See the topic "Recoding Values with the Derive Node" on page 142 for more information.

Reading in mixed data. Note that when reading in fields with numeric storage (either integer, real, time, timestamp, or date), any non-numeric values are set to null or system missing. This is because unlike some applications, IBM SPSS Modeler does not allow mixed storage types within a field. To avoid this, any fields with mixed data should be read in as strings, either by changing the storage type in the source node or in the external application as necessary.

Note: Generated User Input nodes may already contain storage information garnered from the source node if instantiated. An uninstantiated node does not contain storage or usage type information.

Rules for Specifying Values

For symbolic fields, you should leave spaces between multiple values, such as:

```
HIGH MEDIUM LOW
```

For numeric fields, you can either enter multiple values in the same manner (listed with spaces between):

```
10 12 14 16 18 20
```

Or you can specify the same series of numbers by setting its limits (10, 20) and the steps in between (2). Using this method, you would type:

```
10,20,2
```

These two methods can be combined by embedding one within the other, such as:

```
1 5 7 10,20,2 21 23
```

This entry will produce the following values:

```
1 5 7 10 12 14 16 18 20 21 23
```

Simulation Generate Node

The Simulation Generate node provides an easy way to generate simulated data, either without historical data using user specified statistical distributions, or automatically using the distributions obtained from running a Simulation Fitting node on existing historical data. Generating simulated data is useful when you want to evaluate the outcome of a predictive model in the presence of uncertainty in the model inputs.

Creating data without historical data

The Simulation Generate node is available from the Sources palette and can be added directly to the stream canvas.

1. Click the **Sources** tab of the nodes palette.
2. Drag and drop or double-click to add the Simulation Generate node to the stream canvas.
3. Double-click to open its dialog box and specify fields, storage types, statistical distributions and distribution parameters.

Note: Simulation Generate nodes that are selected from the Sources palette will be completely blank, with no fields and no distribution information. This enables you to entirely create simulated data without historical data.

Generating simulated data using existing historical data

A Simulation Generate node can also be created by executing a Simulation Fitting terminal node:

1. Right-click on the Simulation Fitting node and choose **Run** from the menu.
2. The Simulation Generate node appears on the stream canvas with an update link to the Simulation Fitting node.
3. When generated, the Simulation Generate node inherits all of the fields, storage types, and statistical distribution information from the Simulation Fitting node.

Defining an update link to a simulation fitting node

You can create a link between a Simulation Generate node and a Simulation Fitting node. This is useful if you want to update one or more fields with the information of the best fitting distribution, determined by fitting to historical data.

1. Right click on the Simulation Generate node.
2. From the menu, select **Define Update Link**. The cursor changes to the Link cursor.
3. Click on another node. If this node is a Simulation Fitting node, a link is made. If this node is not a Simulation Fitting node, no link is made, and the cursor changes back to a normal cursor.

If the fields in the Simulation Fitting node are different to those in the Simulation Generate node, a message is displayed informing you that there is a difference.

When the Simulation Fitting node is used to update the linked Simulation Generate node, the result depends on whether the same fields are present in both nodes, and if the fields are unlocked in the

Simulation Generate node. The results of updating a Simulation Fitting node are shown in the following table.

Table 10. Results of updating a Simulation Fitting node

	Field in Simulation Fitting node	
Field in Simulation Generate node	**Present**	**Missing**
Present and unlocked.	Field is overwritten.	Field is deleted.
Missing.	Field is added.	No change.
Present and locked.	The distribution of the field is not overwritten. The information in the **Fit Details** dialog box and the correlations are updated.	The field is not overwritten. The correlations are set to zero.
Do not clear Min/Max when refitting check box is selected.	The field is overwritten, apart from	the values in the Min, Max column.
Do not recalculate correlations when refitting check box is selected.	If the field is unlocked, it is overwritten.	The correlations are not overwritten.

Removing an update link to a simulation fitting node

You can remove a link between a Simulation Generate node and a Simulation Fitting node by taking the following steps:

1. Right click on the Simulation Generate node.
2. From the menu, select **Remove Update Link**. The link is removed.

Setting Options for the Simulation Generate Node

You can use the options on the Data tab of the Simulation Generate node dialog box to do the following:

- View, specify, and edit the statistical distribution information for the fields.
- View, specify, and edit the correlations between the fields.
- Specify the number of iterations and cases to simulate.

Select an item. Enables you to switch between the three views of the Simulation Generate node: Simulated Fields, Correlations, and Advanced Options.

Simulated Fields view

If the Simulation Generate node has been generated or updated from a Simulation Fitting node using historical data, in the Simulated Fields view you can view and edit the statistical distribution information for each field. The following information about each field is copied to the **Types** tab of the Simulation Generate node from the Simulation Fitting node:

- Measurement level
- Values
- Missing
- Check
- Role

If you do not have historical data, you can define fields and specify their distributions by selecting a storage type, and selecting a distribution type and entering the required parameters. Generating data in

this way means that information about the measurement level of each field will not be available until the data are instantiated, for example, on the **Types** tab or in a Type node.

The Simulated Fields view contains several tools, which you can use to perform the following tasks:

- Add and remove fields.
- Change the order of fields on the display.
- Specify a storage type for each field.
- Specify a statistical distribution for each field.
- Specify parameter values for the statistical distribution of each field.

Simulated Fields. This table contains one empty row if the Simulation Generate node has been added to the stream canvas from the Sources palette. When this row is edited, a new empty row is added to the bottom of the table. If the Simulation Generate node has been created from a Simulation Fitting node, this table will contain one row for each field of the historical data. Extra rows can be added to the table by clicking the **Add new field** icon.

The Simulated Fields table is made up of the following columns:

- **Field.** Contains the names of the fields. The field names can be edited by typing in the cells.
- **Storage.** The cells in this column contain a drop-down list of storage types. Available storage types are **String**, **Integer**, **Real**, **Time**, **Date**, and **Timestamp**. The choice of storage type determines which distributions are available in the Distribution column. If the Simulation Generate node has been created from a Simulation Fitting node, the storage type is copied over from the Simulation Fitting node.

  Note: For fields with datetime storage types, you must specify the distribution parameters as integers. For example, to specify a mean date of 1 January 1970, use the integer 0. The signed integer represents the number of seconds since (or before) midnight on 1 January 1970.
- **Status.** Icons in the Status column indicate the fit status for each field.

No distribution has been specified for the field or one or more distribution parameter is missing. In order to run the simulation, you must specify a distribution for this field and enter valid values for the parameters.

The field is set to the closest fitting distribution.

Note: This icon can only ever be displayed if the Simulation Generate node is created from a Simulation Fitting node.

The closest fitting distribution has been replaced with an alternative distribution from the **Fit Details** sub-dialog box. See the topic "Fit Details" on page 52 for more information.

The distribution has been manually specified or edited, and might include a parameter specified at more than one level.

- **Locked.** Locking a simulated field, by selecting the check box in the column with the lock icon, excludes the field from automatic updating by a linked Simulation Fitting node. This is most useful when you manually specify a distribution and want to ensure that it will not be affected by automatic distribution fitting when a linked Simulation Fitting node is executed.
- **Distribution.** The cells in this column contain a drop-down list of statistical distributions. The choice of storage type determines which distributions are available in this column for a given field. See the topic "Distributions" on page 55 for more information.

  Note: You cannot specify the Fixed distribution for every field. If you want every field in your generated data to be fixed, you can use a User Input node followed by a Balance node.
- **Parameters.** The distribution parameters associated with each fitted distribution are displayed in this column. Multiple values of a parameter are comma separated. Specifying multiple values for a parameter generates multiple iterations for the simulation. See the topic "Iterations" on page 55 for more information. If parameters are missing, this is reflected in the icon displayed in the Status column. To specify values for the parameters, click this column in the row corresponding to the field of interest and choose **Specify** from the list. This opens the **Specify Parameters** sub-dialog box. See the topic "Specify Parameters" on page 53 for more information. This column is disabled if Empirical is chosen in the Distribution column.
- **Min, Max.** In this column, for some distributions, you can specify a minimum value, a maximum value, or both for the simulated data. Simulated data that are smaller than the minimum value and larger than the maximum value will be rejected, even though they would be valid for the specified distribution. To specify minimum and maximum values, click this column in the row corresponding to the field of interest and choose **Specify** from the list. This opens the **Specify Parameters** sub-dialog box. See the topic "Specify Parameters" on page 53 for more information. This column is disabled if Empirical is chosen in the Distribution column.

Use Closest Fit. Only enabled if the Simulation Generate node has been created automatically from a Simulation Fitting node using historical data, and a single row in the Simulated Fields table is selected. Replaces the information for the field in the selected row with the information of the best fitting distribution for the field. If the information in the selected row has been edited, pressing this button will reset the information back to the best fitting distribution determined from the Simulation Fitting node.

Fit Details. Only enabled if the Simulation Generate node has been created automatically from a Simulation Fitting node. Opens the **Fit Details** sub-dialog box. See the topic "Fit Details" on page 52 for more information.

Several useful tasks can be carried out using the icons on the right of the Simulated Fields view. These icons are described in the following table.

Table 11. Icons on the Simulated Fields view

Icon	Tooltip	Description
	Edit distribution parameters	Only enabled when a single row in the Simulated Fields table is selected. Opens the **Specify Parameters** sub-dialog box for the selected row. See the topic "Specify Parameters" on page 53 for more information.
	Add new field	Only enabled when a single row in the Simulated Fields table is selected. Adds a new empty row to the bottom of the Simulated Fields table.

Table 11. Icons on the Simulated Fields view (continued)		
Icon	**Tooltip**	**Description**
	Create multiple copies	Only enabled when a single row in the Simulated Fields table is selected. Opens the **Clone Field** sub-dialog box. See the topic "Clone Field" on page 52 for more information.
	Delete selected field	Deletes the selected row from the Simulated Fields table.
	Move to top	Only enabled if the selected row is not already the top row of the Simulated Fields table. Moves the selected row to the top of the Simulated Fields table. This action affects the order of the fields in the simulated data.
	Move up	Only enabled if the selected row is not the top row of the Simulated Fields table. Moves the selected row up one position in the Simulated Fields table. This action affects the order of the fields in the simulated data.
	Move down	Only enabled if the selected row is not the bottom row of the Simulated Fields table. Moves the selected row down one position in the Simulated Fields table. This action affects the order of the fields in the simulated data.
	Move to bottom	Only enabled if the selected row is not already the bottom row of the Simulated Fields table. Moves the selected row to the bottom of the Simulated Fields table. This action affects the order of the fields in the simulated data.

Do not clear Min and Max when refitting. When selected, the minimum and maximum values are not overwritten when the distributions are updated by executing a connected Simulation Fitting node.

Correlations view

Input fields to predictive models are often known to be correlated--for example, height and weight. Correlations between fields that will be simulated must be accounted for in order to ensure that the simulated values preserve those correlations.

If the Simulation Generate node has been generated or updated from a Simulation Fitting node using historical data, in the Correlations view you can view and edit the calculated correlations between pairs of fields. If you do not have historical data, you can specify the correlations manually based on your knowledge of how the fields are correlated.

Note: Before any data are generated, the correlation matrix is automatically checked to see if it is positive semi-definite, and can therefore be inverted. A matrix can be inverted if its columns are linearly independent. If the correlation matrix cannot be inverted, it will be automatically adjusted to make it invertible.

You can choose to display the correlations in a matrix or list format.

Correlations matrix. Displays the correlations between pairs of fields in a matrix. The field names are listed, in alphabetical order, down the left and along the top of the matrix. Only the cells below the diagonal can be edited; a value between -1.000 and 1.000, inclusive, must be entered. The cell above the diagonal is updated when focus is changed away from its mirrored cell below the diagonal; both cells then display the same value. The diagonal cells are always disabled and always have a correlation of 1.000. The default value for all other cells is 0.000. A value of 0.000 specifies that there is no correlation between the associated pair of fields. Only continuous and ordinal fields are included in the matrix. Nominal, categorical and flag fields, and fields that are assigned the Fixed distribution are not shown in the table.

Correlations list. Displays the correlations between pairs of fields in a table. Each row of the table shows the correlation between a pair of fields. Rows cannot be added or deleted. The columns with the headings Field 1 and Field 2 contain the field names, which cannot be edited. The Correlation column contains the correlations, which can be edited; a value between -1.000 and 1.000, inclusive, must be entered. The default value for all cells is 0.000. Only continuous and ordinal fields are included in the list. Nominal, categorical and flag fields, and fields that are assigned the Fixed distribution are not shown in the list.

Reset correlations. Opens the **Reset Correlations** dialog box. If historical data is available, you can choose one of three options:

- **Fitted.** Replaces the current correlations with those calculated using the historical data.
- **Zeroes.** Replaces the current correlations with zeroes.
- **Cancel.** Closes the dialog box. The correlations are unchanged.

If historical data is not available, but you have made changes to the correlations, you can choose to replace the current correlations with zeroes, or cancel.

Show As. Select **Table** to display the correlations as a matrix. Select **List** to display the correlations as a list.

Do not recalculate correlations when refitting. Select this option if you want to manually specify correlations and prevent them from being overwritten when automatically fitting distributions using a Simulation Fitting node and historical data.

Use fitted multiway contingency table for inputs with a categorical distribution. By default, all fields with a categorical distribution are included in a contingency table (or multiway contingency table, depending on the number of fields with a categorical distribution). The contingency table is constructed, like the correlations, when a Simulation Fitting node is executed. The contingency table cannot be viewed. When this option is selected, fields with a categorical distribution are simulated using the actual percentages from the contingency table. That is, any associations between nominal fields are recreated in the new, simulated data. When this option is cleared, fields with categorical distributions are simulated using the expected percentages from the contingency table. If you modify a field, the field is removed from the contingency table.

Advanced Options view

Number of cases to simulate. Displays the options for specifying the number of cases to be simulated, and how any iterations will be named.

- **Maximum number of cases.** This specifies the maximum number of cases of simulated data, and associated target values, to generate. The default value is 100,00, minimum value is 1000, and maximum value is 2,147,483,647.
- **Iterations.** This number is calculated automatically and cannot be edited. An iteration is created automatically each time a distribution parameter has multiple values specified.

- **Total rows.** Only enabled when the number of iterations is greater than 1. The number is calculated automatically, using the equation shown, and cannot be edited.
- **Create iteration field.** Only enabled when the number of iterations is greater than 1. When selected, the **Name** field is enabled. See the topic "Iterations" on page 55 for more information.
- **Name.** Only enabled when the **Create iteration field** check box is selected, and the number of iterations is greater than 1. Edit the name of the iteration field by typing in this text field. See the topic "Iterations" on page 55 for more information.

Random seed. Setting a random seed allows you to replicate your simulation.

- **Replicate results.** When selected, the **Generate** button and **Random seed** field are enabled.
- **Random seed.** Only enabled when the **Replicate results** check box is selected. In this field you can specify an integer to be used as the random seed. The default value is 629111597.
- **Generate.** Only enabled when the **Replicate results** check box is selected. Creates a pseudo-random integer between 1 and 999999999, inclusive, in the **Random seed** field.

Clone Field

In the **Clone Field** dialog box you can specify how many copies of the selected field to create, and how each copy is named. It is useful to have multiple copies of fields when investigating compounded effects, for example interest or growth rates over a number of successive time periods.

The title bar of the dialog box contains the name of the selected field.

Number of copies to make. Contains the number of copies of the field to create. Click the arrows to select the number of copies to create. The minimum number of copies is 1 and the maximum is 512. The number of copies is initially set to 10.

Copy suffix character(s). Contains the characters that are added to the end of the field name for each copy. These characters separate the field name from the copy number. The suffix characters can be edited by typing in this field. This field can be left empty; in this case there will be no characters between the field name and copy number. The default character is an underscore.

Initial copy number. Contains the suffix number for the first copy. Click the arrows to select the initial copy number. The minimum initial copy number is 1 and the maximum is 1000. The default initial copy number is 1.

Copy number step. Contains the increment for the suffix numbers. Click the arrows to select the increment. The minimum increment is 1 and the maximum is 255. The increment is initially set to 1.

Fields. Contains a preview of the field names for the copies, which is updated if any of the fields of the **Clone Field** dialog box are edited. This text is generated automatically and cannot be edited.

OK. Generates all the copies as specified in the dialog box. The copies are added to the Simulated Fields table in the Simulation Generate node dialog box, directly below the row that contains the field that was copied.

Cancel. Closes the dialog box. Any changes that have been made are discarded.

Fit Details

The **Fit Details** dialog box is only available if the Simulation Generate node was created or updated by executing a Simulation Fitting node. It displays the results of automatic distribution fitting for the selected field. Distributions are ordered by goodness of fit, with the closest fitting distribution listed first. In this dialog box you can perform the following tasks:

- Examine the distributions fitted to the historical data.
- Select one of the fitted distributions.

Field. Contains the name of the selected field. This text cannot be edited.

Treat as (Measure). Displays the measurement type of the selected field. This is taken from the Simulated Fields table in the Simulation Generate node dialog box. The measurement type can be

changed by clicking the arrow and selecting a measurement type from the drop-down list. There are three options: **Continuous**, **Nominal**, and **Ordinal**.

Distributions. The Distributions table shows all the distributions that are appropriate for the measurement type. The distributions that have been fitted to the historical data are ordered by goodness of fit, from best fitting to worst fitting. Goodness of fit is determined by the fit statistic chosen in the Simulation Fitting node. The distributions that were not fitted to the historical data are listed in the table in alphabetical order below the distributions that were fitted.

The Distribution table contains the following columns:

- **Use.** The selected radio button indicates which distribution is currently chosen for the field. You can override the closest fitting distribution by selecting the radio button for the desired distribution in the Use column. Selecting a radio button in the Use column also displays a plot of the distribution superimposed on a histogram (or bar chart) of the historical data for the selected field. Only one distribution can be selected at a time.
- **Distribution.** Contains the name of the distribution. This column cannot be edited.
- **Fit Statistics.** Contains the calculated fit statistics for the distribution. This column cannot be edited. The contents of the cell depends on the measurement type of the field:
  - **Continuous.** Contains the results of the Anderson-Darling and Kolmogorov-Smirnoff tests. The p-values associated with the tests are also shown. The fit statistic chosen as the Goodness of fit criteria in the Simulation Fitting node is shown first and is used to order the distributions. The Anderson-Darling statistics are displayed as A=*aval* P=*pval*. The Kolmogorov-Smirnoff statistics are displayed as K=*kval* P=*pval*. If a statistic cannot be calculated, a dot is displayed instead of a number.
  - **Nominal and Ordinal.** Contains the results of the chi-squared test. The p-value associated with the test is also shown. The statistics are displayed as Chi-Sq=*val* P=*pval*. If the distribution was not fitted, `Not fitte`d is displayed. If the distribution cannot be mathematically fit, `Cannot be fit` is displayed.

  Note: The cell is always empty for the Empirical distribution.
- **Parameters.** Contains the distribution parameters associated with each fitted distribution. The parameters are displayed as *parameter_name* = *parameter_value*, with the parameters separated by a single space. For the categorical distribution, the parameter names are the categories and the parameter values are the associated probabilities. If the distribution was not fitted to the historical data, the cell is empty. This column cannot be edited.

Histogram thumbnail. Shows a plot of the selected distribution superimposed on a histogram of the historical data of the selected field.

Distribution thumbnail. Shows an explanation and illustration of the selected distribution.

OK. Closes the dialog box and updates the values of the Measurement, Distribution, Parameters and Min, Max columns of the Simulated Fields table for the selected field with the information from the selected distribution. The icon in the Status column is also updated to reflect whether the selected distribution is the distribution with the closest fit to the data.

Cancel. Closes the dialog box. Any changes that have been made are discarded.

Specify Parameters

In the **Specify Parameters** dialog box you can manually specify the parameter values for the distribution of the selected field. You can also choose a different distribution for the selected field.

The Specify Parameters dialog box can be opened in three ways:

- Double click on a field name in the Simulated Fields table in the Simulation Generate node dialog box.
- Click the Parameters or Min, Max column of the Simulated Fields table, and choose **Specify** from the list.
- In the Simulated Fields table, select a row, and then click the **Edit distribution parameters** icon.

Field. Contains the name of the selected field. This text cannot be edited.

Distribution. Contains the distribution of the selected field. This is taken from the Simulated Fields table. The distribution can be changed by clicking the arrow and selecting a distribution from the drop-down list. The available distributions depend on the storage type of the selected field.

Sides. This option is only available for the when the dice distribution is selected in the **Distribution** field. Click the arrows to select the number of sides, or categories, to split the field in to. The minimum number of sides is two, and the maximum is 20. The number of sides is initially set to 6.

Distribution parameters. The Distribution parameters table contains one row for each parameter of the chosen distribution.

Note: Distribution uses a rate parameter, with a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = 1/\theta$.

The table contains two columns:

- **Parameter.** Contains the names of the parameters. This column cannot be edited.
- **Value(s).** Contains the values of the parameters. If the Simulation Generate node has been created or updated from a Simulation Fitting node, the cells in this column contain the parameter values that have been determined by fitting the distribution to the historical data. If the Simulation Generate node has been added to the stream canvas from the Source nodes palette, the cells in this column are empty. The values can be edited by typing in the cells. See the topic "Distributions" on page 55 for more information about the parameters needed by each distribution, and acceptable parameter values.

  Multiple values for a parameter must be separated by commas. Specifying multiple values for a parameter will define multiple iterations of the simulation. You can only specify multiple values for one parameter.

  Note: For fields with datetime storage types, you must specify the distribution parameters as integers. For example, to specify a mean date of 1 January 1970, use the integer 0.

Note: When the dice distribution is selected, the Distribution parameters table is slightly different. The table contains one row for each side (or category). The table contains a Value column and a Probability column. The Value column contains a label for each category. The default values for the labels are the integers 1-N, where N is the number of sides. The labels can be edited by typing in the cells. Any value can be entered into the cells. If you want to use a value that is not a number, the storage type of the data field must be changed to string if the storage type is not already set to string. The Probability column contains the probability of each category. The probabilities cannot be edited, and are calculated as 1/N.

Preview. Shows a sample plot of the distribution, based on the specified parameters. If two or more values have been specified for one parameter, sample plots for each value of the parameter are shown. If historical data is available for the selected field, the plot of the distribution is superimposed on a histogram of the historical data.

Optional Settings. Use these options to specify a minimum value, a maximum value, or both for the simulated data. Simulated data that are smaller than the minimum value and larger than the maximum value will be rejected, even though they would be valid for the specified distribution.

- **Specify minimum.** Select to enable the **Reject values below** field. The check box is disabled if the Empirical distribution is selected.
- **Reject values below.** Only enabled if **Specify minimum** is selected. Enter a minimum value for the simulated data. Any simulated values smaller than this value will be rejected.
- **Specify maximum.** Select to enable the **Reject values above** field. The check box is disabled if the Empirical distribution is selected.
- **Reject values above.** Only enabled if **Specify maximum** is selected. Enter a maximum value for the simulated data. Any simulated values larger than this value will be rejected.

OK. Closes the dialog box and updates the values of the Distribution, Parameters and Min, Max columns of the Simulated Fields table for the selected field. The icon in the Status column is also updated to reflect the selected distribution.

Cancel. Closes the dialog box. Any changes that have been made are discarded.

Iterations

If you have specified more than one value for a fixed field or a distribution parameter, an independent set of simulated cases - effectively, a separate simulation - is generated for each specified value. This allows you to investigate the effect of varying the field or parameter. Each set of simulated cases is referred to as an *iteration*. In the simulated data, the iterations are stacked up.

If the **Create iteration field** check box on the Advanced Options view of the Simulation Generate node dialog is selected, an iteration field is added to the simulated data as a nominal field with numeric storage. The name of this field can be edited by typing in the **Name** field on the Advanced Options view. This field contains a label which indicates to which iteration each simulated case belongs. The form of the labels depends on the type of iteration:

- **Iterating a fixed field.** The label is the name of the field, followed by an equals sign, followed by the value of the field for that iteration, that is

  field_name = *field_value*
- **Iterating a distribution parameter.** The label is the name of the field, followed by a colon, followed by the name of the iterated parameter, followed by an equals sign, followed by the value of the parameter for that iteration, that is

  field_name:*parameter_name* = *parameter_value*
- **Iterating a distribution parameter for a categorical or range distribution.** The label is the name of the field, followed by a colon, followed by "Iteration", followed by the iteration number, that is

  field_name: Iteration *iteration_number*

Distributions

You can manually specify the probability distribution for any field by opening the **Specify Parameters** dialog box for that field, selecting the desired distribution from the **Distribution** list, and entering the distribution parameters in the **Distribution parameters** table. Following are some notes on particular distributions:

- **Categorical.** The categorical distribution describes an input field that has a fixed number of numeric values, referred to as categories. Each category has an associated probability such that the sum of the probabilities over all categories equals one.

  Note: If you specify probabilities for the categories that do not sum to 1, you will receive a warning.
- **Negative Binomial - Failures.** Describes the distribution of the number of failures in a sequence of trials before a specified number of successes are observed. The parameter *Threshold* is the specified number of successes and the parameter *Probability* is the probability of success in any given trial.
- **Negative Binomial - Trials.** Describes the distribution of the number of trials that are required before a specified number of successes are observed. The parameter *Threshold* is the specified number of successes and the parameter *Probability* is the probability of success in any given trial.
- **Range.** This distribution consists of a set of intervals with a probability assigned to each interval such that the sum of the probabilities over all intervals equals 1. Values within a given interval are drawn from a uniform distribution defined on that interval. Intervals are specified by entering a minimum value, a maximum value and an associated probability.

  For example, you believe that the cost of a raw material has a 40% chance of falling in the range of $10 - $15 per unit, and a 60% chance of falling in the range of $15 - $20 per unit. You would model the cost with a Range distribution consisting of the two intervals [10 - 15] and [15 - 20], setting the probability associated with the first interval to 0.4 and the probability associated with the second interval to 0.6. The intervals do not have to be contiguous and they can even be overlapping. For example, you might have specified the intervals $10 - $15 and $20 - $25 or $10 - $15 and $13 - $16.
- **Weibull.** The parameter *Location* is an optional location parameter, which specifies where the origin of the distribution is located.

The following table shows the distributions that are available for custom distribution fitting, and the acceptable values for the parameters. Some of these distributions are available for custom fitting to particular storage types, even though they are not fitted automatically to these storage types by the Simulation Fitting node.

Table 12. Distributions available for custom fitting

Distribution	Storage type supported for custom fitting	Parameters	Parameter limits	Notes
Bernoulli	Integer, real, datetime	Probability	0 ≤ *Probability* ≤ 1	
Beta	Integer, real, datetime	Shape 1 Shape 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum and maximum are optional.
Binomial	Integer, real, datetime	Number of trials (n) Probability Minimum Maximum	> 0, integer 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	Number of trials must be an integer. Minimum and maximum are optional.
Categorical	Integer, real, datetime, string	Category name (or label)	0 ≤ *Value* ≤ 1	Value is the probability of the category. The values must sum to 1, otherwise a warning is generated.
Dice	Integer, string	Sides	2 ≤ *Sides* ≤ 20	The probability of each category (side) is calculated as 1/N, where N is the number of sides. The probabilities cannot be edited.
Empirical	Integer, real, datetime			You cannot edit the empirical distribution, or select it as a type. The Empirical distribution is only available when there is historical data.
Exponential	Integer, real, datetime	Scale Minimum Maximum	> 0 < *Maximum* > *Minimum*	Minimum and maximum are optional.

Table 12. Distributions available for custom fitting (continued)

Distribution	Storage type supported for custom fitting	Parameters	Parameter limits	Notes
Fixed	Integer, real, datetime, string	Value		You cannot specify the Fixed distribution for every field. If you want every field in your generated data to be fixed, you can use a User Input node followed by a Balance node.
Gamma	Integer, real, datetime	Shape Scale Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum and maximum are optional. Distribution uses a rate parameter, with a shape parameter α = k and an inverse scale parameter β = 1/θ.
Lognormal	Integer, real, datetime	Shape 1 Shape 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum and maximum are optional.
Negative Binomial - Failures	Integer, real, datetime	Threshold Probability Minimum Maximum	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	Minimum and maximum are optional.
Negative Binomial - Trials	Integer, real, datetime	Threshold Probability Minimum Maximum	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	Minimum and maximum are optional.
Normal	Integer, real, datetime	Mean Standard deviation Minimum Maximum	≥ 0 > 0 < *Maximum* > *Minimum*	Minimum and maximum are optional.
Poisson	Integer, real, datetime	Mean Minimum Maximum	≥ 0 < *Maximum* > *Minimum*	Minimum and maximum are optional.

Table 12. Distributions available for custom fitting (continued)

Distribution	Storage type supported for custom fitting	Parameters	Parameter limits	Notes
Range	Integer, real, datetime	Begin(X) End(X) Probability(X)	 0 ≤ *Value* ≤ 1	X is the index of each bin. The probability values must sum to 1.
Triangular	Integer, real, datetime	Mode Minimum Maximum	*Minimum* ≤ *Value* ≤ *Maximum* < *Maximum* > *Minimum*	
Uniform	Integer, real, datetime	Minimum Maximum	< *Maximum* > *Minimum*	
Weibull	Integer, real, datetime	Rate Scale Location Minimum Maximum	> 0 > 0 ≥ 0 < *Maximum* > *Minimum*	Location, maximum and minimum are optional.

Extension Import node

With the Extension Import node, you can run R or Python for Spark scripts to import data.

Extension Import node - Syntax tab

Select your type of syntax – **R** or **Python for Spark**. Then enter or paste your custom script for importing data. When your syntax is ready, you can click **Run** to execute the Extension Import node.

Extension Import node - Console Output tab

The **Console Output** tab contains any output that is received when the R script or Python for Spark script on the Syntax tab runs (for example, if using an R script, it shows output received from the R console when the R script in the **R Syntax** field on the **Syntax** tab is executed). This output might include R or Python error messages or warnings that are produced when the R or Python script is executed. The output can be used, primarily, to debug the script. The **Console Output** tab also contains the script from the **R Syntax** or **Python Syntax** field.

Every time the Extension Import script is executed, the content of the **Console Output** tab is overwritten with the output received from the R console or Python for Spark. The output cannot be edited.

Filtering or renaming fields

You can rename or exclude fields at any point in a stream. For example, as a medical researcher, you may not be concerned about the potassium level (field-level data) of patients (record-level data); therefore, you can filter out the K (potassium) field. This can be done using a separate Filter node or using the Filter tab on a source or output node. The functionality is the same regardless of which node it is accessed from.

- From source nodes, such as Variable File, Fixed File, Statistics File, XML, or Extension Import, you can rename or filter fields as the data are read into IBM SPSS Modeler.
- Using a Filter node, you can rename or filter fields at any point in the stream.

- From Statistics Export, Statistics Transform, Statistics Model, and Statistics Output nodes, you can filter or rename fields to conform to IBM SPSS Statistics naming standards.
- You can use the Filter tab in any of the above nodes to define or edit multiple response sets.
- Finally, you can use a Filter node to map fields from one source node to another.

Geospatial Source Node

You use the Geospatial source node to bring map or spatial data into your data mining session. You can import data in one of two ways:

- In a shape file (`.shp`)
- By connecting to an ESRI server that contains a hierarchical file system which includes map files.

  Note: You can only connect to public map services.

Spatio-Temporal Prediction (STP) models can include map or spatial elements in their predictions. For more information on these models, see the topic titled "Spatio-Temporal Prediction modeling node" in the Time Series Models section of the Modeler Modeling Nodes guide (ModelerModelingNodes.pdf).

Setting Options for the Geospatial Source Node

Datasource type You can import data from either a **Shape file** (`.shp`), or connect to a **Map service**.

If you are using a **Shape file**, either enter the file name and the filepath to it, or browse to select the file. The file must either be on the local directory or accessed from a mapped drive; you cannot access the file by using a uniform naming convention (UNC) path.

Note: Shape data requires both a `.shp` and a `.dbf` file. The two files must have the same name and be in the same folder. The `.dbf` file is automatically imported when you select the `.shp` file. In addition, there might be a `.prj` file that specifies the coordinate system for the shape data.

If you are using a **Map service**, enter the URL to the service and click **Connect**. After you connect to the service, the layers within that service are displayed in the bottom of the dialog box in a tree structure in the **Available maps** pane; expand the tree and select the layer that you require.

Note: You can only connect to public map services.

Automatic definition of geospatial data

By default, SPSS Modeler automatically defines, when possible, any geospatial data fields in the source node with the correct metadata. Metadata might include the measurement level of the geospatial field (such as Point or Polygon) and the coordinate system that is used by the fields, including details such as the origin point (for example, latitude 0, longitude 0) and the units of measure. For more information about measurement levels, see “Geospatial measurement sublevels” on page 124.

The `.shp` and `.dbf` files that make up the shape file contain a common identifier field that is used as a key. For example, the `.shp` file might contain countries, with the country name field used as an identifier, and the `.dbf` file might contain information about those countries with the name of the country that is also used as the identifier.

Note: If the coordinate system is not the same as the default SPSS Modeler coordinate system you may have to reproject the data to use the required coordinate system. For more information, see “Reprojection Node” on page 164.

JSON Source node

Use the JSON source node to import data from a JSON file into an SPSS Modeler stream, using UTF-8 encoding. Data in a JSON file can be in form of *object*, *array*, or *value*. This JSON source node only supports reading in an *array* of objects, and the object can't be nested.

Example JSON data:

```
[
    {
        "After": 122762,
        "Promotion": 1467,
        "Cost": 23.99,
        "Class": "Confection",
        "Before": 114957
    },
    {
        "After": 137097,
        "Promotion": 1745,
        "Cost": 79.29,
        "Class": "Drink",
        "Before": 123378
    }
]
```

When SPSS Modeler reads data from a JSON file, it performs the following translations.

Table 13. JSON data storage translation

JSON Value	SPSS Modeler Data Storage
string	String
number(int)	Integer
number(real)	Real
true	1(Integer)
false	0(Integer)
null	Missing values

The following options are available in the JSON source node dialog.

JSON data source. Select the JSON file to import.

JSON string format. Specify the format of the JSON string. Select **Records** if your JSON file is a collection of name and value pairs. The JSON source node imports names as field names in SPSS Modeler. Or select **Values** if your JSON data only uses values (no names).

Common Source Node Tabs

The following options can be specified for all source nodes by clicking the corresponding tab:

- **Data tab.** Used to change the default storage type.
- **Filter tab.** Used to eliminate or rename data fields. This tab offers the same functionality as the Filter node. See the topic "Setting filtering options" on page 134 for more information.
- **Types tab.** Used to set measurement levels. This tab offers the same functionality as the Type node.
- **Annotations tab.** Used for all nodes, this tab offers options to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

Setting Measurement Levels in the Source Node

Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes. The following properties are available:

- **Field** Double-click any field name to specify value and field labels for data in IBM SPSS Modeler. For example, field metadata imported from IBM SPSS Statistics can be viewed or modified here. Similarly, you can create new labels for fields and their values. The labels that you specify here are displayed throughout IBM SPSS Modeler depending on the selections you make in the stream properties dialog box.

- **Measurement** This is the measurement level, used to describe characteristics of the data in a given field. If all of the details of a field are known, it is called **fully instantiated**. For more information, see "Measurement levels" on page 122.

  Note: The measurement level of a field is different from its storage type, which indicates whether the data are stored as strings, integers, real numbers, dates, times, timestamps, or lists.
- **Values** This column enables you to specify options for reading data values from the data set, or use the **Specify** option to specify measurement levels and values in a separate dialog box. You can also choose to pass fields without reading their values. For more information, see "Data Values" on page 127.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.
- **Missing** Used to specify how missing values for the field will be handled. For more information, see "Defining Missing Values" on page 130.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.
- **Check** In this column, you can set options to ensure that field values conform to the specified values or ranges. For more information, see "Checking Type Values" on page 131.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.
- **Role** Used to tell modeling nodes whether fields will be **Input** (predictor fields) or **Target** (predicted fields) for a machine-learning process. **Both** and **None** are also available roles, along with **Partition**, which indicates a field used to partition records into separate samples for training, testing, and validation. The value **Split** specifies that separate models will be built for each possible value of the field. For more information, see "Setting the field role" on page 131.

See the topic "Type Node" on page 121 for more information.

When to instantiate at the source node

There are two ways you can learn about the data storage and values of your fields. This *instantiation* can occur at either the source node, when you first bring data into IBM SPSS Modeler, or by inserting a Type node into the data stream.

Instantiating at the source node is useful when:

- The dataset is small.
- You plan to derive new fields using the Expression Builder (instantiating makes field values available from the Expression Builder).

Generally, if your dataset is not very large and you don't plan to add fields later in the stream, instantiating at the source node is the most convenient method.

Note: If you export data in a database export node, the data must be fully instantiated.

Filtering Fields from the Source Node

The Filter tab on a source node dialog box enables you to exclude fields from downstream operations based on your initial examination of the data. This is useful, for example, if there are duplicate fields in the data or if you are already familiar enough with the data to exclude irrelevant fields. Alternatively, you can add a separate Filter node later in the stream. The functionality is similar in both cases. See the topic "Setting filtering options" on page 134 for more information.

Chapter 3. Record Operations Nodes

Overview of Record Operations

Record operations nodes are used to make changes to data at the record level. These operations are important during the **Data Understanding** and **Data Preparation** phases of data mining because they allow you to tailor the data to your particular business need.

For example, based on the results of the data audit conducted using the Data Audit node (Output palette), you might decide that you would like customer purchase records for the past three months to be merged. Using a Merge node, you can merge records based on the values of a key field, such as *Customer ID*. Or you might discover that a database containing information about Web site hits is unmanageable with over one million records. Using a Sample node, you can select a subset of data for use in modeling.

The Record Operations palette contains the following nodes:

The Select node selects or discards a subset of records from the data stream based on a specific condition. For example, you might select the records that pertain to a particular sales region.

The Sample node selects a subset of records. A variety of sample types are supported, including stratified, clustered, and nonrandom (structured) samples. Sampling can be useful to improve performance, and to select groups of related records or transactions for analysis.

The Balance node corrects imbalances in a dataset, so it conforms to a specified condition. The balancing directive adjusts the proportion of records where a condition is true by the factor specified.

The Aggregate node replaces a sequence of input records with summarized, aggregated output records.

The Recency, Frequency, Monetary (RFM) Aggregate node enables you to take customers' historical transactional data, strip away any unused data, and combine all of their remaining transaction data into a single row that lists when they last dealt with you, how many transactions they have made, and the total monetary value of those transactions.

The Sort node sorts records into ascending or descending order based on the values of one or more fields.

The Merge node takes multiple input records and creates a single output record containing some or all of the input fields. It is useful for merging data from different sources, such as internal customer data and purchased demographic data.

The Append node concatenates sets of records. It is useful for combining datasets with similar structures but different data.

The Distinct node removes duplicate records, either by passing the first distinct record to the data stream or by discarding the first record and passing any duplicates to the data stream instead.

The Streaming Time Series node builds and scores time series models in one step. You can use the node with data in either a local or distributed environment; in a distributed environment you can harness the power of IBM SPSS Analytic Server

The Spectral Clustering© algorithm uses several eigenvectors to project data into a space with fewer dimensions. Then a k-means clustering algorithm is applied in the new space to separate the data into clusters. It's reasonably fast for small records with many fields, and computationally expensive for large data sets. The Spectral Clustering node in SPSS Modeler exposes the core features and commonly used parameters of the Spectral Clustering library. The node is implemented in Python.

Space-Time-Boxes (STB) are an extension of Geohashed spatial locations. More specifically, an STB is an alphanumeric string that represents a regularly shaped region of space and time.

The Streaming TCM node builds and scores temporal causal models in one step.

The CPLEX Optimization node provides the ability to use complex mathematical (CPLEX) based optimization via an Optimization Programming Language (OPL) model file. This functionality was available in the IBM Analytical Decision Management product, which is no longer supported. But you can also use the CPLEX node in SPSS Modeler without requiring IBM Analytical Decision Management.

Many of the nodes in the Record Operations palette require you to use a CLEM expression. If you are familiar with CLEM, you can type an expression in the field. However, all expression fields provide a button that opens the CLEM Expression Builder, which helps you create such expressions automatically.

Figure 1. Expression Builder button

Select Node

You can use Select nodes to select or discard a subset of records from the data stream based on a specific condition, such as BP (blood pressure) = `"HIGH"`.

Mode. Specifies whether records that meet the condition will be included or excluded from the data stream.

- **Include.** Select to include records that meet the selection condition.

- **Discard.** Select to exclude records that meet the selection condition.

Condition. Displays the selection condition that will be used to test each record, which you specify using a CLEM expression. Either enter an expression in the window or use the Expression Builder by clicking the calculator (Expression Builder) button to the right of the window.

If you choose to discard records based on a condition, such as the following:

```
(var1='value1' and var2='value2')
```

the Select node by default also discards records having null values for all selection fields. To avoid this, append the following condition to the original one:

```
and not(@NULL(var1) and @NULL(var2))
```

Select nodes are also used to choose a proportion of records. Typically, you would use a different node, the Sample node, for this operation. However, if the condition you want to specify is more complex than the parameters provided, you can create your own condition using the Select node. For example, you can create a condition such as:

```
BP = "HIGH" and random(10) <= 4
```

This will select approximately 40% of the records showing high blood pressure and pass those records downstream for further analysis.

Sample Node

You can use Sample nodes to select a subset of records for analysis, or to specify a proportion of records to discard. A variety of sample types are supported, including stratified, clustered, and nonrandom (structured) samples. Sampling can be used for several reasons:

- To improve performance by estimating models on a subset of the data. Models estimated from a sample are often as accurate as those derived from the full dataset, and may be more so if the improved performance allows you to experiment with different methods you might not otherwise have attempted.
- To select groups of related records or transactions for analysis, such as selecting all the items in an online shopping cart (or market basket), or all the properties in a specific neighborhood.
- To identify units or cases for random inspection in the interest of quality assurance, fraud prevention, or security.

Note: If you simply want to partition your data into training and test samples for purposes of validation, a Partition node can be used instead. See the topic "Partition Node" on page 155 for more information.

Types of Samples

Clustered samples. Sample groups or clusters rather than individual units. For example, suppose you have a data file with one record per student. If you cluster by school and the sample size is 50%, then 50% of schools will be chosen and all students from each selected school will be picked. Students in unselected schools will be rejected. On average, you would expect about 50% of students to be picked, but because schools vary in size, the percentage may not be exact. Similarly, you could cluster shopping cart items by transaction ID to make sure that all items from selected transactions are maintained. For an example that clusters properties by town, see the *complexsample_property.str* sample stream.

Stratified samples. Select samples independently within non-overlapping subgroups of the population, or strata. For example, you can ensure that men and women are sampled in equal proportions, or that every region or socioeconomic group within an urban population is represented. You can also specify a different sample size for each stratum (for example, if you think that one group has been under-represented in the original data). For an example that stratifies properties by county, see the *complexsample_property.str* sample stream.

Systematic or 1-in-n sampling. When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

Sampling weights. Sampling weights are automatically computed while drawing a complex sample and roughly correspond to the "frequency" that each sampled unit represents in the original data. Therefore, the sum of the weights over the sample should estimate the size of the original data.

Sampling Frame

A sampling frame defines the potential source of cases to be included in a sample or study. In some cases, it may be feasible to identify every single member of a population and include any one of them in a sample--for example, when sampling items that come off a production line. More often, you will not be able to access every possible case. For example, you cannot be sure who will vote in an election until after the election has happened. In this case, you might use the electoral register as your sampling frame, even though some registered people won't vote, and some people may vote despite not having been listed at the time you checked the register. Anybody not in the sampling frame has no prospect of being sampled. Whether your sampling frame is close enough in nature to the population you are trying to evaluate is a question that must be addressed for each real-life case.

Sample node options

You can choose the **Simple** or **Complex** method as appropriate for your requirements.

Simple sampling options

The Simple method allows you to select a random percentage of records, select contiguous records, or select every *nth* record.

Mode. Select whether to pass (include) or discard (exclude) records for the following modes:

- **Include sample.** Includes selected records in the data stream and discards all others. For example, if you set the mode to **Include sample** and set the **1-in-n** option to 5, then every fifth record will be included, yielding a dataset that is roughly one-fifth the original size. This is the default mode when sampling data, and the only mode when using the complex method.
- **Discard sample.** Excludes selected records and includes all others. For example, if you set the mode to **Discard sample** and set the **1-in-n** option to 5, then every fifth record will be discarded. This mode is only available with the simple method.

Sample. Select the method of sampling from the following options:

- **First.** Select to use contiguous data sampling. For example, if the maximum sample size is set to 10000, then the first 10,000 records will be selected.
- **1-in-n.** Select to sample data by passing or discarding every *n*th record. For example, if *n* is set to 5, then every fifth record will be selected.
- **Random %.** Select to sample a random percentage of the data. For example, if you set the percentage to 20, then 20% of the data will either be passed to the data stream or discarded, depending on the mode selected. Use the field to specify a sampling percentage. You can also specify a seed value using the **Set random seed** control.

  Use block level sampling (in-database only). This option is enabled only if you choose random percentage sampling when performing in-database mining on an Oracle or IBM Db2 database. In these circumstances, block-level sampling can be more efficient.

  Note: You do not get an exact number of rows returned each time you run the same random sample settings. This is because each input record has a probability of `N/100` of being included in the sample (where `N` is the **Random %** you specify in the node) and the probabilities are independent; therefore the results are not exactly `N%`.

Maximum sample size. Specifies the maximum number of records to include in the sample. This option is redundant and therefore disabled when **First** and **Include** are selected. Also note that when used in combination with the **Random %** option, this setting may prevent certain records from being selected. For example, if you have 10 million records in your dataset, and you select 50% of records with a maximum sample size of three million records, then 50% of the first six million records will be selected, and the remaining four million records have no chance of being selected. To avoid this limitation, select the

Complex sampling method, and request a random sample of three million records without specifying a cluster or stratify variable.

Complex sampling options

Complex sample options allow for finer control of the sample, including clustered, stratified, and weighted samples along with other options.

Cluster and stratify. Allows you to specify cluster, stratify, and input weight fields if needed. See the topic "Cluster and Stratify Settings" on page 67 for more information.

Sample type.

- **Random.** Selects clusters or records randomly within each strata.
- **Systematic.** Selects records at a fixed interval. This option works like the *1 in n* method, except the position of the first record changes depending on a random seed. The value of *n* is determined automatically based on the sample size or proportion.

Sample units. You can select proportions or counts as the basic sample units.

Sample size. You can specify the sample size in several ways:

- **Fixed.** Allows you to specify the overall size of the sample as a count or proportion.
- **Custom.** Allows you to specify the sample size for each subgroup or strata. This option is only available if a stratification field has been specified in the Cluster and Stratify sub dialog box.
- **Variable.** Allows the user to pick a field that defines the sample size for each subgroup or strata. This field should have the same value for each record within a particular stratum; for example, if the sample is stratified by county, then all records with *county = Surrey* must have the same value. The field must be numeric and its values must match the selected sample units. For proportions, values should be greater than 0 and less than 1; for counts, the minimum value is 1.

Minimum sample per stratum. Specifies a minimum number of records (or minimum number of clusters if a cluster field is specified).

Maximum sample per stratum. Specifies a maximum number of records or clusters. If you select this option without specifying a cluster or stratify field, a random or systematic sample of the specified size will be selected.

Set random seed. When sampling or partitioning records based on a random percentage, this option allows you to duplicate the same results in another session. By specifying the starting value used by the random number generator, you can ensure the same records are assigned each time the node is executed. Enter the desired seed value, or click the **Generate** button to automatically generate a random value. If this option is not selected, a different sample will be generated each time the node is executed.

Note: When using the **Set random seed** option with records read from a database, a Sort node may be required prior to sampling in order to ensure the same result each time the node is executed. This is because the random seed depends on the order of records, which is not guaranteed to stay the same in a relational database. See the topic "Sort Node" on page 74 for more information.

Cluster and Stratify Settings

The Cluster and Stratify dialog box enables you to select cluster, stratification, and weight fields when drawing a complex sample.

Clusters. Specifies a categorical field used to cluster records. Records are sampled based on cluster membership, with some clusters included and others not. But if any record from a given cluster is included, all are included. For example, when analyzing product associations in shopping carts, you could cluster items by transaction ID to make sure that all items from selected transactions are maintained. Instead of sampling records—which would destroy information about what items are sold together—you can sample transactions to make sure that all records for selected transactions are preserved.

Stratify by. Specifies a categorical field used to stratify records so that samples are selected independently within non-overlapping subgroups of the population, or strata. If you select a 50% sample

stratified by gender, for example, then two 50% samples will be taken, one for the men and one for the women. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups, allowing you to ensure adequate sample sizes for subgroups of interest. If there are three times more women than men in the original dataset, this ratio will be preserved by sampling separately from each group. Multiple stratification fields can also be specified (for example, sampling product lines within regions or vice-versa).

Note: If you stratify by a field that has missing values (null or system missing values, empty strings, white space, and blank or user-defined missing values), then you cannot specify custom sample sizes for strata. If you want to use custom sample sizes when stratifying by a field with missing or blank values, then you need to fill them upstream.

Use input weight. Specifies a field used to weight records prior to sampling. For example, if the weight field has values ranging from 1 to 5, records weighted 5 are five times as likely to be selected. The values of this field will be overwritten by the final output weights generated by the node (see following paragraph).

New output weight. Specifies the name of the field where final weights are written if no input weight field is specified. (If an input weight field is specified, its values are replaced by the final weights as noted above, and no separate output weight field is created.) The output weight values indicate the number of records represented by each sampled record in the original data. The sum of the weight values gives an estimate of the sample size. For example, if a random 10% sample is taken, the output weight will be 10 for all records, indicating that each sampled record represents roughly ten records in the original data. In a stratified or weighted sample, the output weight values may vary based on the sample proportion for each stratum.

Comments

- Clustered sampling is useful if you cannot get a complete list of the population you want to sample, but can get complete lists for certain groups or clusters. It is also used when a random sample would produce a list of test subjects that it would be impractical to contact. For example, it would be easier to visit all farmers in one county than a selection of farmers scattered across every county in the nation.
- You can specify both cluster and stratify fields in order to sample clusters independently within each strata. For example, you could sample property values stratified by county, and cluster by town within each county. This will ensure that an independent sample of towns is drawn from within each county. Some towns will be included and others will not, but for each town that is included, all properties within the town are included.
- To select a random sample of units from within each cluster, you can string two Sample nodes together. For example, you could first sample townships stratified by county as described above. Then attach a second Sample node and select *town* as a stratify field, allowing you to sample a proportion of records from within each township.
- In cases where a combination of fields is required to uniquely identify clusters, a new field can be generated using a Derive node. For example, if multiple shops use the same numbering system for transactions, you could derive a new field that concatenates the shop and transaction IDs.

Sample Sizes for Strata

When drawing a stratified sample, the default option is to sample the same proportion of records or clusters from each stratum. If one group outnumbers another by a factor of 3, for example, you typically want to preserve the same ratio in the sample. If this is not the case, however, you can specify the sample size separately for each stratum.

The Sample Sizes for Strata dialog box lists each value of the stratification field, allowing you to override the default for that stratum. If multiple stratification fields are selected, every possible combination of values is listed, allowing you to specify the size for each ethnic group within each city, for example, or each town within each county. Sizes are specified as proportions or counts, as determined by the current setting in the Sample node.

To Specify Sample Sizes for Strata

1. In the Sample node, select **Complex**, and select one or more stratification fields. See the topic "Cluster and Stratify Settings" on page 67 for more information.
2. Select **Custom**, and select **Specify Sizes**.
3. In the Sample Sizes for Strata dialog box, click the **Read Values** button at lower left to populate the display. If necessary, you may need to instantiate values in an upstream source or Type node. See the topic "What is instantiation?" on page 126 for more information.
4. Click in any row to override the default size for that stratum.

Notes on Sample Size

Custom sample sizes may be useful if different strata have different variances, for example, in order to make sample sizes proportional to the standard deviation. (If the cases within the stratum are more varied, you need to sample more of them to get a representative sample.) Or if a stratum is small, you may wish to use a higher sample proportion to ensure that a minimum number of observations is included.

Note: If you stratify by a field that has missing values (null or system missing values, empty strings, white space, and blank or user-defined missing values), then you cannot specify custom sample sizes for strata. If you want to use custom sample sizes when stratifying by a field with missing or blank values, then you need to fill them upstream.

Balance Node

You can use Balance nodes to correct imbalances in datasets so they conform to specified test criteria. For example, suppose that a dataset has only two values--*low* or *high*--and that 90% of the cases are *low* while only 10% of the cases are *high*. Many modeling techniques have trouble with such biased data because they will tend to learn only the *low* outcome and ignore the *high* one, since it is more rare. If the data are well balanced with approximately equal numbers of *low* and *high* outcomes, models will have a better chance of finding patterns that distinguish the two groups. In this case, a Balance node is useful for creating a balancing directive that reduces cases with a *low* outcome.

Balancing is carried out by duplicating and then discarding records based on the conditions you specify. Records for which no condition holds are always passed through. Because this process works by duplicating and/or discarding records, the original sequence of your data is lost in downstream operations. Be sure to derive any sequence-related values before adding a Balance node to the data stream.

Note: Balance nodes can be generated automatically from distribution charts and histograms. For example, you can balance your data to show equal proportions across all categories of a categorical field, as shown in a distribution plot.

Example. When building an RFM stream to identify recent customers who have positively responded to previous marketing campaigns, the marketing department of a sales company uses a Balance node to balance the differences between true and false responses in the data.

Setting Options for the Balance Node

Record balancing directives. Lists the current balancing directives. Each directive includes both a factor and a condition that tells the software to "increase the proportion of records by a factor specified where the condition is true." A factor lower than 1.0 means that the proportion of indicated records will be decreased. For example, if you want to decrease the number of records where drug Y is the treatment drug, you might create a balancing directive with a factor of 0.7 and a condition `Drug = "drugY"`. This directive means that the number of records where drug Y is the treatment drug will be reduced to 70% for all downstream operations.

Note: Balance factors for reduction may be specified to four decimal places. Factors set below 0.0001 will result in an error, since the results do not compute correctly.

- **Create conditions** by clicking the button to the right of the text field. This inserts an empty row for entering new conditions. To create a CLEM expression for the condition, click the Expression Builder button.

- **Delete directives** using the red delete button.
- **Sort directives** using the up and down arrow buttons.

Only balance training data. If a partition field is present in the stream, this option balances data in the training partition only. In particular, this may be useful if generating adjusted propensity scores, which require an unbalanced testing or validation partition. If no partition field is present in the stream (or if multiple partition fields are specified), then this option is ignored and all data are balanced.

Aggregate Node

Aggregation is a data preparation task frequently used to reduce the size of a dataset. Before proceeding with aggregation, you should take time to clean the data, concentrating especially on missing values. Once you have aggregated, potentially useful information regarding missing values may be lost.

You can use an Aggregate node to replace a sequence of input records with summary, aggregated output records. For example, you might have a set of input sales records such as those shown in the following table.

Table 14. Sales record input example

Age	Sex	Region	Branch	Sales
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	2	11
29	F	S	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

You can aggregate these records with *Sex* and *Region* as key fields. Then choose to aggregate *Age* with the mode **Mean** and *Sales* with the mode **Sum**. Select **Include record count in field** in the Aggregate node dialog box and your aggregated output would be as shown in the following table.

Table 15. Aggregated record example

Age (mean)	Sex	Region	Sales (sum)	Record Count
35.5	F	N	25	4
29	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

From this you learn, for example, that the average age of the four female sales staff in the North region is 35.5, and the sum total of their sales was 25 units.

Note: Fields such as *Branch* are automatically discarded when no aggregate mode is specified.

Setting options for the Aggregate node

On the Aggregate node you specify the following.

- One or more key fields to use as categories for the aggregation
- One or more aggregate fields for which to calculate the aggregate values
- One or more aggregation modes (types of aggregation) to output for each aggregate field

You can also specify the default aggregation modes to use for newly added fields, and use expressions (similar to formulae) to categorize aggregation.

Note that for added performance, aggregations operations may benefit from enabling parallel processing.

Key fields. Lists fields that can be used as categories for aggregation. Both continuous (numeric) and categorical fields can be used as keys. If you choose more than one key field, the values will be combined to produce a key value for aggregating records. One aggregated record will be generated for each unique key field. For example, if `Sex` and `Region` are your key fields, each unique combination of `M` and `F` with regions `N` and `S` (four unique combinations) will have an aggregated record. To add a key field, use the Field Chooser button to the right of the window.

The rest of the dialog box is split into two main areas - **Basic Aggregates** and **Aggregate Expressions**.

Basic Aggregates

Aggregate fields. Lists the fields for which values will be aggregated as well as the selected modes of aggregation. To add fields to this list, use the Field Chooser button on the right. The following aggregation modes are available.

Note: Some modes are not applicable to non-numeric fields (for example, **Sum** for a date/time field). Modes that cannot be used with a selected aggregate field are disabled.

- **Sum.** Select to return summed values for each key field combination. The sum is the total of the values, across all cases with nonmissing values.
- **Mean.** Select to return the mean values for each key field combination. The mean is a measure of central tendency, and is the arithmetic average (the sum divided by the number of cases).
- **Min.** Select to return minimum values for each key field combination.
- **Max.** Select to return maximum values for each key field combination.
- **SDev.** Select to return the standard deviation for each key field combination. The standard deviation is a measure of dispersion around the mean, and is equal to the square root of the variance measurement.
- **Median.** Select to return the median values for each key field combination. The median is a measure of central tendency that is not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values). Also known as the 50th percentile or 2nd quartile.
- **Count.** Select to return the count of non-null values for each key field combination.
- **Variance.** Select to return the variance values for each key field combination. The variance is a measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases.
- **1st Quartile.** Select to return the 1st quartile (25th percentile) values for each key field combination.
- **3rd Quartile.** Select to return the 3rd quartile (75th percentile) values for each key field combination.

Note: When running a stream containing an Aggregate node, the values returned for 1st and 3rd Quartiles when pushing back SQL to an Oracle database may differ from those returned in native mode.

Default mode. Specify the default aggregation mode to be used for newly added fields. If you frequently use the same aggregation, select one or more modes here and use the Apply to All button on the right to apply the selected modes to all fields listed above.

New field name extension. Select to add a suffix or prefix, such as **1** or **new**, to duplicate aggregated fields. For example, the result of a minimum values aggregation on the field **Age** will produce a field name called `Age_Min_1` if you have selected the suffix option and specified **1** as the extension. Note that

aggregation extensions such as _Min or Max_ are automatically added to the new field, indicating the type of aggregation performed. Select **Suffix** or **Prefix** to indicate your preferred extension style.

Include record count in field. Select to include an extra field in each output record called Record_Count, by default. This field indicates how many input records were aggregated to form each aggregate record. Create a custom name for this field by typing in the edit field.

Note: System null values are excluded when aggregates are computed, but they are included in the record count. Blank values, on the other hand, are included in both aggregation and record count. To exclude blank values, you can use a Filler node to replace blanks with null values. You can also remove blanks using a Select node.

Aggregate Expressions

Expressions are similar to formulas that are created from values, field names, operators, and functions. Unlike functions that work on a single record at a time, aggregate expressions operate on a group, set, or collection of records.

Note: You can only create aggregate expressions if the stream includes a database connection (by means of a Database source node).

New expressions are created as derived fields; to create an expression you use the *Database Aggregates* functions which are available from the Expression Builder.

For more information about the Expression Builder, see the IBM SPSS Modeler User's Guide (ModelerUsersGuide.pdf).

Note that there is a connection between the **Key Fields** and any aggregate expressions you create because the aggregate expressions are grouped by the key field.

Valid aggregate expressions are ones that evaluate to aggregate outcomes; a couple of examples of valid aggregate expressions, and the rules that govern them, are as follows:

- You can use scalar functions to combine multiple aggregation functions together to produce a single aggregation result. For example:

```
max(C01) - min(C01)
```

- An aggregation function can operate on the result of multiple scalar functions. For example:

```
sum (C01*C01)
```

Aggregate optimization settings

On the Optimization tab you specify the following.

Keys are contiguous. Select this option if you know that all records with the same key values are grouped together in the input (for example, if the input is sorted on the key fields). Doing so can improve performance.

Allow approximation for Median and Quartiles. The Order statistics (Median, 1st Quartile, and 3rd Quartile) are currently not supported when processing data in Analytic Server. If you are using Analytic Server you can select this check box to use an approximated value for these statistics instead which is calculated by binning the data and then computing an estimate for the statistic based on the distribution across the bins. By default, this option is unchecked.

Number of bins. Only available if you select the **Allow approximation for Median and Quartiles** check box. Select the number of bins to be used when estimating the statistic; the number of bins affects the **Maximum error %**. By default, the number of bins is 1000, which corresponds to a maximum error of 0.1 percent of the range.

RFM Aggregate Node

The Recency, Frequency, Monetary (RFM) Aggregate node enables you to take customers' historical transactional data, strip away any unused data, and combine all of their remaining transaction data into a single row, using their unique customer ID as a key, that lists when they last dealt with you (recency), how many transactions they have made (frequency), and the total value of those transactions (monetary).

Before proceeding with any aggregation, you should take time to clean the data, concentrating especially on any missing values.

Once you have identified and transformed the data using the RFM Aggregate node, you may use an RFM Analysis node to carry out further analysis. See the topic "RFM Analysis Node" on page 152 for more information.

Note that once the data file has been run through the RFM Aggregate node, it will not have any target values; therefore, before being able to use it as inputs for further predictive analysis with any modeling nodes such as C5.0 or CHAID, you will need to merge it with other customer data (for example, by matching the customer IDs). See the topic "Merge Node" on page 74 for more information.

The RFM Aggregate and RFM Analysis nodes in IBM SPSS Modeler are set up to use independent binning; that is, they rank and bin data on each measure of recency, frequency, and monetary value, without regard to their values or the other two measures.

Setting Options for the RFM Aggregate Node

The Settings tab of the RFM Aggregate node contains the following fields.

Calculate Recency relative to Specify the date from which the recency of transactions will be calculated. This may be either a **Fixed date** that you enter, or **Today's date**, as set by your system. **Today's date** is entered by default and is automatically updated when the node is executed.

Note: The display of the **Fixed date** may be different for different locales. For example, if the value `2007-8-10` is stored in your stream as `Fri Aug 10 00:00:00 CST 2007` it is a time and date in the time zone 'UTC+8'. However, it displays as `Thu Aug 9 12:00:00 EDT 2007` in the time zone 'UTC-8'.

IDs are contiguous If your data are presorted so that all records with the same ID appear together in the data stream, select this option to speed up processing. If your data are not presorted (or you are not sure), leave this option unselected, and the node will sort the data automatically.

ID Select the field to be used to identify the customer and their transactions. To display the fields from which you can select, use the Field Chooser button on the right.

Date Select the date field to be used to calculate recency against. To display the fields from which you can select, use the Field Chooser button on the right.

Note that this requires a field with a storage of date, or timestamp, in the appropriate format to use as input. For example, if you have a string field with values like *Jan 2007*, *Feb 2007*, and so on, you can convert this to a date field using a Filler node and the `to_date()` function. See the topic "Storage Conversion Using the Filler Node" on page 143 for more information.

Value Select the field to be used to calculate the total monetary value of the customer's transactions. To display the fields from which you can select, use the Field Chooser button on the right. *Note*: This must be a numeric value.

New field name extension Select to append either a suffix or prefix, such as "12_month", to the newly generated recency, frequency, and monetary fields. Select **Suffix** or **Prefix** to indicate your preferred extension style. For example, this may be useful when examining several time periods.

Discard records with value below If required, you can specify a minimum value below which any transaction details are not used when calculating the RFM totals. The units of value relate to the **Value** field selected.

Include only recent transactions If you are analyzing a large database, you can specify that only the latest records are used. You can chose to use the data recorded either after a certain date or within a recent period:

- **Transaction date after** Specify the transaction date after which records will be included in your analysis.
- **Transaction within the last** Specify the number and type of periods (days, weeks, months, or years) back from the **Calculate recency relative to** date after which records will be included in your analysis.

Save date of second most recent transaction If you want to know the date of the second most recent transaction for each customer, select this box. In addition, you can then select the **Save date of third most recent transaction** box as well. For example, this can help you identify customers who may have carried out many transactions some considerable time ago, but only one recent transaction.

Sort Node

You can use Sort nodes to sort records into ascending or descending order based on the values of one or more fields. For example, Sort nodes are frequently used to view and select records with the most common data values. Typically, you would first aggregate the data using the Aggregate node and then use the Sort node to sort the aggregated data into descending order of record counts. Displaying these results in a table will allow you to explore the data and to make decisions, such as selecting the records of the top 10 best customers.

The Settings tab of the Sort node contains the following fields.

Sort by. All fields selected to use as sort keys are displayed in a table. A key field works best for sorting when it is numeric.

- **Add fields** to this list using the Field Chooser button on the right.
- **Select an order** by clicking the **Ascending** or **Descending** arrow in the table's *Order* column.
- **Delete fields** using the red delete button.
- **Sort directives** using the up and down arrow buttons.

Default sort order. Select either **Ascending** or **Descending** to use as the default sort order when new fields are added above.

Note: The Sort node is not applied if there is a Distinct node down the model stream. For information about the Distinct node, see “Distinct node” on page 82.

Sort Optimization Settings

If you are working with data you know are already sorted by some key fields, you can specify which fields are already sorted, allowing the system to sort the rest of the data more efficiently. For example, you want to sort by *Age* (descending) and *Drug* (ascending) but know your data are already sorted by *Age* (descending).

Data is presorted. Specifies whether the data are already sorted by one or more fields.

Specify existing sort order. Specify the fields that are already sorted. Using the Select Fields dialog box, add fields to the list. In the *Order* column, specify whether each field is sorted in ascending or descending order. If you are specifying multiple fields, make sure that you list them in the correct sorting order. Use the arrows to the right of the list to arrange the fields in the correct order. If you make a mistake in specifying the correct existing sort order, an error will appear when you run the stream, displaying the record number where the sorting is inconsistent with what you specified.

Note: Sorting speed may benefit from enabling parallel processing.

Merge Node

The function of a Merge node is to take multiple input records and create a single output record containing all or some of the input fields. This is a useful operation when you want to merge data from different

sources, such as internal customer data and purchased demographic data. You can merge data in the following ways.

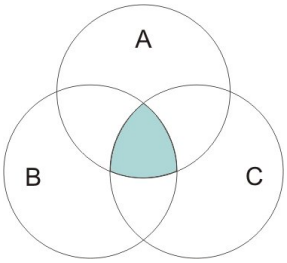
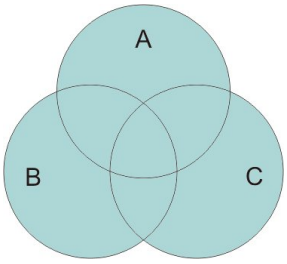
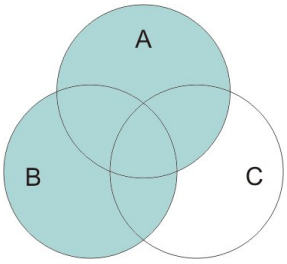
- Merge by **Order** concatenates corresponding records from all sources in the order of input until the smallest data source is exhausted. It is important if using this option that you have sorted your data using a Sort node.
- Merge using a **Key** field, such as *Customer ID*, to specify how to match records from one data source with records from the other(s). Several types of joins are possible, including inner join, full outer join, partial outer join, and anti-join. See the topic "Types of Joins" on page 75 for more information.
- Merge by **Condition** means that you can specify a condition to be satisfied for the merge to take place. You can specify the condition directly in the node, or build the condition using the Expression Builder.
- Merge by **Ranked Condition** is a left sided outer join in which you specify a condition to be satisfied for the merge to take place and a ranking expression which sorts into order from low to high. Most often used to merge geospatial data, you can specify the condition directly in the node, or build the condition using the Expression Builder.

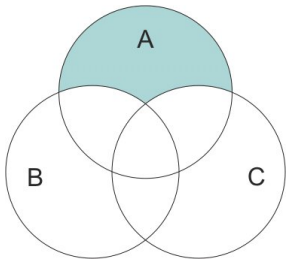
Types of Joins

When using a key field for data merging, it is useful to spend some time thinking about which records will be excluded and which will be included. There are a variety of joins, which are discussed in detail below.

The two basic types of joins are referred to as inner and outer joins. These methods are frequently used to merge tables from related datasets based on common values of a key field, such as *Customer ID*. Inner joins allow for clean merging and an output dataset that includes only complete records. Outer joins also include complete records from the merged data, but they also allow you to include unique data from one or more input tables.

The types of joins allowed are described in greater detail below.

	An **inner join** includes only records in which a value for the key field is common to all input tables. That is, unmatched records will not be included in the output dataset.
	A **full outer join** includes all records, both matching and nonmatching, from the input tables. Left and right outer joins are referred to as partial outer joins and are described below.
	A **partial outer join** includes all records matched using the key field as well as unmatched records from specified tables. (Or, to put it another way, all records from some tables and only matching records from others.) Tables (such as A and B shown here) can be selected for inclusion in the outer join using the Select button on the Merge tab. Partial joins are also called left or right outer joins when only two tables are being merged. Since IBM SPSS Modeler allows the merging of more than two tables, we refer to this as a partial outer join.



An **anti-join** includes only unmatched records for the first input table (Table A shown here). This type of join is the opposite of an inner join and does not include complete records in the output dataset.

For example, if you have information about farms in one dataset and farm-related insurance claims in another, you can match the records from the first source to the second source using the Merge options.

To determine if a customer in your farm sample has filed an insurance claim, use the inner join option to return a list showing where all IDs match from the two samples.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

Figure 2. Sample output for an inner join merge

Using the full outer join option returns both matching and nonmatching records from the input tables. The system-missing value ($null$) will be used for any incomplete values.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.1(
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	$null$	$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	$null$	$null$	$null$	$null$	$null$	$null$	$null$	arable_de	122135.0

Figure 3. Sample output for a full outer join merge

A partial outer join includes all records matched using the key field as well as unmatched records from specified tables. The table displays all of the records matched from the ID field as well as the records matched from the first dataset.

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	$null$	$null$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	$null$	$null$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	$null$	$null$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	$null$	$null$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	$null$	$null$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	$null$	$null$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

Figure 4. Sample output for a partial outer join merge

If you are using the anti-join option, the table returns only unmatched records for the first input table.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

Figure 5. Sample output for an anti-join merge

Specifying a Merge Method and Keys

The Merge tab of the Merge node contains the following fields.

Merge Method Select the method to be used for merging records. Selecting either **Keys** or **Condition** activates the bottom half of the dialog box.

- **Order** Merges records by order such that the *n*th record from each input is merged to produce the *n*th output record. When any record runs out of a matching input record, no more output records are produced. This means that the number of records that are created is the number of records in the smallest dataset.
- **Keys** Uses a key field, such as *Transaction ID*, to merge records with the same value in the key field. This is equivalent to a database "equi-join." If a key value occurs more than once, all possible combinations are returned. For example, if records with the same key field value *A* contain differing values *B*, *C*, and *D* in other fields, the merged fields produce a separate record for each combination of *A* with value *B*, *A* with value *C*, and *A* with value *D*.

  Note: Null values are not considered identical in the merge-by-key method and will not join.
- **Condition** Use this option to specify a condition for the merge. For more information, see “Specifying Conditions for a Merge” on page 78.
- **Ranked condition** Use this option to specify whether each row pairing in the primary and all secondary data sets are to be merged; use the ranking expression to sort any multiple matches into order from low to high. For more information, see “Specifying Ranked Conditions for a Merge” on page 78.

Possible keys Lists only those fields with exactly matching field names in all input data sources. Select a field from this list and use the arrow button to add it as a key field used for merging records. More than one key field can be used. You can rename non-matching input fields by using a Filter node, or the Filter tab of a source node.

Keys for merge Lists all fields that are used to merge records from all input data sources based on values of the key fields. To remove a key from the list, select one and use the arrow button to return it to the Possible Keys list. When more than one key field is selected, the option below is enabled.

Combine duplicate key fields When more than one key field is selected above, this option ensures that there is only one output field of that name. This option is enabled by default except in the case when streams have been imported from earlier versions of IBM SPSS Modeler. When this option is disabled, duplicate key fields must be renamed or excluded by using the Filter tab in the Merge node dialog box.

Include only matching records (inner join) Select to merge only complete records.

Include matching and non-matching records (full outer join) Select to perform a "full outer join." This means that if values for the key field are not present in all input tables, the incomplete records are still retained. The undefined value (`$null$`) is added to the key field and included in the output record.

Include matching and selected non-matching records (partial outer join) Select to perform a "partial outer join" of the tables you select in a subdialog box. Click **Select** to specify tables for which incomplete records will be retained in the merge.

Include records in the first dataset not matching any others (anti-join) Select to perform a type of "anti-join," where only nonmatching records from the first dataset are passed downstream. You can specify the order of input datasets using arrows on the Inputs tab. This type of join does not include complete records in the output dataset. For more information, see “Types of Joins” on page 75.

Selecting Data for Partial Joins

For a partial outer join, you must select the table(s) for which incomplete records will be retained. For example, you may want to retain all records from a Customer table while retaining only matched records from the Mortgage Loan table.

Outer Join column. In the *Outer Join* column, select datasets to include in their entirety. For a partial join, overlapping records will be retained as well as incomplete records for datasets selected here. See the topic “Types of Joins” on page 75 for more information.

Specifying Conditions for a Merge

By setting the merge method to **Condition**, you can specify one or more conditions to be satisfied for the merge to take place.

You can either enter the conditions directly into the Condition field, or build them with the aid of the Expression Builder by clicking the calculator symbol to the right of the field.

Add tags to duplicate field names to avoid merge conflicts If two or more of the datasets to be merged contain the same field names, select this check box to add a different prefix tag to the start of the field column headers. For example, if there are two fields called *Name* the result of the merge would contain *1_Name* and *2_Name*. If the tag has been renamed in the data source, the new name is used instead of the numbered prefix tag. If you do not select this check box, and there are duplicate names in the data, a warning is displayed to the right of the check box.

Specifying Ranked Conditions for a Merge

A Ranked Condition merge can be considered as a left sided outer join merge by condition; the left side of the merge is the primary data set where each record is an event. For example, in a model that is used to find patterns in crime data, each record in the primary data set would be a crime and its associated information (location, type, and so on). In this example, the right side might contain the relevant geospatial data sets.

The merge uses both a merge condition and a ranking expression. The merge condition can use a geospatial function such as *within* or *close_to*. During the merge, all of the fields in the right side data sets are added to the left side data set but multiple matches result in a list field. For example:

- Left side: Crime data
- Right side: Counties data set and roads data set
- Merge conditions: Crime data *within* counties and *close_to* roads, along with a definition of what counts as *close_to*.

In this example, if a crime occurred within the required *close_to* distance of three roads (and the number of matches to be returned is set to at least three), then all three roads are returned as a list item.

By setting the merge method to **Ranked condition**, you can specify one or more conditions to be satisfied for the merge to take place.

Primary dataset Select the primary data set for the merge; the fields from all other data sets are added to the data set you select. This can be considered as the left side of an outer join merge.

When you select a primary data set, all the other input data sets that are connected to the Merge node are automatically listed in the **Merges** table.

Add tags to duplicate field names to avoid merge conflicts If two or more of the data sets to be merged contain the same field names, select this check box to add a different prefix tag to the start of the field column headers. For example, if there are two fields that are called *Name* the result of the merge would contain *1_Name* and *2_Name*. If the tag is renamed in the data source, the new name is used instead of the numbered prefix tag. If you do not select this check box, and there are duplicate names in the data, a warning is displayed to the right of the check box.

Merges

Dataset

Shows the name of the secondary data sets that are connected as inputs to the Merge node. By default, where there is more than one secondary data set, they are listed in the order in which they were connected to the Merge node.

Merge Condition

Enter the unique conditions for merging each of the data sets in the table with the primary data set. You can either type the conditions directly into the cell, or build them with the aid of the Expression Builder by clicking the calculator symbol to the right of the cell. For example, you might use geospatial

predicates to create a merge condition that places crime data from one data set within the county data of another data set. The default merge condition depends on the geospatial measurement level, as shown in the list below.

- Point, LineString, MultiPoint, MultiLineString - default condition of *close_to*.
- Polygon, MultiPolygon - default condition of *within*.

For more information about these levels, see "Geospatial measurement sublevels" on page 124.

If a data set contains multiple geospatial fields of different types, the default condition that is used depends on the first measurement level that is found in the data, in the following descending order.

- Point
- LineString
- Polygon

Note: Defaults are only available when there is a geospatial data field in the secondary database.

Ranking Expression

Specify an expression to rank the merging of the data sets; this expression is used to sort multiple matches into an order that is based on the ranking criteria. You can either type the conditions directly into the cell, or build them with the aid of the Expression Builder by clicking the calculator symbol to the right of the cell.

Default ranking expressions of distance and area are provided in the Expression Builder and both rank low to high, meaning that, for example, the top match for distance is the smallest value. An example of ranking by distance is when the primary data set contains crimes and their associated location and each other data set contains objects with locations; in this case the distance between the crimes and the objects can be used as a ranking criteria. The default ranking expression depends on the geospatial measurement level, as shown in the list below.

- Point, LineString, MultiPoint, MultiLineString - the default expression is *distance*.
- Polygon, MultiPolygon - the default expression is *area*.

Note: Defaults are only available when there is a geospatial data field in the secondary database.

Number of Matches

Specify the number of matches that are returned, based on the condition and ranking expressions. The default number of matches depend on the geospatial measurement level in the secondary data set, as shown in the list below; however, you can double-click in the cell to enter your own value, up to a maximum of 100.

- Point, LineString, MultiPoint, MultiLineString - default value of 3.
- Polygon, MultiPolygon - default value of 1.
- Data set that contains no geospatial fields - default value of 1.

As an example, if you set up a merge that is based on a **Merge Condition** of *close_to* and a **Ranking Expression** of *distance*, the top three (closest) matches from the secondary data sets to each record in the primary data set are returned as the values in the resultant list field.

Filtering Fields from the Merge Node

Merge nodes include a convenient way of filtering or renaming duplicate fields as a result of merging multiple data sources. Click the **Filter** tab in the dialog box to select filtering options.

The options presented here are nearly identical to those for the Filter node. There are, however, additional options not discussed here that are available on the Filter menu. See the topic "Filtering or renaming fields" on page 133 for more information.

Field. Displays the input fields from currently connected data sources.

Tag. Lists the tag name (or number) associated with the data source link. Click the **Inputs** tab to alter active links to this Merge node.

Source node. Displays the source node whose data is being merged.

Connected node. Displays the node name for the node that is connected to the Merge node. Frequently, complex data mining requires several merge or append operations that may include the same source node. The connected node name provides a way of differentiating these.

Filter. Displays the current connections between input and output field. Active connections show an unbroken arrow. Connections with a red X indicate filtered fields.

Field. Lists the output fields after merging or appending. Duplicate fields are displayed in red. Click in the Filter field above to disable duplicate fields.

View current fields. Select to view information on fields selected to be used as key fields.

View unused field settings. Select to view information on fields that are not currently in use.

Setting Input Order and Tagging

Using the Inputs tab in the Merge and Append node dialog boxes, you can specify the order of input data sources and make any changes to the tag name for each source.

Tags and order of input datasets. Select to merge or append only complete records.

- **Tag.** Lists current tag names for each input data source. Tag names, or **tags**, are a way of uniquely identifying the data links for the merge or append operation. For example, imagine water from various pipes that is combined at one point and flows through a single pipe. Data in IBM SPSS Modeler flows similarly, and the merging point is often a complex interaction between the various data sources. Tags provide a way of managing the inputs ("pipes") to a Merge or Append node so that if the node is saved or disconnected, the links remain and are easily identifiable.

When you connect additional data sources to a Merge or Append node, default tags are automatically created using numbers to represent the order in which you connected the nodes. This order is unrelated to the order of fields in the input or output datasets. You can change the default tag by entering a new name in the *Tag* column.

- **Source Node.** Displays the source node whose data is being combined.
- **Connected Node.** Displays the node name for the node that is connected to the Merge or Append node. Frequently, complex data mining requires several merge operations that may include the same source node. The connected node name provides a way of differentiating these.
- **Fields.** Lists the number of fields in each data source.

View current tags. Select to view tags that are actively being used by the Merge or Append node. In other words, current tags identify links to the node that have data flowing through. Using the pipe metaphor, current tags are analogous to pipes with existing water flow.

View unused tag settings. Select to view tags, or links, that were previously used to connect to the Merge or Append node but are not currently connected with a data source. This is analogous to empty pipes still intact within a plumbing system. You can choose to connect these "pipes" to a new source or remove them. To remove unused tags from the node, click **Clear**. This clears all unused tags at once.

Merge Optimization Settings

The system provides two options that can help you merge data more efficiently in certain situations. These options allow you to optimize merging when one input dataset is significantly larger than the other datasets or when your data are already sorted by all or some of the key fields that you are using for the merge.

Note: Optimizations from this tab only apply to IBM SPSS Modeler native node execution only; that is, when the Merge node does not pushback to SQL. Optimization settings have no effect on SQL generation.

One input dataset is relatively large. Select to indicate that one of the input datasets is much larger than the others. The system will cache the smaller datasets in memory and then perform the merge by processing the large dataset without caching or sorting it. You will commonly use this type of join with data designed using a star-schema or similar design, where there is a large central table of shared data

(for example, in transactional data). If you select this option, click **Select** to specify the large dataset. Note that you can select only *one* large dataset. The following table summarizes which joins can be optimized using this method.

Table 16. Summary of join optimizations

Type of Join	Can be optimized for a large input dataset?
Inner	Yes
Partial	Yes, if there are no incomplete records in the large dataset.
Full	No
Anti-join	Yes, if the large dataset is the first input.

All inputs are already sorted by key field(s). Select to indicate that the input data are already sorted by one or more of the key fields that you are using for the merge. Make sure *all* your input datasets are sorted.

Specify existing sort order. Specify the fields that are already sorted. Using the Select Fields dialog box, add fields to the list. You can select from only the key fields that are being used for the merge (specified in the Merge tab). In the *Order* column, specify whether each field is sorted in ascending or descending order. If you are specifying multiple fields, make sure that you list them in the correct sorting order. Use the arrows to the right of the list to arrange the fields in the correct order. If you make a mistake in specifying the correct existing sort order, an error will appear when you run the stream, displaying the record number where the sorting is inconsistent with what you specified.

Depending on the case sensitivity of the collation method used by the database, optimization may not function correctly where one or more inputs are sorted by the database. For example, if you have two inputs where one is case sensitive and the other is case insensitive, the results of sorting could be different. Merge optimization causes records to be processed using their sorted order. As a result, if inputs are sorted using different collation methods, the Merge node reports an error and displays the record number where sorting is inconsistent. When all inputs are from one source, or are sorted using mutually inclusive collations, records can be merged successfully.

Note: Merging speed may benefit from enabling parallel processing.

Append Node

You can use Append nodes to concatenate sets of records. Unlike Merge nodes, which join records from different sources together, Append nodes read and pass downstream all of the records from one source until there are no more. Then the records from the next source are read using the same data structure (number of records, number of fields, and so on) as the first, or primary, input. When the primary source has more fields than another input source, the system null string (`$null$`) will be used for any incomplete values.

Append nodes are useful for combining datasets with similar structures but different data. For example, you might have transaction data stored in different files for different time periods, such as a sales data file for March and a separate one for April. Assuming that they have the same structure (the same fields in the same order), the Append node will join them together into one large file, which you can then analyze.

Note: In order to append files, the field measurement levels must be similar. For example, a *Nominal* field cannot be appended with a field whose measurement level is *Continuous*.

Setting Append Options

Match fields by. Select a method to use when matching fields to append.

- **Position.** Select to append datasets based on the position of fields in the main data source. When using this method, your data should be sorted to ensure proper appending.

- **Name.** Select to append datasets based on the name of fields in the input datasets. Also select **Match case** to enable case sensitivity when matching field names.

Output Field. Lists the source nodes that are connected to the Append node. The first node on the list is the primary input source. You can sort the fields in the display by clicking on the column heading. This sorting does not actually reorder the fields in the dataset.

Include fields from. Select **Main dataset only** to produce output fields based on the fields in the main dataset. The main dataset is the first input, specified on the Inputs tab. Select **All datasets** to produce output fields for all fields in all datasets regardless of whether there is a matching field across all input datasets.

Tag records by including source dataset in field. Select to add an additional field to the output file whose values indicate the source dataset for each record. Specify a name in the text field. The default field name is *Input*.

Distinct node

Duplicate records in a data set must be removed before data mining can begin. For example, in a marketing database, individuals may appear multiple times with different address or company information. You can use the Distinct node to find or remove duplicate records in your data, or to create a single, composite record from a group of duplicate records.

To use the Distinct node, you must first define a set of key fields that determine when two records are considered to be duplicates.

If you do not pick all your fields as key fields, then two "duplicate" records may not be truly identical because they can still differ in the values of the remaining fields. In this case, you can also define a sort order that is applied within each group of duplicate records. This sort order gives you fine control over which record is treated as the first within a group. Otherwise, all duplicates are considered to be interchangeable and any record might be selected. The incoming order of the records is not taken into account, so it does not help to use an upstream Sort node (see "Sorting records within the distinct node" below).

Mode. Specify whether to create a composite record, or to either include or exclude (discard) the first record.

- **Create a composite record for each group.** Provides a way for you to aggregate non-numeric fields. Selecting this option makes the Composite tab available where you specify how to create the composite records. See “Distinct Composite Settings” on page 84 for more information.
- **Include only the first record in each group.** Selects the first record from each group of duplicate records and discards the rest. The *first* record is determined by the sort order defined below, and not by the incoming order of the records.
- **Discard only the first record in each group.** Discards the first record from each group of duplicate records and selects the remainder instead. The *first* record is determined by the sort order defined below, and not by the incoming order of the records. This option is useful for *finding* duplicates in your data so that you can examine them later in the stream.

Key fields for grouping. Lists the field or fields used to determine whether records are identical. You can:

- Add fields to this list using the field picker button on the right.
- Delete fields from the list by using the red X (remove) button.

Within groups, sort records by. Lists the fields used to determine how records are sorted within each group of duplicates, and whether they are sorted in ascending or descending order. You can:

- Add fields to this list using the field picker button on the right.
- Delete fields from the list by using the red X (remove) button.
- Move fields using the up or down buttons, if you are sorting by more than one field.

You must specify a sort order if you have chosen to include or exclude the first record in each group, and it matters to you which record is treated as the first.

Default sort order. Specify whether, by default, records are sorted in **Ascending** or **Descending** order of the sort key values.

Sorting records within the Distinct node

If the order of records within a group of duplicates is important to you, then you must specify the order using the **Within groups, sort records by** option in the Distinct node. Do not rely on an upstream Sort node. Remember that the incoming order of the records is not taken into account -- only the order specified within the node.

If you do not specify any sort fields (or you specify insufficient sort fields), then the records within each group of duplicates will be unordered (or incompletely ordered) and the results may be unpredictable.

For example, assume we have a very large set of log records pertaining to a number of machines. The log contains data such as the following:

Table 17. Machine log data

Timestamp	Machine	Temperature
17:00:22	Machine A	31
13:11:30	Machine B	26
16:49:59	Machine A	30
18:06:30	Machine X	32
16:17:33	Machine A	29
19:59:04	Machine C	35
19:20:55	Machine Y	34
15:36:14	Machine X	28
12:30:41	Machine Y	25
14:45:49	Machine C	27
19:42:00	Machine B	34
20:51:09	Machine Y	36
19:07:23	Machine X	33

To reduce the number of records down to the latest record for each machine, use `Machine` as the key field and use `Timestamp` as the sort field (in descending order). The input order does not affect the result because the sort selection specifies which of the many rows for a given Machine is to be returned, and the final data output would be as follows.

Table 18. Sorted machine log data

Timestamp	Machine	Temperature
17:00:22	Machine A	31
19:42:00	Machine B	34
19:59:04	Machine C	35
19:07:23	Machine X	33
20:51:09	Machine Y	36

Distinct Optimization Settings

If the data on which you are working has only a small number of records, or has already been sorted, you can optimize the way in which it is handled to enable IBM SPSS Modeler to process the data more efficiently.

Note: If you either select **Input dataset has a low number of distinct keys**, or use SQL generation for the node, any row within the distinct key value can be returned; to control which row is returned within a distinct key you need to specify the sort order by using the **Within groups, sort records by** fields on the Settings tab. The optimization options do not affect the results output by the Distinct node as long as you have specified a sort order on the Settings tab.

Input dataset has a low number of distinct keys. Select this option if you have a small number of records, or a small number of unique values of the key field(s), or both. Doing so can improve performance.

Input dataset is already ordered by grouping fields and sorting fields on the Settings tab. Only select this option if your data is already sorted by all of the fields listed under **Within groups, sort records by** on the Settings tab, and if the ascending or descending sort order of the data is the same. Doing so can improve performance.

Disable SQL generation. Select this option to disable SQL generation for the node.

Distinct Composite Settings

If the data on which you are working has multiple records, for example for the same person, you can optimize the way in which the data is handled by creating a single composite, or aggregate, record to process.

Note: This tab is only available when you select **Create a composite record for each group** on the Settings tab.

Setting options for the Composite tab

Field. This column shows all fields, except key fields in the data model, in their natural sort order. If the node is not connected, no fields are shown. To sort the rows alphabetically by field name, click the column header. You can select more than one row using either Shift-click or Ctrl-click. In addition, if you right-click a field, a menu is displayed from which you can choose to select all rows, sort the rows by ascending or descending field name or value, select fields by either measure or storage type, or select a value to automatically add the same **Fill with values based on** entry to every selected row.

Fill with values based on. Select the value type to be used for the composite record for the **Field**. The available options depend on the field type.

- For numeric range fields you can choose from:
  - First record in group
  - Last record in group
  - Total
  - Mean
  - Minimum
  - Maximum
  - Custom
- For time or date fields you can choose from:
  - First record in group
  - Last record in group
  - Earliest
  - Most Recent

  - Custom
- For string or typeless fields you can choose from:
  - First record in group
  - Last record in group
  - First alphanumeric
  - Last alphanumeric
  - Custom

In each case, you can use the **Custom** option to exercise more control over which value is used to fill the composite record. See "Distinct Composite - Custom Tab" on page 85 for more information.

Include record count in field. Select this option to include an extra field in each output record, called `Record_Count` by default. This field indicates how many input records were aggregated to form each aggregate record. To create a custom name for this field, type your entry in the edit field.

Distinct Composite - Custom Tab

The Custom Fill dialog box gives you more control over which value is used to complete the new composite record. Note that must instantiate your data first before using this option if only customizing a single Field row on the Composite tab.

Note: This dialog box is only available when you select the Custom value in the **Fill with values based on** column on the Composite tab.

Depending on the field type, you can choose from one of the following options.

- **Select by frequency.** Choose a value based on the frequency with which it occurs in the data record.

  Note: Not available for Fields with a with a type of Continuous, Typeless, or Date/Time.

  - **Use.** Select from either Most or Least Frequent.
  - **Ties.** If there are two or more records with the same frequency of occurrence, specify how to select the required record. You can choose from one of four options: Use First, Use Last, Use Lowest, or Use Highest.
- **Includes value (T/F).** Select this to convert a field to a flag which identifies if any of the records in a group has a specified value. You can then select the **Value** from the list of those for the selected field.

  Note: Not available if you select more than one Field row on the Composite tab
- **First match in list.** Select this to prioritize which value to give to the composite record. You can then select one of the **Items** from the list of those for the selected field.

  Note: Not available if you select more than one Field row on the Composite tab
- **Concatenate values.** Select this to retain all the values in a group by concatenating them into a string. You must specify a delimiter to be used between each value.

  Note: This is the only option available if you select one or more Field rows with a type of Continuous, Typeless, or Date/Time.
- **Use delimiter.** You can choose to use a **Space** or **Comma** as a delimiter value in the concatenated string. Alternatively, in the **Other** field, you can enter your own delimiter value character.

  Note: Only available if you select the **Concatenate values** option.

Streaming Time Series node

You use the Streaming Time Series node to build and score time series models in one step. A separate time series model is built for each target field, however model nuggets are not added to the generated models palette and the model information cannot be browsed.

Methods for modeling time series data require a uniform interval between each measurement, with any missing values indicated by empty rows. If your data does not already meet this requirement, you will have to transform values as needed.

Other points to note in connection with Time Series nodes are:

- Fields must be numeric.
- Date fields cannot be used as inputs.
- Partitions are ignored.

The Streaming Time Series node estimates exponential smoothing, univariate Autoregressive Integrated Moving Average (ARIMA), and multivariate ARIMA (or transfer function) models for time series and produces forecasts based on the time series data. Also available is an Expert Modeler, which attempts to automatically identify and estimate the best-fitting ARIMA or exponential smoothing model for one or more target fields.

For more information about time series modeling, see the Time Series Models section of the SPSS Modeler Modeling Nodes guide.

The Streaming Time Series node is supported for use in a streaming deployment environment, through IBM SPSS Modeler Solution Publisher, using the IBM SPSS Collaboration and Deployment Services Scoring Service.

Streaming Time Series node - field options

Use predefined roles This option uses the role settings (targets, predictors and so on) from an upstream Type node (or the Types tab of an upstream source node).

Use custom field assignments. To manually assign targets, predictors and other roles, select this option.

Note: If you have partitioned your data, the partitions are taken into account if you select **Use predefined roles**, but not if you select **Use custom field assignments**.

Fields. Use the arrow buttons to assign items manually from this list to the various role fields on the right of the screen. The icons indicate the valid measurement levels for each role field.

To select all the fields in the list, click the **All** button, or click an individual measurement level button to select all fields with that measurement level.

Targets Select one or more of the fields as your target for the prediction.

Candidate inputs Select one or more fields as inputs for the prediction.

Events and Interventions Use this area to designate certain input fields as event or intervention fields. This designation identifies a field as containing time series data that can be affected by events (predictable recurring situations; for example, sales promotions) or interventions (one-time incidents; for example, power outage or employee strike).

Streaming Time Series node - data specification options

The Data Specifications tab is where you set all the options for the data to be included in your model. As long as you specify both a **Date/time field** and **Time interval**, you can click the **Run** button to build a model with all the default options, but normally you will want to customize the build for your own purposes.

The tab contains several different panes on which you set the customizations that are specific to your model.

Streaming Time Series node - observations

Use the settings in this pane to specify the fields that define the observations.

Observations that are specified by a date/time field

You can specify that the observations are defined by a date, time, or timestamp field. In addition to the field that defines the observations, select the appropriate time interval that describes the observations. Depending on the specified time interval, you can also specify other settings, such as the interval between observations (increment) or the number of days per week. The following considerations apply to the time interval:

- Use the value **Irregular** when the observations are irregularly spaced in time, such as the time at which a sales order is processed. When **Irregular** is selected, you must specify the time interval that is used for the analysis, from the **Time Interval** settings on the Data Specifications tab.
- When the observations represent a date and time and the time interval is hours, minutes, or seconds, then use **Hours per day**, **Minutes per day**, or **Seconds per day**. When the observations represent a time (duration) without reference to a date and the time interval is hours, minutes, or seconds, then use **Hours (non-periodic)**, **Minutes (non-periodic)**, or **Seconds (non-periodic)**.
- Based on the selected time interval, the procedure can detect missing observations. Detecting missing observations is necessary since the procedure assumes that all observations are equally spaced in time and that no observations are missing. For example, if the time interval is Days and the date 2015-10-27 is followed by 2015-10-29, then an observation is missing for 2015-10-28. Values are imputed for any missing observations; use the **Missing Value Handling** area of the Data Specifications tab to specify settings for handling missing values.
- The specified time interval allows the procedure to detect multiple observations in the same time interval that need to be aggregated together and to align observations on an interval boundary, such as the first of the month, to ensure that the observations are equally spaced. For example, if the time interval is Months, then multiple dates in the same month are aggregated together. This type of aggregation is referred to as *grouping*. By default, observations are summed when grouped. You can specify a different method for grouping, such as the mean of the observations, from the **Aggregation and Distribution** settings on the Data Specifications tab.
- For some time intervals, the additional settings can define breaks in the normal equally spaced intervals. For example, if the time interval is Days, but only weekdays are valid, you can specify that there are five days in a week, and the week begins on Monday.

Observations that are defined as periods or cyclic periods

Observations can be defined by one or more integer fields that represent periods or repeating cycles of periods, up to an arbitrary number of cycle levels. With this structure, you can describe series of observations that don't fit one of the standard time intervals. For example, a fiscal year with only 10 months can be described with a cycle field that represents years and a period field that represents months, where the length of one cycle is 10.

Fields that specify cyclic periods define a hierarchy of periodic levels, where the lowest level is defined by the **Period** field. The next highest level is specified by a cycle field whose level is 1, followed by a cycle field whose level is 2, and so on. Field values for each level, except the highest, must be periodic with respect to the next highest level. Values for the highest level cannot be periodic. For example, in the case of the 10-month fiscal year, months are periodic within years and years are not periodic.

- The length of a cycle at a particular level is the periodicity of the next lowest level. For the fiscal year example, there is only one cycle level and the cycle length is 10 since the next lowest level represents months and there are 10 months in the specified fiscal year.
- Specify the starting value for any periodic field that does not start from 1. This setting is necessary for detecting missing values. For example, if a periodic field starts from 2 but the starting value is specified as 1, then the procedure assumes that there is a missing value for the first period in each cycle of that field.

Streaming Time Series node - time interval for analysis

The time interval that you use for analysis can differ from the time interval of the observations. For example, if the time interval of the observations is Days, you might choose Months for the time interval for

analysis. The data is then aggregated from daily to monthly data before the model is built. You can also choose to distribute the data from a longer to a shorter time interval. For example, if the observations are quarterly then you can distribute the data from quarterly to monthly data.

Use the settings in this pane to specify the time interval for the analysis. The method by which the data is aggregated or distributed is specified from the **Aggregation and Distribution** settings on the Data Specifications tab.

The available choices for the time interval at which the analysis is done depend on how the observations are defined and the time interval of those observations. In particular, when the observations are defined by cyclic periods, only aggregation is supported. In that case, the time interval of the analysis must be greater than or equal to the time interval of the observations.

Streaming Time Series node - aggregation and distribution options

Use the settings in this pane to specify settings for aggregating or distributing the input data with respect to the time intervals of the observations.

Aggregation functions
When the time interval that is used for the analysis is longer than the time interval of the observations, the input data are aggregated. For example, aggregation is done when the time interval of the observations is Days and the time interval for analysis is Months. The following aggregation functions are available: mean, sum, mode, min, or max.

Distribution functions
When the time interval that is used for the analysis is shorter than the time interval of the observations, the input data are distributed. For example, distribution is done when the time interval of the observations is Quarters and the time interval for analysis is Months. The following distribution functions are available: mean or sum.

Grouping functions
Grouping is applied when observations are defined by date/times and multiple observations occur in the same time interval. For example, if the time interval of the observations is Months, then multiple dates in the same month are grouped and associated with the month in which they occur. The following grouping functions are available: mean, sum, mode, min, or max. Grouping is always done when the observations are defined by date/times and the time interval of the observations is specified as Irregular.

Note: Although grouping is a form of aggregation, it is done before any handling of missing values whereas formal aggregation is done after any missing values are handled. When the time interval of the observations is specified as Irregular, aggregation is done only with the grouping function.

Aggregate cross-day observations to previous day
Specifies whether observations with times that cross a day boundary are aggregated to the values for the previous day. For example, for hourly observations with an eight-hour day that starts at 20:00, this setting specifies whether observations between 00:00 and 04:00 are included in the aggregated results for the previous day. This setting applies only if the time interval of the observations is Hours per day, Minutes per day or Seconds per day and the time interval for analysis is Days.

Custom settings for specified fields
You can specify aggregation, distribution, and grouping functions on a field by field basis. These settings override the default settings for the aggregation, distribution, and grouping functions.

Streaming Time Series node - missing value options

Use the settings in this pane to specify how any missing values in the input data are to be replaced with an imputed value. The following replacement methods are available:

Linear interpolation
Replaces missing values by using a linear interpolation. The last valid value before the missing value and the first valid value after the missing value are used for the interpolation. If the first or last observation in the series has a missing value, then the two nearest non-missing values at the beginning or end of the series are used.

Replaces missing values by using a linear interpolation.

- For non-seasonal data, the last valid value before the missing value and the first valid value after the missing value are used for the interpolation. If missing values are at the beginning or the end of a time series, then a linear extrapolation method is used based on the two nearest valid values.
- For seasonal data, a missing value is linearly interpolated using the last valid value of the same period before the missing value and the first valid value of the same period after the missing value. If one of the two values of the same period can't be found for the missing value, then the data will be regarded as non-seasonal data and linear interpolation of non-seasonal data is used to impute the missing value.

Series mean
Replaces missing values with the mean for the entire series.

Mean of nearby points
Replaces missing values with the mean of valid surrounding values. The span of nearby points is the number of valid values before and after the missing value that are used to compute the mean.

Median of nearby points
Replaces missing values with the median of valid surrounding values. The span of nearby points is the number of valid values before and after the missing value that are used to compute the median.

Linear trend
This option uses all non-missing observations in the series to fit a simple linear regression model, which is then used to impute the missing values.

Other settings:

Lowest data quality score (%)
Computes data quality measures for the time variable and for input data corresponding to each time series. If the data quality score is lower that this threshold, the corresponding time series will be discarded.

Streaming Time Series node - estimation period

In the Estimation Period pane, you can specify the range of records to be used in model estimation. By default, the estimation period starts at the time of the earliest observation and ends at the time of the latest observation across all series.

By start and end times
You can specify both the start and end of the estimation period or you can specify just the start or just the end. If you omit the start or the end of the estimation period, the default value is used.

- If the observations are defined by a date/time field, enter values for start and end in the same format that is used for the date/time field.
- For observations that are defined by cyclic periods, specify a value for each of the cyclic periods fields. Each field is displayed in a separate column.

By latest or earliest time intervals

Defines the estimation period as a specified number of time intervals that start at the earliest time interval or end at the latest time interval in the data, with an optional offset. In this context, the time interval refers to the time interval of the analysis. For example, assume that the observations are monthly but the time interval of the analysis is quarters. Specifying **Latest** and a value of 24 for the **Number of time intervals** means the latest 24 quarters.

Optionally, you can exclude a specified number of time intervals. For example, specifying the latest 24 time intervals and 1 for the number to exclude means that the estimation period consists of the 24 intervals that precede the last one.

Streaming Time Series node - build options

The Build Options tab is where you set all the options for building your model. You can, of course, just click the **Run** button to build a model with all the default options, but normally you will want to customize the build for your own purposes.

The tab contains two different panes on which you set the customizations that are specific to your model.

Streaming Time Series node - general build options

The options available on this pane depend on which of the following three settings you choose from the **Method** list:

- **Expert Modeler.** Choose this option to use the Expert Modeler, which automatically finds the best-fitting model for each dependent series.
- **Exponential Smoothing.** Use this option to specify a custom exponential smoothing model.
- **ARIMA.** Use this option to specify a custom ARIMA model.

Expert Modeler

Under **Model Type**, select the type of models you want to build:

- **All models.** The Expert Modeler considers both ARIMA and exponential smoothing models.
- **Exponential smoothing models only.** The Expert Modeler considers only exponential smoothing models.
- **ARIMA models only.** The Expert Modeler considers only ARIMA models.

Expert Modeler considers seasonal models. This option is only enabled if a periodicity is defined for the active dataset. When this option is selected, the Expert Modeler considers both seasonal and nonseasonal models. If this option is not selected, the Expert Modeler considers only nonseasonal models.

Expert Modeler considers sophisticated exponential smoothing models. When this option is selected, the Expert Modeler searches a total of 13 exponential smoothing models (7 of them existed in the original Time Series node, and 6 of them were added in version 18.1). If this option is not selected, the Expert Modeler only searches the original 7 exponential smoothing models.

Under **Outliers**, select from the following options

Detect outliers automatically. By default, automatic detection of outliers is not performed. Select this option to perform automatic detection of outliers, then select the desired outlier types.

Input fields must have a measurement level of *Flag*, *Nominal*, or *Ordinal* and must be numeric (for example, 1/0, not True/False, for a flag field), before they are included in this list.

The Expert Modeler considers only simple regression and not arbitrary transfer functions for inputs that are identified as event or intervention fields on the **Fields** tab.

Exponential Smoothing

Model Type. Exponential smoothing models are classified as either seasonal or nonseasonal.[1] Seasonal models are only available if the periodicity defined by using the Time Intervals pane on the Data Specifications tab is seasonal. The seasonal periodicities are: cyclic periods, years, quarters, months, days per week, hours per day, minutes per day, and seconds per day. The following model types are available:

- **Simple.** This model is appropriate for a series in which there is no trend or seasonality. Its only relevant smoothing parameter is level. Simple exponential smoothing is most similar to an ARIMA with zero orders of autoregression, one order of differencing, one order of moving average, and no constant.

[1] Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **Holt's linear trend.** This model is appropriate for a series in which there is a linear trend and no seasonality. Its relevant smoothing parameters are level and trend, and, in this model, they are not constrained by each other's values. Holt's model is more general than Brown's model but may take longer to compute estimates for large series. Holt's exponential smoothing is most similar to an ARIMA with zero orders of autoregression, two orders of differencing, and two orders of moving average.
- **Damped trend.** This model is appropriate for a series in which there is a linear trend that is dying out and no seasonality. Its relevant smoothing parameters are level, trend, and damping trend. Damped exponential smoothing is most similar to an ARIMA with one order of autoregression, one order of differencing, and two orders of moving average.
- **Multiplicative trend.** This model is appropriate for a series in which there is a trend that changes with the magnitude of the series and no seasonality. Its relevant smoothing parameters are level and trend. Multiplicative trend exponential smoothing is not similar to any ARIMA model.
- **Brown's linear trend.** This model is appropriate for a series in which there is a linear trend and no seasonality. Its relevant smoothing parameters are level and trend, but, in this model, they are assumed to be equal. Brown's model is therefore a special case of Holt's model. Brown's exponential smoothing is most similar to an ARIMA with zero orders of autoregression, two orders of differencing, and two orders of moving average, with the coefficient for the second order of moving average equal to one half of the coefficient for the first order squared.
- **Simple seasonal.** This model is appropriate for a series in which there is no trend and a seasonal effect that is constant over time. Its relevant smoothing parameters are level and season. Seasonal exponential smoothing is most similar to an ARIMA with zero orders of autoregression; one order of differencing; one order of seasonal differencing; and orders 1, *p*, and *p*+1 of moving average, where *p* is the number of periods in a seasonal interval. For monthly data, *p* = 12.
- **Winters' additive.** This model is appropriate for a series in which there is a linear trend and a seasonal effect that is constant over time. Its relevant smoothing parameters are level, trend, and season. Winters' additive exponential smoothing is most similar to an ARIMA with zero orders of autoregression; one order of differencing; one order of seasonal differencing; and *p*+1 orders of moving average, where *p* is the number of periods in a seasonal interval. For monthly data, *p*=12.
- **Damped trend with additive seasonal.** This model is appropriate for a series in which there is a linear trend that is dying out and a seasonal effect that is constant over time. Its relevant smoothing parameters are level, trend, damping trend, and season. Damped trend and additive seasonal exponential smoothing is not similar to any ARIMA model.
- **Multiplicative trend with additive seasonal.** This model is appropriate for a series in which there is a trend that changes with the magnitude of the series and a seasonal effect that is constant over time. Its relevant smoothing parameters are level, trend, and season. Multiplicative trend and additive seasonal exponential smoothing is not similar to any ARIMA model.
- **Multiplicative seasonal.** This model is appropriate for a series in which there is no trend and a seasonal effect that changes with the magnitude of the series. Its relevant smoothing parameters are level and season. Multiplicative seasonal exponential smoothing is not similar to any ARIMA model.
- **Winters' multiplicative.** This model is appropriate for a series in which there is a linear trend and a seasonal effect that changes with the magnitude of the series. Its relevant smoothing parameters are level, trend, and season. Winters' multiplicative exponential smoothing is not similar to any ARIMA model.
- **Damped trend with multiplicative seasonal.** This model is appropriate for a series in which there is a linear trend that is dying out and a seasonal effect that changes with the magnitude of the series. Its relevant smoothing parameters are level, trend, damping trend, and season. Damped trend and multiplicative seasonal exponential smoothing is not similar to any ARIMA model.
- **Multiplicative trend with multiplicative seasonal.** This model is appropriate for a series in which there are a trend and a seasonal effect that both change with the magnitude of the series. Its relevant smoothing parameters are level, trend, and season. Multiplicative trend and multiplicative seasonal exponential smoothing is not similar to any ARIMA model.

Target Transformation. You can specify a transformation to be performed on each dependent variable before it is modeled.

- **None.** No transformation is performed.
- **Square root.** Square root transformation is performed.
- **Natural log.** Natural log transformation is performed.

ARIMA

Specify the structure of a custom ARIMA model.

ARIMA Orders. Enter values for the various ARIMA components of your model into the corresponding cells of the grid. All values must be non-negative integers. For autoregressive and moving average components, the value represents the maximum order. All positive lower orders are included in the model. For example, if you specify 2, the model includes orders 2 and 1. Cells in the Seasonal column are only enabled if a periodicity is defined for the active dataset.

- **Autoregressive (p).** The number of autoregressive orders in the model. Autoregressive orders specify which previous values from the series are used to predict current values. For example, an autoregressive order of 2 specifies that the value of the series two time periods in the past is used to predict the current value.
- **Difference (d).** Specifies the order of differencing applied to the series before estimating models. Differencing is necessary when trends are present (series with trends are typically nonstationary and ARIMA modeling assumes stationarity) and is used to remove their effect. The order of differencing corresponds to the degree of series trend; first-order differencing accounts for linear trends, second-order differencing accounts for quadratic trends, and so on.
- **Moving Average (q).** The number of moving average orders in the model. Moving average orders specify how deviations from the series mean for previous values are used to predict current values. For example, moving-average orders of 1 and 2 specify that deviations from the mean value of the series from each of the last two time periods be considered when predicting current values of the series.

Seasonal. Seasonal autoregressive, moving average, and differencing components play the same roles as their nonseasonal counterparts. For seasonal orders, however, current series values are affected by previous series values that are separated by one or more seasonal periods. For example, for monthly data (seasonal period of 12), a seasonal order of 1 means that the current series value is affected by the series value 12 periods before the current one. A seasonal order of 1, for monthly data, is then the same as specifying a nonseasonal order of 12.

Detect outliers automatically. Select this option to perform automatic detection of outliers, and select one or more of the outlier types available.

Type of Outliers to Detect. Select the outlier type(s) you want to detect. The supported types are:

- Additive (default)
- Level shift (default)
- Innovational
- Transient
- Seasonal additive
- Local trend
- Additive patch

Transfer Function Orders and Transformations. To specify transformations and to define transfer functions for any or all of the input fields in your ARIMA model, click **Set**; a separate dialog box is displayed in which you enter the transfer and transformation details.

Include constant in model. Inclusion of a constant is standard unless you are sure that the overall mean series value is 0. Excluding the constant is recommended when differencing is applied.

Transfer and transformation functions

Use the Transfer Function Orders and Transformations dialog box to specify transformations and to define transfer functions for any or all of the input fields in your ARIMA model.

Target Transformations. In this pane you can specify a transformation to be performed on each target variable before it is modeled.

- **None.** No transformation is performed.
- **Square root.** Square root transformation is performed.
- **Natural log.** Natural log transformation is performed.

Candidate Inputs Transfer functions and Transformation. You use transfer functions to specify the manner in which past values of the input fields are used to forecast future values of the target series. The list on the left hand side of the pane shows all input fields. The remaining information in this pane is specific to the input field you select.

Transfer Function Orders. Enter values for the various components of the transfer function into the corresponding cells of the **Structure.** grid. All values must be non-negative integers. For numerator and denominator components, the value represents the maximum order. All positive lower orders are included in the model. In addition, order 0 is always included for numerator components. For example, if you specify 2 for numerator, the model includes orders 2, 1, and 0. If you specify 3 for denominator, the model includes orders 3, 2, and 1. Cells in the Seasonal column are only enabled if a periodicity is defined for the active dataset.

Numerator. The numerator order of the transfer function specifies which previous values from the selected independent (predictor) series are used to predict current values of the dependent series. For example, a numerator order of 1 specifies that the value of an independent series one time period in the past, in addition to the current value of the independent series, is used to predict the current value of each dependent series.

Denominator. The denominator order of the transfer function specifies how deviations from the series mean, for previous values of the selected independent (predictor) series, are used to predict current values of the dependent series. For example, a denominator order of 1 specifies that deviations from the mean value of an independent series one time period in the past is considered when predicting the current value of each dependent series.

Difference. Specifies the order of differencing applied to the selected independent (predictor) series before estimating models. Differencing is necessary when trends are present and is used to remove their effect.

Seasonal. Seasonal numerator, denominator, and differencing components play the same roles as their nonseasonal counterparts. For seasonal orders, however, current series values are affected by previous series values that are separated by one or more seasonal periods. For example, for monthly data (seasonal period of 12), a seasonal order of 1 means that the current series value is affected by the series value 12 periods prior to the current one. A seasonal order of 1, for monthly data, is then the same as specifying a nonseasonal order of 12.

Delay. Setting a delay causes the input field's influence to be delayed by the number of intervals specified. For example, if the delay is set to 5, the value of the input field at time *t* doesn't affect forecasts until five periods have elapsed (*t* + 5).

Transformation. Specification of a transfer function for a set of independent variables also includes an optional transformation to be performed on those variables.

- **None.** No transformation is performed.
- **Square root.** Square root transformation is performed.
- **Natural log.** Natural log transformation is performed.

Streaming Time Series node - model options

Confidence limit width (%). Confidence intervals are computed for the model predictions and residual autocorrelations. You can specify any positive value less than 100. By default, a 95% confidence interval is used.

The option to **Extend records into the future** sets the number of time intervals to forecast beyond the end of the estimation period. The time interval in this case is the time interval of the analysis, which you

specify on the Data Specifications tab. When forecasts are requested, autoregressive models are automatically built for any input series that are not also targets. These models are then used to generate values for those input series in the forecast period. There is no maximum limit for this setting.

Future Values to Use in Forecasting

- **Compute future values of inputs** If you select this option, the forecast values for predictors, noise predictions, variance estimation, and future time values are calculated automatically. When forecasts are requested, autoregressive models are automatically built for any input series that are not also targets. These models are then used to generate values for those input series in the forecast period.
- **Select fields whose values you wish to add to the data.** For each record that you want to forecast (excluding holdouts), if you are using predictor fields (with the role set to `Input`), you can specify estimated values for the forecast period for each predictor. You can either specify values manually, or choose from a list.
  - **Field.** Click the field selector button and choose any fields that may be used as predictors. Note that fields selected here may or may not be used in modeling; to actually use a field as a predictor, it must be selected in a downstream modeling node. This dialog box simply gives you a convenient place to specify future values so they can be shared by multiple downstream modeling nodes without specifying them separately in each node. Also note that the list of available fields may be constrained by selections on the Build Options tab.

    Note that if future values are specified for a field that is no longer available in the stream (because it has been dropped or because of updated selections made on the Build Options tab), the field is shown in red.
  - **Values.** For each field, you can choose from a list of functions, or click **Specify** to either enter values manually or choose from a list of predefined values. If the predictor fields relate to items that are under your control, or which are otherwise knowable in advance, you should enter values manually. For example, if you are forecasting next month's revenues for a hotel based on the number of room reservations, you could specify the number of reservations you actually have for that period. Conversely, if a predictor field relates to something outside your control, such as a stock price, you could use a function such as the most recent value or the mean of recent points.

    The available functions depend on the measurement level of the field.

Table 19. Functions available for measurement levels

Measurement level	Functions
Continuous or Nominal field	Blank Mean of recent points Most recent value Specify
Flag field	Blank Most recent value True False Specify

Mean of recent points calculates the future value from the mean of the last three data points.

Most recent value sets the future value to that of the most recent data point.

True/False sets the future value of a flag field to True or False as specified.

Specify opens a dialog box for specifying future values manually, or choosing them from a predefined list.

Make Available for Scoring

You can set the default values here for the scoring options that appear on the dialog box for the model nugget.

- **Calculate upper and lower confidence limits.** If selected, this option creates new fields (with the default prefixes $TSLCI- and $TSUCI-) for the lower and upper confidence intervals, for each target field.
- **Calculate noise residuals.** If selected, this option creates a new field (with the default prefix $TSResidual-) for the model residuals for each target field, together with a total of these values.

Model Settings

Maximum number of models to be displayed in output. Specify the maximum number of models you want to include in the output. Note that if the number of models built exceeds this threshold, the models are not shown in the output but they're still available for scoring. Default value is 10. Displaying a large number of models may result in poor performance or instability.

SMOTE node

The Synthetic Minority Over-sampling Technique (SMOTE) node provides an over-sampling algorithm to deal with imbalanced data sets. It provides an advanced method for balancing data. The SMOTE process node is implemented in Python and requires the `imbalanced-learn`© Python library. For details about the imbalanced-learn library, see https://imbalanced-learn.org/stable/[1].

The Python tab on the Nodes Palette contains the SMOTE node and other Python nodes.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

SMOTE node Settings

Define the following settings on the SMOTE node's **Settings** tab.

Target Setting

Target Field. Select the target field. All flag, nominal, ordinal, and discrete measurement types are supported. If the **Use partitioned data** option is selected in the Partition section, only training data will be over-sampled.

Over Sample Ratio

Select **Auto** to automatically select an over-sample ratio, or select **Set Ratio (minority over majority)** to set a custom ratio value. The ratio is the number of samples in the minority class over the number of samples in the majority class. The value must be greater than **0** and less than or equal to **1**.

Random Seed

Set random seed. Select this option and click **Generate** to generate the seed used by the random number generator.

Methods

Algorithm Kind. Select the type of SMOTE algorithm you wish to use.

Samples Rules

K Neighbours. Specify the number of the nearest neighbors to use for constructing synthetic samples

M Neighbours. Specify the number of nearest neighbors to use for determining if a minority sample is in danger. This will only be used if the **Borderline1** or **Borderline2** SMOTE algorithm type is selected.

Partition

Use partitioned data. Select this option if you only want training data to be over-sampled.

The SMOTE node requires the `imbalanced-learn`© Python library. The following table shows the relationship between the settings in the SPSS Modeler SMOTE node dialog and the Python algorithm.

Table 20. Node properties mapped to Python library parameters

SPSS Modeler setting	Script name (property name)	Python API parameter name
Over sample ratio (number input control)	`sample_ratio_value`	`ratio`
Random seed	`random_seed`	`random_state`
K_Neighbours	`k_neighbours`	`k`
M_Neighbours	`m_neighbours`	`m`
Algorithm kind	`algorithm_kind`	`kind`

Extension Transform node

With the Extension Transform node, you can take data from an IBM SPSS Modeler stream and apply transformations to the data using R scripting or Python for Spark scripting. When the data has been modified, it is returned to the stream for further processing, model building and model scoring. The Extension Transform node makes it possible to transform data using algorithms that are written in R or Python for Spark, and enables the user to develop data transformation methods that are tailored to a particular problem.

To use this node with R, you must install IBM SPSS Modeler - Essentials for R. See the *IBM SPSS Modeler - Essentials for R: Installation Instructions* for installation instructions and compatibility information. You must also have a compatible version of R installed on your computer.

Extension Transform node - Syntax tab

Select your type of syntax – **R** or **Python for Spark**. See the following sections for more information. When your syntax is ready, you can click **Run** to execute the Extension Transform node.

R Syntax

R Syntax. You can enter, or paste, custom R scripting syntax for data analysis into this field.

Convert flag fields. Specifies how flag fields are treated. There are two options: **Strings to factor, Integers and Reals to double**, and **Logical values (True, False)**. If you select **Logical values (True, False)** the original values of the flag fields are lost. For example, if a field has values Male and Female, these are changed to True and False.

Convert missing values to the R 'not available' value (NA). When selected, any missing values are converted to the R NA value. The value NA is used by R to identify missing values. Some R functions that you use might have an argument that can be used to control how the function behaves when the data contain NA. For example, the function might allow you to choose to automatically exclude records that contain NA. If this option is not selected, any missing values are passed to R unchanged, and might cause errors when your R script is executed.

Convert date/time fields to R classes with special control for time zones. When selected, variables with date or datetime formats are converted to R date/time objects. You must select one of the following options:

- **R POSIXct.** Variables with date or datetime formats are converted to R `POSIXct` objects.
- **R POSIXlt (list).** Variables with date or datetime formats are converted to R `POSIXlt` objects.

Note: The POSIX formats are advanced options. Use these options only if your R script specifies that datetime fields are treated in ways that require these formats. The POSIX formats do not apply to variables with time formats.

Python Syntax

Python Syntax. You can enter, or paste, custom Python scripting syntax for data analysis into this field. For more information about Python for Spark, see Python for Spark and Scripting with Python for Spark.

Extension Transform node - Console Output tab

The **Console Output** tab contains any output that is received when the R script or Python for Spark script on the Syntax tab runs (for example, if using an R script, it shows output received from the R console when the R script in the **R Syntax** field on the **Syntax** tab is executed). This output might include R or Python error messages or warnings that are produced when the R or Python script is executed. The output can be used, primarily, to debug the script. The **Console Output** tab also contains the script from the **R Syntax** or **Python Syntax** field.

Every time the Extension Transform script is executed, the content of the **Console Output** tab is overwritten with the output received from the R console or Python for Spark. The output cannot be edited.

Space-Time-Boxes Node

Space-Time-Boxes (STB) are an extension of Geohashed spatial locations. More specifically, an STB is an alphanumeric string that represents a regularly shaped region of space and time.

For example, the STB **dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00** is made up of the following three parts:

- The geohash **dr5ru7**
- The start timestamp **2013-01-01 00:00:00**
- The end timestamp **2013-01-01 00:15:00**

As an example, you could use space and time information to improve confidence that two entities are the same because they are virtually in the same place at the same time. Alternatively, you could improve the accuracy of relationship identification by showing that two entities are related due to their proximity in space and time.

You can choose the **Individual Records** or **Hangouts** mode as appropriate for your requirements. Both modes require the same basic details, as follows:

Latitude field. Select the field that identifies the latitude (in WGS84 coordinate system).

Longitude field. Select the field that identifies the longitude (in WGS84 coordinate system).

Timestamp field. Select the field that identifies the time or date.

Individual Record Options

Use this mode to add an additional field to a record to identify its location at a given time.

Derive. Select one or more densities of space and time from which to derive the new field. See “Defining Space-Time-Box density” on page 99 for more information.

Field name extension. Type the extension that you would like added to the new field name(s). You can choose to add this extension as either a **Suffix** or **Prefix**.

Hangout Options

A hangout can be thought of as a location and/or time in which an entity is continually or repeatedly found. For example, this could be used to identify a vehicle that makes regular transportation runs and identify any deviations from the norm.

The hangout detector monitors the movement of entities and flags conditions where an entity is observed to be "hanging out" in the area. The hangout detector automatically assigns each flagged hangout to one or more STBs, and uses in-memory entity and event tracking to detect hangouts with optimum efficiency.

STB Density. Select the density of space and time from which to derive the new field. For example, a value of **STB_GH4_10MINS** would correspond to a four-character geohash box of size approximately 20 km by 20 km and a 10-minute time window. See “Defining Space-Time-Box density” on page 99 for more information.

Entity ID field. Select the entity to be used as the hangout identifier. This ID field identifies the event.

Minimum number of events. An event is a row in the data. Select the minimum number of occurrences of an event for the entity to be considered to be hanging out. A hangout must also qualify based on the following **Dwell time is at least** field.

Dwell time is at least. Specify the minimum duration over which the entity must dwell in the same location. This can help exclude, for example, a car waiting at a traffic light from being considered as hanging out. A hangout must also qualify based on the previous **Minimum number of events** field.

Following is more detail about what qualifies as a hangout:

Let $e_1, \ldots, e_n$ denote all time ordered events that are received from a given entity ID during a time duration (t_1, t_n). These events qualify as a hangout if:

- n >= *minimum number of events*
- $t_n - t_1$ >= *minimum dwell time*
- All events $e_1, \ldots, e_n$ occur in the same STB

Allow hangouts to span STB boundaries. If this option is selected the definition of a hangout is less strict and could include, for example, an entity that hangs out in more than one Space-Time-Box. For example, if your STBs are defined as whole hours, selecting this option would recognize an entity that hangs out for an hour as valid, even if the hour consisted of the 30 minutes before midnight and the 30 minutes after midnight. If this option is not selected, 100% of hangout time must be within a single Space-Time-Box.

Min proportion of events in qualifying timebox (%). Only available if **Allow hangouts to span STB boundaries** is selected. Use this to control the degree to which a hangout reported in one STB might in fact overlap another. Select the minimum proportion of events that must occur within a single STB to identify a hangout. If set to 25%, and the proportion of events is 26%, this qualifies as being a hangout.

For example, suppose you configure the hangout detector to require at least two events (minimum number of events = 2) and a contiguous hover time of at least 2 minutes in a 4-byte-geohash space box and a 10-minute time box (STB_NAME = STB_GH4_10MINS). When a hangout is detected, say the entity hovers in the same 4-byte-geohash space box while the three qualifying events occur within a 10-minute time span between 4:57pm and 5:07pm at 4:58pm, 5:01pm, and 5:03pm. The qualifying timebox percent value specifies the STBs that be credit for the hangout, as follows:

- **100%.** Hangout is reported in the 5:00 - 5:10pm time-box and not in the 4:50 - 5:00pm time-box (events 5:01pm and 5:03pm meet all conditions that are required for a qualifying hangout and 100% of these events occurred in the 5:00 - 5:10 time box).
- **50%.** Hangouts in both the time-boxes are reported (events 5:01pm and 5:03pm meet all conditions that are required for a qualifying hangout and at least 50% of these events occurred in the 4:50 - 5:00 time box and at least 50% of these events occurred in the 5:00 - 5:10 time box).
- **0%.** Hangouts in both the time-boxes are reported.

When 0% is specified, hangout reports include the STBs representing every time box that is touched by the qualifying duration. The qualifying duration needs to be less than or equal to the corresponding

duration of the time box in the STB. In other words, there should never be a configuration where a 10-minute STB is configured in tandem with a 20-minute qualifying duration.

A hangout is reported as soon as the qualifying conditions are met, and is not reported more than once per STB. Suppose that three events qualify for a hangout, and 10 total events occur within a qualifying duration all within the same STB. In that case, the hangout is reported when the third qualifying event occurs. None of the additional seven events trigger a hangout report.

Note:

- The hangout detector's in-memory event data is not shared across processes. Therefore, a particular entity has affinity with a particular hangout detector node. That is, incoming motion data for an entity must always be consistently passed to the hangout detector node tracking that entity, which is ordinarily the same node throughout the run.
- The hangout detector's in-memory event data is volatile. Whenever the hangout detector is exited and restarted, any work-in-progress hangouts are lost. This means stopping and restarting the process might cause the system to miss reporting real hangouts. A potential remedy involves replaying some of the historical motion data (for example, going back 48 hours and replaying the motion records that are applicable to any node that was restarted).
- The hangout detector must be fed data in time-sequential order.

Defining Space-Time-Box density

Choose the size (density) of your Space-Time-Boxes (STB) by specifying the physical area and elapsed time to be included in each.

Geo density. Select the size of the area to be included in each STB.

Time interval. Select the number of hours to be included in each STB.

Field name. Prefixed with STB, this is automatically completed based on your selections in the preceding two fields.

Streaming TCM Node

The Streaming TCM node can be used to build and score temporal causal models in one step.

For more information about temporal causal modeling, see the Temporal Causal Models topic in the Time Series Models section of the SPSS Modeler Modeling Nodes guide.

Streaming TCM Node - Time Series options

On the Fields tab, use the **Time Series** settings to specify the series to include in the model system.

Select the option for the data structure that applies to your data. For multidimensional data, click **Select Dimensions** to specify the dimension fields. The specified order of the dimension fields defines the order in which they appear in all subsequent dialogs and output. Use the up and down arrow buttons on the Select Dimensions subdialog to reorder the dimension fields.

For column-based data, the term *series* has the same meaning as the term *field*. For multidimensional data, fields that contain time series are referred to as *metric* fields. A time series, for multidimensional data, is defined by a metric field and a value for each of the dimension fields. The following considerations apply to both column-based and multidimensional data.

- Series that are specified as candidate inputs or as both target and input are considered for inclusion in the model for each target. The model for each target always includes lagged values of the target itself.
- Series that are specified as forced inputs are always included in the model for each target.
- At least one series must be specified as either a target or as both target and input.
- When **Use predefined roles** is selected, fields that have a role of Input are set as candidate inputs. No predefined role maps to a forced input.

Multidimensional data

For multidimensional data, you specify metric fields and associated roles in a grid, where each row in the grid specifies a single metric and role. By default, the model system includes series for all combinations of the dimension fields for each row in the grid. For example, if there are dimensions for *region* and *brand* then, by default, specifying the metric *sales* as a target means that there is a separate sales target series for each combination of *region* and *brand*.

For each row in the grid, you can customize the set of values for any of the dimension fields by clicking the ellipsis button for a dimension. This action opens the Select Dimension Values subdialog. You can also add, delete, or copy grid rows.

The **Series Count** column displays the number of sets of dimension values that are currently specified for the associated metric. The displayed value can be larger than the actual number of series (one series per set). This condition occurs when some of the specified combinations of dimension values do not correspond to series contained by the associated metric.

Streaming TCM Node - Select Dimension Values

For multidimensional data, you can customize the analysis by specifying which dimension values apply to a particular metric field with a particular role. For example, if *sales* is a metric field and *channel* is a dimension with values 'retail' and 'web,' you can specify that 'web' sales is an input and 'retail' sales is a target. You can also specify dimension subsets that apply to all metric fields used in the analysis. For example, if *region* is a dimension field that indicates geographical region, then you can limit the analysis to particular regions.

All values
Specifies that all values of the current dimension field are included. This option is the default.

Select values to include or exclude
Use this option to specify the set of values for the current dimension field. When **Include** is selected for the **Mode**, only values that are specified in the **Selected values** list are included. When **Exclude** is selected for the **Mode**, all values other than the values that are specified in the **Selected values** list are included.

You can filter the set of values from which to choose. Values that meet the filter condition appear in the **Matched** tab and values that do not meet the filter condition appear in the **Unmatched** tab of the **Unselected values** list. The **All** tab lists all unselected values, regardless of any filter condition.

- You can use asterisks (*) to indicate wildcard characters when you specify a filter.
- To clear the current filter, specify an empty value for the search term on the Filter Displayed Values dialog.

Streaming TCM Node - Observations options

On the Fields tab, use the **Observations** settings to specify the fields that define the observations.

Observations that are defined by date/times

You can specify that the observations are defined by a date, time, or timestamp field. In addition to the field that defines the observations, select the appropriate time interval that describes the observations. Depending on the specified time interval, you can also specify other settings, such as the interval between observations (increment) or the number of days per week. The following considerations apply to the time interval:

- Use the value **Irregular** when the observations are irregularly spaced in time, such as the time at which a sales order is processed. When **Irregular** is selected, you must specify the time interval that is used for the analysis, from the **Time Interval** settings on the Data Specifications tab.
- When the observations represent a date and time and the time interval is hours, minutes, or seconds, then use **Hours per day**, **Minutes per day**, or **Seconds per day**. When the observations represent a time (duration) without reference to a date and the time interval is hours, minutes, or seconds, then use **Hours (non-periodic)**, **Minutes (non-periodic)**, or **Seconds (non-periodic)**.

- Based on the selected time interval, the procedure can detect missing observations. Detecting missing observations is necessary since the procedure assumes that all observations are equally spaced in time and that there are no missing observations. For example, if the time interval is Days and the date 2014-10-27 is followed by 2014-10-29, then there is a missing observation for 2014-10-28. Values are imputed for any missing observations. Settings for handling missing values can be specified from the Data Specifications tab.
- The specified time interval allows the procedure to detect multiple observations in the same time interval that need to be aggregated together and to align observations on an interval boundary, such as the first of the month, to ensure that the observations are equally spaced. For example, if the time interval is Months, then multiple dates in the same month are aggregated together. This type of aggregation is referred to as *grouping*. By default, observations are summed when grouped. You can specify a different method for grouping, such as the mean of the observations, from the **Aggregation and Distribution** settings on the Data Specifications tab.
- For some time intervals, the additional settings can define breaks in the normal equally spaced intervals. For example, if the time interval is Days, but only weekdays are valid, you can specify that there are five days in a week, and the week begins on Monday.

Observations that are defined by periods or cyclic periods

Observations can be defined by one or more integer fields that represent periods or repeating cycles of periods, up to an arbitrary number of cycle levels. With this structure, you can describe series of observations that don't fit one of the standard time intervals. For example, a fiscal year with only 10 months can be described with a cycle field that represents years and a period field that represents months, where the length of one cycle is 10.

Fields that specify cyclic periods define a hierarchy of periodic levels, where the lowest level is defined by the **Period** field. The next highest level is specified by a cycle field whose level is 1, followed by a cycle field whose level is 2, and so on. Field values for each level, except the highest, must be periodic with respect to the next highest level. Values for the highest level cannot be periodic. For example, in the case of the 10-month fiscal year, months are periodic within years and years are not periodic.

- The length of a cycle at a particular level is the periodicity of the next lowest level. For the fiscal year example, there is only one cycle level and the cycle length is 10 since the next lowest level represents months and there are 10 months in the specified fiscal year.
- Specify the starting value for any periodic field that does not start from 1. This setting is necessary for detecting missing values. For example, if a periodic field starts from 2 but the starting value is specified as 1, then the procedure assumes that there is a missing value for the first period in each cycle of that field.

Streaming TCM Node - Time Interval options

The time interval that is used for the analysis can differ from the time interval of the observations. For example, if the time interval of the observations is Days, you might choose Months for the time interval for analysis. The data are then aggregated from daily to monthly data before the model is built. You can also choose to distribute the data from a longer to a shorter time interval. For example, if the observations are quarterly then you can distribute the data from quarterly to monthly data.

The available choices for the time interval at which the analysis is done depend on how the observations are defined and the time interval of those observations. In particular, when the observations are defined by cyclic periods, then only aggregation is supported. In that case, the time interval of the analysis must be greater than or equal to the time interval of the observations.

The time interval for the analysis is specified from the **Time Interval** settings on the Data Specifications tab. The method by which the data are aggregated or distributed is specified from the **Aggregation and Distribution** settings on the Data Specifications tab.

Streaming TCM Node - Aggregation and Distribution options

Aggregation functions
When the time interval that is used for the analysis is longer than the time interval of the observations, the input data are aggregated. For example, aggregation is done when the time interval of the observations is Days and the time interval for analysis is Months. The following aggregation functions are available: mean, sum, mode, min, or max.

Distribution functions
When the time interval that is used for the analysis is shorter than the time interval of the observations, the input data are distributed. For example, distribution is done when the time interval of the observations is Quarters and the time interval for analysis is Months. The following distribution functions are available: mean or sum.

Grouping functions
Grouping is applied when observations are defined by date/times and multiple observations occur in the same time interval. For example, if the time interval of the observations is Months, then multiple dates in the same month are grouped and associated with the month in which they occur. The following grouping functions are available: mean, sum, mode, min, or max. Grouping is always done when the observations are defined by date/times and the time interval of the observations is specified as Irregular.

Note: Although grouping is a form of aggregation, it is done before any handling of missing values whereas formal aggregation is done after any missing values are handled. When the time interval of the observations is specified as Irregular, aggregation is done only with the grouping function.

Aggregate cross-day observations to previous day
Specifies whether observations with times that cross a day boundary are aggregated to the values for the previous day. For example, for hourly observations with an eight-hour day that starts at 20:00, this setting specifies whether observations between 00:00 and 04:00 are included in the aggregated results for the previous day. This setting applies only if the time interval of the observations is Hours per day, Minutes per day or Seconds per day and the time interval for analysis is Days.

Custom settings for specified fields
You can specify aggregation, distribution, and grouping functions on a field by field basis. These settings override the default settings for the aggregation, distribution, and grouping functions.

Streaming TCM Node - Missing Value options

Missing values in the input data are replaced with an imputed value. The following replacement methods are available:

Linear interpolation
Replaces missing values by using a linear interpolation. The last valid value before the missing value and the first valid value after the missing value are used for the interpolation. If the first or last observation in the series has a missing value, then the two nearest non-missing values at the beginning or end of the series are used.

Series mean
Replaces missing values with the mean for the entire series.

Mean of nearby points
Replaces missing values with the mean of valid surrounding values. The span of nearby points is the number of valid values before and after the missing value that are used to compute the mean.

Median of nearby points
Replaces missing values with the median of valid surrounding values. The span of nearby points is the number of valid values before and after the missing value that are used to compute the median.

Linear trend
This option uses all non-missing observations in the series to fit a simple linear regression model, which is then used to impute the missing values.

Other settings:

Maximum percentage of missing values (%)
Specifies the maximum percentage of missing values that are allowed for any series. Series with more missing values than the specified maximum are excluded from the analysis.

Streaming TCM Node - General Data options

Maximum number of distinct values per dimension field
This setting applies to multidimensional data and specifies the maximum number of distinct values that are allowed for any one dimension field. By default, this limit is set to 10000 but it can be increased to an arbitrarily large number.

Streaming TCM Node - General Build options

Confidence interval width (%)
This setting controls the confidence intervals for both forecasts and model parameters. You can specify any positive value less than 100. By default, a 95% confidence interval is used.

Maximum number of inputs for each target
This setting specifies the maximum number of inputs that are allowed in the model for each target. You can specify an integer in the range 1 - 20. The model for each target always includes lagged values of itself, so setting this value to 1 specifies that the only input is the target itself.

Model tolerance
This setting controls the iterative process that is used for determining the best set of inputs for each target. You can specify any value that is greater than zero. The default is 0.001. Model tolerance is a stop criterion for predictor selection. It can affect the number of predictors that are included in the final model. But if a target can predict itself very well, other predictors may not be included in the final model. Some trial and error may be required (for example, if you have this value set to high, you can try setting it to a smaller value to see if other predictors can be included or not).

Outlier threshold (%)
An observation is flagged as an outlier if the probability, as calculated from the model, that it is an outlier exceeds this threshold. You can specify a value in the range 50 - 100.

Number of Lags for Each Input
This setting specifies the number of lag terms for each input in the model for each target. By default, the number of lag terms is automatically determined from the time interval that is used for the analysis. For example, if the time interval is months (with an increment of one month) then the number of lags is 12. Optionally, you can explicitly specify the number of lags. The specified value must be an integer in the range 1 - 20.

Continue estimation using existing models
If you already generated a temporal causal model, select this option to reuse the criteria settings that are specified for that model, rather than building a new model. In this way, you can save time by reestimating and producing a new forecast that is based on the same model settings as before but using more recent data.

Streaming TCM Node - Estimation Period options

By default, the estimation period starts at the time of the earliest observation and ends at the time of the latest observation across all series.

By start and end times
You can specify both the start and end of the estimation period or you can specify just the start or just the end. If you omit the start or the end of the estimation period, the default value is used.

- If the observations are defined by a date/time field, then enter values for start and end in the same format that is used for the date/time field.
- For observations that are defined by cyclic periods, specify a value for each of the cyclic periods fields. Each field is displayed in a separate column.

By latest or earliest time intervals

Defines the estimation period as a specified number of time intervals that start at the earliest time interval or end at the latest time interval in the data, with an optional offset. In this context, the time interval refers to the time interval of the analysis. For example, assume that the observations are monthly but the time interval of the analysis is quarters. Specifying **Latest** and a value of 24 for the **Number of time intervals** means the latest 24 quarters.

Optionally, you can exclude a specified number of time intervals. For example, specifying the latest 24 time intervals and 1 for the number to exclude means that the estimation period consists of the 24 intervals that precede the last one.

Streaming TCM Node - Model options

Model Name

You can specify a custom name for the model or accept the automatically generated name, which is *TCM*.

Forecast

The option to **Extend records into the future** sets the number of time intervals to forecast beyond the end of the estimation period. The time interval in this case is the time interval of the analysis, which is specified on the Data Specifications tab. When forecasts are requested, autoregressive models are automatically built for any input series that are not also targets. These models are then used to generate values for those input series in the forecast period. There is no maximum limit for this setting.

CPLEX Optimization node

The CPLEX Optimization node provides the ability to use complex mathematical (CPLEX) based optimization via an Optimization Programming Language (OPL) model file. This functionality was available in the IBM Analytical Decision Management product that's no longer supported, but you can also use the CPLEX node in SPSS Modeler without requiring IBM Analytical Decision Management.

For more information about CPLEX optimization and OPL, see the IBM ILOG CPLEX Optimization Studio documentation.

The CPLEX Optimization node supports multiple data sources, or multiple dimensional incoming data. You can connect several nodes to the CPLEX Optimization node, and each prior node can be utilized to provide data to OPL model computing – set as individual tuple sets with individual field mappings.

When outputting the data generated by the CPLEX Optimization node, the original data from the data sources can be output together as single indexes, or as multiple dimensional indexes of the result.

Note: When running a stream containing a CPLEX Optimization node on **IBM SPSS Modeler Server**, by default the embedded Community edition CPLEX library is used. It has a limitation of 1000 variables and 1000 constraints. If you install the full edition of IBM ILOG CPLEX and want to use the full edition's CPLEX engine instead, which doesn't have such limitations, complete the following step for your platform.

- On Windows, edit `options.cfg` and add the OPL library path. For example:

```
cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
```

  Where `<CPLEX_path>` is the CPLEX installation directory such as `C:\Program Files\IBM\ILOG\CPLEX_Studio127`, and `<Platform_dir>` is the platform-specific directory such as `x64_win64`.

- On Linux, edit `modelersrv.sh` and add the OPL library path. For example:

```
CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
```

  Where `<CPLEX_path>` is the CPLEX installation directory such as `/root/Libs_127_FullEdition/Linux_x86_64`, and `<Platform_dir>` is the platform-specific directory such as `x86-64_linux`.

Note: When running a stream containing a CPLEX Optimization node in **SPSS Modeler Solution Publisher**, by default the embedded Community edition CPLEX library is used. It has a limitation of 1000

variables and 1000 constraints. If you install the full edition of IBM ILOG CPLEX and want to use the full edition's CPLEX engine instead, which doesn't have such limitations, complete the following step for your platform.

- On Windows, add the OPL library path as a command line argument for `modelerrun.exe`. For example:

```
-o cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
```

  Where <CPLEX_path> is the CPLEX installation directory such as `C:\Program Files\IBM\ILOG \CPLEX_Studio127`, and <Platform_dir> is the platform-specific directory such as `x64_win64`.
- On Linux, edit `modelerrun` and add the OPL library path. For example:

```
CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
```

  Where <CPLEX_path> is the CPLEX installation directory such as `/root/Libs_127_FullEdition/ Linux_x86_64`, and <Platform_dir> is the platform-specific directory such as `x86-64_linux`.

Setting options for the CPLEX Optimization node

The Options tab of the CPLEX Optimization node contains the following fields.

OPL Model File. Select an Optimization Programming Language (OPL) model file.

OPL Model. After selecting an OPL model, the contents are shown here.

Input Data

On the Input Data tab, the **Data Source** drop-down lists all data sources (prior nodes) connected to the current CPLEX Optimization node. Selecting a data source from the drop-down refreshes the **Input Mappings** section below. Click **Apply All Fields** to generate all field mappings for the selected data source automatically. The **Input Mappings** table will be populated automatically.

Enter the tuple set name in the OPL model that corresponds to the incoming data. Then, if needed, verify that all tuple fields are mapped to data input fields according to their order in the tuple definition.

After setting the input mapping for a data source, you can select a different data source from the drop-down and repeat the process. The previous data source mappings will be saved automatically. Click **Apply** or **OK** when finished.

Other Data

On the Other Data tab, use the **OPL Data** section if you need to specify any other data for the optimization.

Output

When the output is a decision variable, it must take the prior data sources (incoming data) as indexes, and the indexes must be predefined in the **Input Mappings** section on the Input Data tab. No other type of decision variables are currently supported. The decision variable can have a single index or multiple indexes. SPSS Modeler will output the CPLEX results with all or part of the original incoming data together, which is consistent with other SPSS Modeler nodes. The referred corresponding indexes must be specified in the **Output Tuple** field described below.

On the Output tab, choose the output mode (**Raw Output** or **Decision Variable**) and specify other options as appropriate. The Raw Output option will output the objective function value directly, regardless of name.

Objective Function Value Variable Name in OPL. This field is enabled if you selected the **Decision Variable** output mode. Enter the name of the objective function value variable from the OPL model.

Objective Function Value Field Name for Output. Enter the field name to use in the output. The default is `_OBJECTIVE`.

Output Tuple. Enter the name of the predefined tuple from the incoming data. This acts as the indexes of the decision variable and is expected to be output with the Variable Outputs. The Output Tuple should be consistent with the decision variable definition in the OPL. If there are multiple indexes, the tuple names must be joined by a comma (,).

Variable Outputs. Add one or more variables to include in the output.

Chapter 4. Field Operations Nodes

Field Operations Overview

After an initial data exploration, you will probably have to select, clean, or construct data in preparation for analysis. The Field Operations palette contains many nodes useful for this transformation and preparation.

For example, using a Derive node, you might create an attribute that is not currently represented in the data. Or you might use a Binning node to recode field values automatically for targeted analysis. You will probably find yourself using a Type node frequently—it allows you to assign a measurement level, values, and a modeling role for each field in the dataset. Its operations are useful for handling missing values and downstream modeling.

The Field Operations palette contains the following nodes:

The Automated Data Preparation (ADP) node can analyze your data and identify fixes, screen out fields that are problematic or not likely to be useful, derive new attributes when appropriate, and improve performance through intelligent screening and sampling techniques. You can use the node in fully automated fashion, allowing the node to choose and apply fixes, or you can preview the changes before they are made and accept, reject, or amend them as desired.

The Type node specifies field metadata and properties. For example, you can specify a measurement level (continuous, nominal, ordinal, or flag) for each field, set options for handling missing values and system nulls, set the role of a field for modeling purposes, specify field and value labels, and specify values for a field.

The Filter node filters (discards) fields, renames fields, and maps fields from one source node to another.

The Derive node modifies data values or creates new fields from one or more existing fields. It creates fields of type formula, flag, nominal, state, count, and conditional.

The Ensemble node combines two or more model nuggets to obtain more accurate predictions than can be gained from any one model.

The Filler node replaces field values and changes storage. You can choose to replace values based on a CLEM condition, such as `@BLANK(@FIELD)`. Alternatively, you can choose to replace all blanks or null values with a specific value. A Filler node is often used together with a Type node to replace missing values.

The Anonymize node transforms the way field names and values are represented downstream, thus disguising the original data. This can be useful if you want to allow other users to build models using sensitive data, such as customer names or other details.

The Reclassify node transforms one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis.

The Binning node automatically creates new nominal (set) fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing groups of income as deviations from the mean. Once you have created bins for the new field, you can generate a Derive node based on the cut points.

The Recency, Frequency, Monetary (RFM) Analysis node enables you to determine quantitatively which customers are likely to be the best ones by examining how recently they last purchased from you (recency), how often they purchased (frequency), and how much they spent over all transactions (monetary).

The Partition node generates a partition field, which splits the data into separate subsets for the training, testing, and validation stages of model building.

The Set to Flag node derives multiple flag fields based on the categorical values defined for one or more nominal fields.

The Restructure node converts a nominal or flag field into a group of fields that can be populated with the values of yet another field. For example, given a field named *payment type*, with values of *credit*, *cash*, and *debit*, three new fields would be created (*credit*, *cash*, *debit*), each of which might contain the value of the actual payment made.

The Transpose node swaps the data in rows and columns so that records become fields and fields become records.

Use the Time Intervals node to specify intervals and derive a new time field for estimating or forecasting. A full range of time intervals is supported, from seconds to years.

The History node creates new fields containing data from fields in previous records. History nodes are most often used for sequential data, such as time series data. Before using a History node, you may want to sort the data using a Sort node.

The Field Reorder node defines the natural order used to display fields downstream. This order affects the display of fields in a variety of places, such as tables, lists, and the Field Chooser. This operation is useful when working with wide datasets to make fields of interest more visible.

Within SPSS Modeler, items such as the Expression Builder spatial functions, the Spatio-Temporal Prediction (STP) Node, and the Map Visualization Node use the projected coordinate system. Use the Reproject node to change the coordinate system of any data that you import that uses a geographic coordinate system.

Several of these nodes can be generated directly from the audit report created by a Data Audit node. See the topic "Generating Other Nodes for Data Preparation" on page 288 for more information.

Automated Data Preparation

Preparing data for analysis is one of the most important steps in any project—and traditionally, one of the most time consuming. Automated Data Preparation (ADP) handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques. You can use the algorithm in fully **automatic** fashion, allowing it to choose and apply fixes, or you can use it in **interactive** fashion, previewing the changes before they are made and accept or reject them as you want.

Using ADP enables you to make your data ready for model building quickly and easily, without needing prior knowledge of the statistical concepts involved. Models will tend to build and score more quickly; in addition, using ADP improves the robustness of automated modeling processes, such as model refresh and champion / challenger.

Note: When ADP prepares a field for analysis, it creates a new field containing the adjustments or transformations, rather than replacing the existing values and properties of the old field. The old field is not used in further analysis; its role is set to None.

Example. An insurance company with limited resources to investigate homeowner's insurance claims wants to build a model for flagging suspicious, potentially fraudulent claims. Before building the model, they will ready the data for modeling using automated data preparation. Since they want to be able to review the proposed transformations before the transformations are applied, they will use automated data preparation in interactive mode.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, they want to establish a relationship between vehicle sales and vehicle characteristics. They will use automated data preparation to prepare the data for analysis, and build models using the data "before" and "after" preparation to see how the results differ.

What is your objective? Automated data preparation recommends data preparation steps that will affect the speed with which other algorithms can build models and improve the predictive power of those models. This can include transforming, constructing and selecting features. The target can also be transformed. You can specify the model-building priorities that the data preparation process should concentrate on.

- **Balance speed and accuracy.** This option prepares the data to give equal priority to both the speed with which data are processed by model-building algorithms and the accuracy of the predictions.
- **Optimize for speed.** This option prepares the data to give priority to the speed with which data are processed by model-building algorithms. When you are working with very large datasets, or are looking for a quick answer, select this option.
- **Optimize for accuracy.** This option prepares the data to give priority to the accuracy of predictions produced by model-building algorithms.
- **Custom analysis.** When you want to manually change the algorithm on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with one of the other objectives.

Training the node

The ADP node is implemented as a process node and works in a similar way to the Type node; **training** the ADP node corresponds to instantiating the Type node. Once analysis has been performed, the specified transformations are applied to the data without further analysis as long as the upstream data

model does not change. Like the Type and Filter nodes, if the ADP node is disconnected it remembers the data model and transformations so that if it is reconnected it does not need to be retrained; this enables you to train it on a subset of typical data and then copy or deploy it for use it on live data as often as required.

Using the toolbar

The toolbar enables you to run and update the display of the data analysis, and generate nodes that you can use in conjunction with the original data.

- **Generate** From this menu you can generate either a Filter or Derive node. Note that this menu is only available when there is an analysis shown on the Analysis tab.

  The Filter node removes transformed input fields. If you configure the ADP node to leave the original input fields in the dataset, this will restore the original set of inputs allowing you to interpret the score field in terms of the inputs. For example, this may be useful if you want to produce a graph of the score field against various inputs.

  The Derive node can restore the original dataset and target units. You can only generate a Derive node when the ADP node contains an analysis which rescales a range target (that is, Box-Cox rescaling is selected on the Prepare Inputs & Target panel). You cannot generate a Derive node if the target is not a range, or if the Box-Cox rescaling is not selected. See the topic "Generating a Derive node" on page 121 for more information.
- **View** Contains options that control what is shown on the Analysis tab. This includes the graph editing controls and the display selections for both the main panel and linked views.
- **Preview** Displays a sample of the transformations that will be applied to the input data.
- **Analyze Data** Initiates an analysis using the current settings and displays the results on the Analysis tab.
- **Clear Analysis** Deletes the existing analysis (only available when a current analysis exists).

Node status

The status of the ADP node on the IBM SPSS Modeler canvas is indicated by either an arrow or tick on the icon that shows whether or not analysis has taken place.

For more information about the calculations performed with the Automated Data Preparation node, see the *Automated Data Preparation algorithms* section of the *IBM SPSS Modeler Algorithms Guide*. The guide is available in PDF format from the `\Documentation` directory of the installation disk, as part of your product download, or on the Web.

Fields tab

Before you can build a model, you need to specify which fields you want to use as targets and as inputs. With a few exceptions, all modeling nodes will use field information from an upstream Type node. If you are using a Type node to select input and target fields, you don't need to change anything on this tab.

Use type node settings. This option tells the node to use field information from an upstream Type node. This is the default.

Use custom settings. This option tells the node to use field information specified here instead of that given in any upstream Type node(s). After selecting this option, specify the fields below as required.

Target. For models that require one or more target fields, select the target field or fields. This is similar to setting the field role to *Target* in a Type node.

Inputs. Select the input field or fields. This is similar to setting the field role to *Input* in a Type node.

Settings tab

The Settings tab comprises several different groups of settings that you can modify to fine-tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the other objectives, the Objective tab is automatically updated to select the **Customize analysis** option.

Field settings

Use frequency field. This option enables you to select a field as a frequency weight. Use this if the records in your training data represent more than one unit each; for example, if you are using aggregated data. The field values should be the number of units represented by each record.

Use weight field. This option enables you to select a field as a case weight. Case weights are used to account for differences in variance across levels of the output field.

How to handle fields that are excluded from modeling. Specify what happens to excluded fields; you can choose either to filter them out of the data or simply set their *Role* to **None**.

Note: This action will also be applied to the target if it is transformed. For example, if the new derived version of the target is used as the **Target** field, the original target is either filtered or set to **None**.

If the incoming fields do not match the existing analysis. Specify what happens if one or more required input fields are missing from the incoming dataset when you execute a trained ADP node.

- **Stop execution and keep the existing analysis.** This stops the execution process, preserves the current analysis information, and displays an error.
- **Clear the existing analysis and analyze the new data.** This clears the existing analysis, analyzes the incoming data, and applies the recommended transformations to that data.

Prepare dates & times

Many modeling algorithms are unable to directly handle date and time details; these settings enable you to derive new duration data that can be used as model inputs from dates and times in your existing data. The fields containing dates and times must be predefined with date or time storage types. The original date and time fields will not be recommended as model inputs following automated data preparation.

Prepare dates and times for modeling. Deselecting this option disables all other Prepare Dates & Times controls while maintaining the selections.

Compute elapsed time until reference date. This produces the number of years/months/days since a reference date for each variable containing dates.

- **Reference Date.** Specify the date from which the duration will be calculated with regard to the date information in the input data. Selecting **Today's date** means that the current system date is always used when ADP is executed. To use a specific date, select **Fixed date** and enter the required date. The current date is automatically entered in the **Fixed date** field when the node is first created.
- **Units for Date Duration.** Specify whether ADP should automatically decide on the date duration unit, or select from **Fixed units** of Years, Months, or Days.

Compute elapsed time until reference time. This produces the number of hours/minutes/seconds since a reference time for each variable containing times.

- **Reference Time.** Specify the time from which the duration will be calculated with regard to the time information in the input data. Selecting **Current time** means that the current system time is always used when ADP is executed. To use a specific time, select **Fixed time** and enter the required details. The current time is automatically entered in the **Fixed time** field when the node is first created.
- **Units for Time Duration.** Specify whether ADP should automatically decide on the time duration unit, or select from **Fixed units** of Hours, Minutes, or Seconds.

Extract Cyclical Time Elements. Use these settings to split a single date or time field into one or more fields. For example if you select all three date checkboxes, the input date field "1954-05-23" is split into three fields: 1954, 5, and 23, each using the suffix defined on the **Field Names** panel, and the original date field is ignored.

- **Extract from dates.** For any date inputs, specify if you want to extract years, months, days, or any combination.
- **Extract from times.** For any time inputs, specify if you want to extract hours, minutes, seconds, or any combination.

Excluding fields

Poor quality data can affect the accuracy of your predictions; therefore, you can specify the acceptable quality level for input features. All fields that are constant or have 100% missing values are automatically excluded.

Exclude low quality input fields. Deselecting this option disables all other Exclude Fields controls while maintaining the selections.

Exclude fields with too many missing values. Fields with more than the specified percentage of missing values are removed from further analysis. Specify a value greater than or equal to 0, which is equivalent to deselecting this option, and less than or equal to 100, though fields with all missing values are automatically excluded. The default is 50.

Exclude nominal fields with too many unique categories. Nominal fields with more than the specified number of categories are removed from further analysis. Specify a positive integer. The default is 100. This is useful for automatically removing fields containing record-unique information from modeling, like ID, address, or name.

Exclude categorical fields with too many values in a single category. Ordinal and nominal fields with a category that contains more than the specified percentage of the records are removed from further analysis. Specify a value greater than or equal to 0, equivalent to deselecting this option, and less than or equal to 100, though constant fields are automatically excluded. The default is 95.

Preparing inputs and targets

Because no data is ever in a perfect state for processing, you may want to adjust some of the settings before running an analysis. For example, this might include the removal of outliers, specifying how to handle missing values, or adjusting the type.

Note: If you change the values on this panel, the **Objectives** tab is automatically updated to select the **Custom analysis** option.

Prepare the input and target fields for modeling. Toggles all fields on the panel either on or off.

Adjust Type and Improve Data Quality. For inputs and the target you can specify several data transformations separately; this is because you may not want to change the values of the target. For example, a prediction of income in dollars is more meaningful than a prediction measured in log(dollars). In addition, if the target has missing values there is no predictive gain to filling missing values, whereas filling missing values in inputs may enable some algorithms to process information that would otherwise be lost.

Additional settings for these transformations, such as the outlier cutoff value, are common to both the target and inputs.

You can select the following settings for either, or both, inputs and target:

- **Adjust the type of numeric fields.** Select this to determine if numeric fields with a measurement level of *Ordinal* can be converted to *Continuous*, or vice versa. You can specify the minimum and maximum threshold values to control the conversion.
- **Reorder nominal fields.** Select this to sort nominal (set) fields into order, from smallest to largest category.
- **Replace outlier values in continuous fields.** Specify whether to replace outliers; use this in conjunction with the **Method for replacing outliers** options below.
- **Continuous fields: replace missing values with mean.** Select this to replace missing values of continuous (range) features.

- **Nominal fields: replace missing values with mode.** Select this to replace missing values of nominal (set) features.
- **Ordinal fields: replace missing values with median.** Select this to replace missing values of ordinal (ordered set) features.

Maximum number of values for ordinal fields. Specify the threshold for redefining ordinal (ordered set) fields as continuous (range). The default is 10; therefore, if an ordinal field has more than 10 categories it is redefined as continuous (range).

Minimum number of values for continuous fields. Specify the threshold for redefining scale or continuous (range) fields as ordinal (ordered set). The default is 5; therefore, if a continuous field has fewer than 5 values it is redefined as ordinal (ordered set).

Outlier cutoff value. Specify the outlier cutoff criterion, measured in standard deviations; the default is 3.

Method for replacing outliers. Select whether outliers are to be replaced by either trimming (coerce) with the cutoff value, or to delete them and set them as missing values. Any outliers set to missing values follow the missing value handling settings selected above.

Put all continuous input fields on a common scale. To normalize continuous input fields, select this check box and choose the normalization method. The default is **z-score transformation**, where you can specify the **Final mean**, which has a default of 0, and the **Final standard deviation**, which has a default of 1. Alternatively, you can choose to use **Min/max transformation** and specify the minimum and maximum values, which have default values of 0 and 100 respectively.

This field is especially useful when you select **Perform feature construction** on the Construct & Select Features panel.

Rescale a continuous target with a Box-Cox transformation. To normalize a continuous (scale or range) target field, select this check box. The Box-Cox transformation has default values of 0 for the **Final mean** and 1 for the **Final standard deviation**.

Note: If you choose to normalize the target, the dimension of the target will be transformed. In this case you may need to generate a Derive node to apply an inverse transformation in order to turn the transformed units back into a recognizable format for further processing. See the topic "Generating a Derive node" on page 121 for more information.

Construction and feature selection

To improve the predictive power of your data, you can transform the input fields, or construct new ones based on the existing fields.

Note: If you change the values on this panel, the **Objectives** tab is automatically updated to select the **Custom analysis** option.

Transform, construct and select input fields to improve predictive power. Toggles all fields on the panel either on or off.

Merge sparse categories to maximize association with target. Select this to make a more parsimonious model by reducing the number of variables to be processed in association with the target. If required, change the probability value from the default of 0.05.

Note that if all categories are merged into one, the original and derived versions of the field are excluded because they have no value as a predictor.

When there is no target, merge sparse categories based on counts. If you are dealing with data that has no target, you can choose to merge sparse categories of either, or both, ordinal (ordered set) and nominal (set) features. Specify the minimum percentage of cases, or records, in the data that identifies the categories to be merged; the default is 10.

Categories are merged using the following rules:

- Merging is not performed on binary fields.
- If there are only two categories during merging, merging stops.

- If there is no original category, nor any category created during merging, with fewer than the specified minimum percent of cases, merging stops.

Bin continuous fields while preserving predictive power. Where you have data that includes a categorical target, you can bin continuous inputs with strong associations to improve processing performance. If required, change the probability value for the homogenous subsets from the default of 0.05.

If the binning operation results in a single bin for a particular field, the original and binned versions of the field are excluded because they have no value as a predictor.

Note: Binning in ADP differs from optimal binning used in other parts of IBM SPSS Modeler. Optimal binning uses entropy information to convert a continuous variable to a categorical variable; this needs to sort data and store it all in memory. ADP uses homogenous subsets to bin a continuous variable, this means that ADP binning does not need to sort data and does not store all data in memory. The use of the homogenous subset method to bin a continuous variable means that the number of categories after binning is always less than or equal to the number of categories of target.

Perform feature selection. Select this option to remove features with a low correlation coefficient. If required, change the probability value from the default of 0.05.

This option only applies to continuous input features where the target is continuous, and to categorical input features.

Perform feature construction. Select this option to derive new features from a combination of several existing features (which are then discarded from modeling).

This option only applies to continuous input features where the target is continuous, or where there is no target.

Field names

To easily identify new and transformed features, ADP creates and applies basic new names, prefixes, or suffixes. You can amend these names to be more relevant to your own needs and data. If you want to specify other labels, you will need to do this in a downstream Type node.

Transformed and Constructed Fields. Specify the name extensions to be applied to transformed target and input fields.

Note that in the ADP node, setting string fields to contain nothing may cause an error depending on how you chose to handle unused fields. If **How to handle fields that are excluded from modeling** is set to **Filter out unused fields** on the Field Settings panel of the Settings tab, the name extensions for inputs and the target can be set to nothing. The original fields are filtered out and the transformed fields saved over them; in this case the new transformed fields will have the same name as your original.

However if you chose **Set the direction of unused fields to 'None'**, then empty, or null, name extensions for the target and inputs will cause an error because you will be trying to create duplicate field names.

In addition, specify the prefix name to be applied to any features that are constructed via the Select and Construct settings. The new name is created by attaching a numeric suffix to this prefix root name. The format of the number depends on how many new features are derived, for example:

- 1-9 constructed features will be named: feature1 to feature9.
- 10-99 constructed features will be named: feature01 to feature99.
- 100-999 constructed features will be named: feature001 to feature999, and so on.

This ensures that the constructed features will sort in a sensible order no matter how many there are.

Durations Computed from Dates and Times. Specify the name extensions to be applied to durations computed from both dates and times.

Cyclical Elements Extracted from Dates and Times. Specify the name extensions to be applied to cyclical elements extracted from both dates and times.

Analysis tab

1. When you are satisfied with the ADP settings, including any changes made on the Objective, Fields, and Settings tabs, click **Analyze Data**; the algorithm applies the settings to the data inputs and displays the results on the Analysis tab.

The Analysis tab contains both tabular and graphical output that summarizes the processing of your data and displays recommendations as to how the data may be modified or improved for scoring. You can then review and either accept or reject those recommendations.

The Analysis tab is made up of two panels, the main view on the left and the linked, or auxiliary, view on the right. There are three main views:

- Field Processing Summary (the default). See the topic "Field Processing Summary " on page 115 for more information.
- Fields. See the topic "Fields " on page 116 for more information.
- Action Summary. See the topic "Action Summary " on page 117 for more information.

There are four linked/auxiliary views:

- Predictive Power (the default). See the topic "Predictive Power " on page 117 for more information.
- Fields Table. See the topic "Fields Table " on page 117 for more information.
- Field Details. See the topic "Field Details " on page 118 for more information.
- Action Details. See the topic "Action Details " on page 119 for more information.

Links between views

Within the main view, underlined text in the tables controls the display in the linked view. Clicking on the text allows you to get details on a particular field, set of fields, or processing step. The link that you last selected is shown in a darker color; this helps you identify the connection between the contents of the two view panels.

Resetting the views

To redisplay the original Analysis recommendations and abandon any changes you have made to the Analysis views, click **Reset** at the bottom of the main view panel.

Field Processing Summary

The Field Processing Summary table gives a snapshot of the projected overall impact of processing, including changes to the state of the features and the number of features constructed.

Note that no model is actually built, so there isn't a measure or graph of the change in overall predictive power before and after data preparation; instead, you can display graphs of the predictive power of individual recommended predictors.

The table displays the following information:

- The number of target fields.
- The number of original (input) predictors.
- The predictors recommended for use in analysis and modeling. This includes the total number of fields recommended; the number of original, untransformed, fields recommended; the number of transformed fields recommended (excluding intermediate versions of any field, fields derived from date/time predictors, and constructed predictors); the number of fields recommended that are derived from date/time fields; and the number of constructed predictors recommended.
- The number of input predictors not recommended for use in any form, whether in their original form, as a derived field, or as input to a constructed predictor.

Where any of the **Fields** information is underlined, click to display more details in a linked view. Details of the **Target**, **Input features**, and **Input features not used** are shown in the Fields Table linked view. See

the topic “Fields Table ” on page 117 for more information. **Features recommended for use in analysis** are displayed in the Predictive Power linked view. See the topic “Predictive Power ” on page 117 for more information.

Fields

The Fields main view displays the processed fields and whether ADP recommends using them in downstream models. You can override the recommendation for any field; for example, to exclude constructed features or include features that ADP recommends excluding. If a field has been transformed, you can decide whether to accept the suggested transformation or use the original version.

The Fields view consists of two tables, one for the target and one for predictors that were either processed or created.

Target table

The **Target** table is only shown if a target is defined in the data.

The table contains two columns:

- **Name.** This is the name or label of the target field; the original name is always used, even if the field has been transformed.
- **Measurement Level.** This displays the icon representing the measurement level; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.

  If the target has been transformed the **Measurement Level** column reflects the final transformed version. *Note*: you cannot turn off transformations for the target.

Predictors table

The **Predictors** table is always shown. Each row of the table represents a field. By default the rows are sorted in descending order of predictive power.

For ordinary features, the original name is always used as the row name. Both original and derived versions of date/time fields appear in the table (in separate rows); the table also includes constructed predictors.

Note that transformed versions of fields shown in the table always represent the final versions.

By default only recommended fields are shown in the Predictors table. To display the remaining fields, select the **Include nonrecommended fields in table** box above the table; these fields are then displayed at the bottom of the table.

The table contains the following columns:

- **Version to Use.** This displays a drop-down list that controls whether a field will be used downstream and whether to use the suggested transformations. By default, the drop-down list reflects the recommendations.

  For ordinary predictors that have been transformed the drop-down list has three choices: **Transformed**, **Original**, and **Do not use**.

  For untransformed ordinary predictors the choices are: **Original** and **Do not use**.

  For derived date/time fields and constructed predictors the choices are: **Transformed** and **Do not use**.

  For original date fields the drop-down list is disabled and set to **Do not use**.

  Note: For predictors with both original and transformed versions, changing between **Original** and **Transformed** versions automatically updates the **Measurement Level** and **Predictive Power** settings for those features.
- **Name.** Each field's name is a link. Click a name to display more information about the field in the linked view. See the topic “Field Details ” on page 118 for more information.

- **Measurement Level.** This displays the icon representing the data type; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.
- **Predictive Power.** Predictive power is displayed only for fields that ADP recommends. This column is not displayed if there is no target defined. Predictive power ranges from 0 to 1, with larger values indicating "better" predictors. In general, predictive power is useful for comparing predictors within an ADP analysis, but predictive power values should not be compared across analyses.

Action Summary

For each action taken by automated data preparation, input predictors are transformed and/or filtered out; fields that survive one action are used in the next. The fields that survive through to the last step are then recommended for use in modeling, whilst inputs to transformed and constructed predictors are filtered out.

The Action Summary is a simple table that lists the processing actions taken by ADP. Where any **Action** is underlined, click to display more details in a linked view about the actions taken. See the topic "Action Details " on page 119 for more information.

Note: Only the original and final transformed versions of each field are shown, not any intermediate versions that were used during analysis.

Predictive Power

Displayed by default when the analysis is first run, or when you select **Predictors recommended for use in analysis** in the Field Processing Summary main view, the chart displays the predictive power of recommended predictors. Fields are sorted by predictive power, with the field with the highest value appearing at the top.

For transformed versions of ordinary predictors, the field name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed*.

Measurement level icons are displayed after the individual field names.

The predictive power of each recommended predictor is computed from either a linear regression or naïve Bayes model, depending upon whether the target is continuous or categorical.

Fields Table

Displayed when you click **Target**, **Predictors**, or **Predictors not used** in the Field Processing Summary main view, the Fields Table view displays a simple table listing the relevant features.

The table contains two columns:

- **Name.** The predictor name.

  For targets, the original name or label of the field is used, even if the target has been transformed.

  For transformed versions of ordinary predictors, the name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed*.

  For fields derived from dates and times, the name of the final transformed version is used; for example: *bdate_years*.

  For constructed predictors, the name of the constructed predictor is used; for example: *Predictor1*.
- **Measurement Level.** This displays the icon representing the data type.

  For the Target, the **Measurement Level** always reflects the transformed version (if the target has been transformed); for example, changed from ordinal (ordered set) to continuous (range, scale), or vice versa.

Field Details

Displayed when you click any **Name** in the Fields main view, the Field Details view contains distribution, missing values, and predictive power charts (if applicable) for the selected field. In addition, the processing history for the field and the name of the transformed field are also shown (if applicable).

For each chart set, two versions are shown side by side to compare the field with and without transformations applied; if a transformed version of the field does not exist, a chart is shown for the original version only. For derived date or time fields and constructed predictors, the charts are only shown for the new predictor.

Note: If a field is excluded due to having too many categories only the processing history is shown.

Distribution Chart

Continuous field distribution is shown as a histogram, with a normal curve overlaid, and a vertical reference line for the mean value; categorical fields are displayed as a bar chart.

Histograms are labeled to show standard deviation and skewness; however, skewness is not displayed if the number of values is 2 or fewer or the variance of the original field is less than 10-20.

Hover the mouse over the chart to display either the mean for histograms, or the count and percentage of the total number of records for categories in bar charts.

Missing Value Chart

Pie charts compare the percentage of missing values with and without transformations applied; the chart labels show the percentage.

If ADP carried out missing value handling, the post-transformation pie chart also includes the replacement value as a label -- that is, the value used in place of missing values.

Hover the mouse over the chart to display the missing value count and percentage of the total number of records.

Predictive Power Chart

For recommended fields, bar charts display the predictive power before and after transformation. If the target has been transformed, the calculated predictive power is in respect to the transformed target.

Note: Predictive power charts are not shown if no target is defined, or if the target is clicked in the main view panel.

Hover the mouse over the chart to display the predictive power value.

Processing History Table

The table shows how the transformed version of a field was derived. Actions taken by ADP are listed in the order in which they were carried out; however, for certain steps multiple actions may have been carried out for a particular field.

Note: This table is not shown for fields that have not been transformed.

The information in the table is broken down into two or three columns:

- **Action.** The name of the action. For example, Continuous Predictors. See the topic “Action Details ” on page 119 for more information.
- **Details.** The list of processing carried out. For example, Transform to standard units.
- **Function.** Only shown for constructed predictors, this displays the linear combination of input fields, for example, .06*age + 1.21*height.

Action Details

Displayed when you select any underlined **Action** in the Action Summary main view, the Action Details linked view displays both action-specific and common information for each processing step that was carried out; the action-specific details are displayed first.

For each action, the description is used as the title at the top of the linked view. The action-specific details are displayed below the title, and may include details of the number of derived predictors, fields recast, target transformations, categories merged or reordered, and predictors constructed or excluded.

As each action is processed, the number of predictors used in the processing may change, for example as predictors are excluded or merged.

Note: If an action was turned off, or no target was specified, an error message is displayed in place of the action details when the action is clicked in the Action Summary main view.

There are nine possible actions; however, not all are necessarily active for every analysis.

Text Fields table

The table displays the number of:

- Trailing blank values trimmed.
- Predictors excluded from analysis.

Date and Time Predictors table

The table displays the number of:

- Durations derived from date and time predictors.
- Date and time elements.
- Derived date and time predictors, in total.

The reference date or time is displayed as a footnote if any date durations were calculated.

Predictor Screening table

The table displays the number of the following predictors excluded from processing:

- Constants.
- Predictors with too many missing values.
- Predictors with too many cases in a single category.
- Nominal fields (sets) with too many categories.
- Predictors screened out, in total.

Check Measurement Level table

The table displays the numbers of fields recast, broken down into the following:

- Ordinal fields (ordered sets) recast as continuous fields.
- Continuous fields recast as ordinal fields.
- Total number recast.

If no input fields (target or predictors) were continuous or ordinal, this is shown as a footnote.

Outliers table

The table displays counts of how any outliers were handled.

- Either the number of continuous fields for which outliers were found and trimmed, or the number of continuous fields for which outliers were found and set to missing, depending on your settings in the Prepare Inputs & Target panel on the Settings tab.
- The number of continuous fields excluded because they were constant, after outlier handling.

One footnote shows the outlier cutoff value; while another footnote is shown if no input fields (target or predictors) were continuous.

Missing Values table

The table displays the numbers of fields that had missing values replaced, broken down into:

- Target. This row is not shown if no target is specified.
- Predictors. This is further broken down into the number of nominal (set), ordinal (ordered set), and continuous.
- The total number of missing values replaced.

Target table

The table displays whether the target was transformed, shown as:

- Box-Cox transformation to normality. This is further broken down into columns that show the specified criteria (mean and standard deviation) and Lambda.
- Target categories reordered to improve stability.

Categorical Predictors table

The table displays the number of categorical predictors:

- Whose categories were reordered from lowest to highest to improve stability.
- Whose categories were merged to maximize association with the target.
- Whose categories were merged to handle sparse categories.
- Excluded due to low association with the target.
- Excluded because they were constant after merging.

A footnote is shown if there were no categorical predictors.

Continuous Predictors table

There are two tables. The first displays one of the following number of transformations:

- Predictor values transformed to standard units. In addition, this shows the number of predictors transformed, the specified mean, and the standard deviation.
- Predictor values mapped to a common range. In addition, this shows the number of predictors transformed using a min-max transformation, as well as the specified minimum and maximum values.
- Predictor values binned and the number of predictors binned.

The second table displays the predictor space construction details, shown as the number of predictors:

- Constructed.
- Excluded due to a low association with the target.
- Excluded because they were constant after binning.
- Excluded because they were constant after construction.

A footnote is shown if no continuous predictors were input.

Generating a Derive node

When you generate a Derive node, it applies the inverse target transformation to the score field. By default the node enters the name of the score field that would be produced by an automodeler (such as Auto Classifier or Auto Numeric) or the Ensemble node. If a scale (range) target has been transformed the score field is shown in transformed units; for example, log($) instead of $. In order to interpret and use the results, you must convert the predicted value back to the original scale.

Note: You can only generate a Derive node when the ADP node contains an analysis which rescales a range target (that is, Box-Cox rescaling is selected on the Prepare Inputs & Target panel). You cannot generate a Derive node if the target is not a range, or if the Box-Cox rescaling is not selected.

The Derive node is created in Multiple mode and uses @FIELD in the expression so you can add the transformed target if required. For example using the following details:

- Target field name: response
- Transformed target field name: response_transformed
- Score field name: $XR-response_transformed

The Derive node would create a new field: $XR-response_transformed_inverse.

Note: If you are not using an automodeler or Ensemble node, you will need to edit the Derive node to transform the correct score field for your model.

Normalized continuous targets

By default, if you select the **Rescale a continuous target with a Box-Cox transformation** check box on the Prepare Inputs & Target panel this transforms the target and you create a new field that will be the target for your model building. For example if your original target was *response*, the new target will be *response_transformed*; models downstream of the ADP node will pick this new target up automatically.

However, this may cause issues, depending on the original target. For example, if the target was *Age*, the values of the new target will not be *Years*, but a transformed version of *Years*. This means you cannot look at the scores and interpret them since they aren't in recognizable units. In this case you can apply an inverse transformation that will turn your transformed units back into whatever they were meant to be. To do this:

1. After clicking **Analyze Data** to run the ADP analysis, select *Derive Node* from the *Generate* menu.
2. Place the Derive node after your nugget on the model canvas.

The Derive node will restore the score field to the original dimensions so that the prediction will be in the original *Years* values.

By default the Derive node transforms the score field generated by an auto-modeler or an ensembled model. If you are building an individual model, you need to edit the Derive node to derive from your actual score field. If you want to evaluate your model, you should add the transformed target to the **Derive From** field in the Derive node. This applies the same inverse transformation to the target and any downstream Evaluation or Analysis node will use the transformed data correctly as long as you switch those nodes to use field names instead of metadata.

If you also want to restore the original name, you can use a Filter node to remove the original target field if it's still there, and rename the target and score fields.

Type Node

Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes. The following properties are available:

- **Field** Double-click any field name to specify value and field labels for data in IBM SPSS Modeler. For example, field metadata imported from IBM SPSS Statistics can be viewed or modified here. Similarly, you can create new labels for fields and their values. The labels that you specify here are displayed

throughout IBM SPSS Modeler depending on the selections you make in the stream properties dialog box.

- **Measurement** This is the measurement level, used to describe characteristics of the data in a given field. If all of the details of a field are known, it is called **fully instantiated**. For more information, see "Measurement levels" on page 122.

  Note: The measurement level of a field is different from its storage type, which indicates whether the data are stored as strings, integers, real numbers, dates, times, timestamps, or lists.

- **Values** This column enables you to specify options for reading data values from the data set, or use the **Specify** option to specify measurement levels and values in a separate dialog box. You can also choose to pass fields without reading their values. For more information, see "Data Values" on page 127.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.

- **Missing** Used to specify how missing values for the field will be handled. For more information, see "Defining Missing Values" on page 130.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.

- **Check** In this column, you can set options to ensure that field values conform to the specified values or ranges. For more information, see "Checking Type Values" on page 131.

  Note: You cannot amend the cell in this column if the corresponding **Field** entry contains a list.

- **Role** Used to tell modeling nodes whether fields will be **Input** (predictor fields) or **Target** (predicted fields) for a machine-learning process. **Both** and **None** are also available roles, along with **Partition**, which indicates a field used to partition records into separate samples for training, testing, and validation. The value **Split** specifies that separate models will be built for each possible value of the field. For more information, see "Setting the field role" on page 131.

Several other options can be specified using the Type node window:

- Using the tools menu button, you can choose to **Ignore Unique Fields** once a Type node has been instantiated (either through your specifications, reading values, or running the stream). Ignoring unique fields will automatically ignore fields with only one value.
- Using the tools menu button, you can choose to **Ignore Large Sets** once a Type node has been instantiated. Ignoring large sets will automatically ignore sets with a large number of members.
- Using the tools menu button, you can choose to **Convert Continuous Integers To Ordinal** once a Type node has been instantiated. See the topic "Converting Continuous Data" on page 126 for more information.
- Using the tools menu button, you can generate a Filter node to discard selected fields.
- Using the sunglasses toggle buttons, you can set the default for all fields to Read or Pass. The Types tab in the source node passes fields by default, while the Type node itself reads values by default.
- Using the **Clear Values** button, you can clear changes to field values made in this node (non-inherited values) and reread values from upstream operations. This option is useful for resetting changes that you may have made for specific fields upstream.
- Using the **Clear All Values** button, you can reset values for **all** fields read into the node. This option effectively sets the **Values** column to **Read** for all fields. This option is useful to reset values for all fields and reread values and types from upstream operations.
- Using the context menu, you can choose to **Copy** attributes from one field to another. See the topic "Copying Type Attributes" on page 132 for more information.
- Using the **View unused field settings** option, you can view type settings for fields that are no longer present in the data or were once connected to this Type node. This is useful when reusing a Type node for datasets that have changed.

Measurement levels

Measurement level (formerly known as "data type" or "usage type") describes the usage of the data fields in IBM SPSS Modeler. The measurement level can be specified on the Types tab of a source or Type node.

For example, you may want to set the measurement level for an integer field with values of 1 and 0 to *Flag*. This usually indicates that 1 = *True* and 0 = *False*.

Storage versus measurement. Note that the measurement level of a field is different from its storage type, which indicates whether data are stored as a string, integer, real number, date, time, or timestamp. While data types can be modified at any point in a stream using a Type node, storage must be determined at the source when reading data into IBM SPSS Modeler (although it can subsequently be changed using a conversion function). See the topic "Setting Field Storage and Formatting" on page 6 for more information.

Some modeling nodes indicate the permitted measurement level types for their input and target fields by means of icons on their Fields tab.

Measurement level icons

Table 21. Measurement level icons

Icon	Measurement level
	Default
	Continuous
	Categorical
	Flag
	Nominal
	Ordinal
	Typeless
[ ]	Collection
	Geospatial

The following measurement levels are available:

- **Default** Data whose storage type and values are unknown (for example, because they have not yet been read) are displayed as **<Default>**.
- **Continuous** Used to describe numeric values, such as a range of 0–100 or 0.75–1.25. A continuous value can be an integer, real number, or date/time.
- **Categorical** Used for string values when an exact number of distinct values is unknown. This is an **uninstantiated** data type, meaning that all possible information about the storage and usage of the data is not yet known. Once data have been read, the measurement level will be *Flag*, *Nominal*, or *Typeless*, depending on the maximum number of members for nominal fields specified in the Stream Properties dialog box.
- **Flag** Used for data with two distinct values that indicate the presence or absence of a trait, such as `true` and `false`, `Yes` and `No` or `0` and `1`. The values used may vary, but one must always be designated as the "true" value, and the other as the "false" value. Data may be represented as text, integer, real number, date, time, or timestamp.
- **Nominal** Used to describe data with multiple distinct values, each treated as a member of a set, such as `small/medium/large`. Nominal data can have any storage—numeric, string, or date/time. Note that setting the measurement level to *Nominal* does not automatically change the values to string storage.
- **Ordinal** Used to describe data with multiple distinct values that have an inherent order. For example, salary categories or satisfaction rankings can be typed as ordinal data. The order is defined by the natural sort order of the data elements. For example, `1, 3, 5` is the default sort order for a set of

integers, while `HIGH, LOW, NORMAL` (ascending alphabetically) is the order for a set of strings. The ordinal measurement level enables you to define a set of categorical data as ordinal data for the purposes of visualization, model building, and export to other applications (such as IBM SPSS Statistics) that recognize ordinal data as a distinct type. You can use an ordinal field anywhere that a nominal field can be used. Additionally, fields of any storage type (real, integer, string, date, time, and so on) can be defined as ordinal.

- **Typeless** Used for data that does not conform to any of the above types, for fields with a single value, or for nominal data where the set has more members than the defined maximum. It is also useful for cases in which the measurement level would otherwise be a set with many members (such as an account number). When you select **Typeless** for a field, the role is automatically set to **None**, with **Record ID** as the only alternative. The default maximum size for sets is 250 unique values. This number can be adjusted or disabled on the Options tab of the Stream Properties dialog box, which can be accessed from the Tools menu.
- **Collection** Used to identify non-geospatial data that is recorded in a list. A collection is effectively a list field of zero depth, where the elements in that list have one of the other measurement levels.

  For more information about lists, see the List Storage and Associated Measurement Levels topic in the Source Nodes section of the SPSS Modeler Source, Process, and Output Nodes Guide.
- **Geospatial** Used with the List storage type to identify geospatial data. Lists can be either List of Integer or List of Real fields with a list depth that is between zero and two, inclusive.

  For more information, see the Geospatial Measurement Sublevels topic in the Type Node section of the SPSS Modeler Source, Process, and Output Nodes Guide.

You can manually specify measurement levels, or you can allow the software to read the data and determine the measurement level based on the values that it reads.

Alternatively, where you have several continuous data fields that should be treated as categorical data, you can choose an option to convert them. See the topic "Converting Continuous Data" on page 126 for more information.

To use auto-typing

1. In either a Type node or the Types tab of a source node, set the **Values** column to **<Read>** for the desired fields. This will make metadata available to all nodes downstream. You can quickly set all fields to **<Read>** or **<Pass>** using the sunglasses buttons on the dialog box.
2. Click **Read Values** to read values from the data source immediately.

To manually set the measurement level for a field

1. Select a field in the table.
2. From the drop-down list in the **Measurement** column, select a measurement level for the field.
3. Alternatively, you can use Ctrl-A or Ctrl-click to select multiple fields before using the drop-down list to select a measurement level.

Geospatial measurement sublevels

The Geospatial measurement level, which is used with the List storage type, has six sublevels that are used to identify different types of geospatial data.

- **Point** - Identifies a specific location; for example, the center of a city.
- **Polygon** - A series of points that identifies the single boundary of a region and its location; for example, a county.
- **LineString** - Also referred to as a Polyline or just a Line, a LineString is a series of points that identifies the route of a line. For example, a LineString might be a fixed item, such as road, river, or railway; or the track of something that moves, such as an aircraft's flight path or a ship's voyage.

- **MultiPoint** - Used when each row in your data contains multiple points per region. For example, if each row represents a city street, the multiple points for each street can be used to identify every street lamp.
- **MultiPolygon** - Used when each row in your data contains several polygons. For example, if each row represents the outline of a country, the US can be recorded as several polygons to identify the different areas such as the mainland, Alaska, and Hawaii.
- **MultiLineString** - Used when each row in your data contains several lines. Because lines cannot branch you can use a MultiLineString to identify a group of lines. For example, data such as the navigable waterways or the railway network in each country.

These measurement sublevels are used with the List storage type. For more information, see "List storage and associated measurement levels" on page 8.

Restrictions

You must be aware of some restrictions when you use geospatial data.

- The coordinate system can affect the format of the data. For example, a Projected coordinate system uses the coordinate values x, y, and (when required) z, whereas a Geographic coordinate system uses the coordinate values longitude, latitude, and (when required) a value for either altitude or depth.

  For more information about coordinate systems, see the Setting Geospatial Options for Streams topic in the Working with Streams section of the SPSS Modeler User's Guide.
- A LineString cannot cross over itself.
- A Polygon is not self-closing; for each Polygon you must ensure that the first and last points are defined as the same.
- The direction of the data in a MultiPolygon is important; clockwise indicates a solid form and clounterclockwise indicates a hole. For example, if you record an area of a country that contains lakes; the main land area border can be recorded in a clockwise direction and the shape of each lake in a clounterclockwise direction.
- A Polygon cannot intersect with itself. An example of this intersection would be if you tried to plot the boundary of the polygon as a continuous line in the form of the figure 8.
- MultiPolygons cannot overlap each other.
- For Geospatial fields, the only relevant storage types are **Real** and **Integer** (the default setting is **Real**).

Geospatial measurement sublevel icons

Table 22. Geospatial measurement sublevel icons

Icon	Measurement level
	Point
	Polygon
	LineString
	MultiPoint
	MultiPolygon
	MultiLineString

Converting Continuous Data

Treating categorical data as continuous can have a serious effect on the quality of a model, especially if it's the target field; for example, producing a regression model rather than a binary model. To prevent this you can convert integer ranges to categorical types such as *Ordinal* or *Flag*.

1. From the Operations and Generate menu button (with the tool symbol), select **Convert Continuous Integers To Ordinal**. The conversion values dialog is displayed.
2. Specify the size of range that will be automatically converted; this applies to any range up to and including the size you enter.
3. Click **OK**. The affected ranges are converted to either *Flag* or *Ordinal* and displayed on the Types tab of the Type node.

Results of the Conversion

- Where a *Continuous* field with integer storage is changed to *Ordinal*, the lower and upper values are expanded to include all of the integer values from the lower value to the upper value. For example, if the range is 1, 5, the set of values is 1, 2, 3, 4, 5.
- If the *Continuous* field changes to *Flag*, the lower and upper values become the false and true values of the flag field.

What is instantiation?

Instantiation is the process of reading or specifying information, such as storage type and values for a data field. To optimize system resources, instantiating is a user-directed process—you tell the software to read values by specifying options on the Types tab in a source node or by running data through a Type node.

- Data with unknown types are also referred to as *uninstantiated*. Data whose storage type and values are unknown are displayed in the *Measurement* column of the Types tab as **<Default>**.
- When you have some information about a field's storage, such as string or numeric, the data are called *partially instantiated*. **Categorical** or **Continuous** are partially instantiated measurement levels. For example, **Categorical** specifies that the field is symbolic, but you don't know whether it is nominal, ordinal, or flag.
- When all of the details about a type are known, including the values, a *fully instantiated* measurement level—nominal, ordinal, flag, or continuous—is displayed in this column. Note that the *continuous* type is used for both partially instantiated and fully instantiated data fields. Continuous data can be either integers or real numbers.

During the execution of a data stream with a Type node, uninstantiated types immediately become partially instantiated, based on the initial data values. Once all of the data have passed through the node, all data become fully instantiated unless values were set to **<Pass>**. If execution is interrupted, the data will remain partially instantiated. Once the Types tab has been instantiated, the values of a field are static at that point in the stream. This means that any upstream changes will not affect the values of a particular field, even if you rerun the stream. To change or update the values based on new data or added manipulations, you need to edit them in the Types tab itself or set the value for a field to **<Read>** or **<Read +>**.

When to instantiate

Generally, if your dataset is not very large and you do not plan to add fields later in the stream, instantiating at the source node is the most convenient method. However, instantiating in a separate Type node is useful when:

- The dataset is large, and the stream filters a subset prior to the Type node.
- Data have been filtered in the stream.
- Data have been merged or appended in the stream.
- New data fields are derived during processing.

Note: If you export data in a database export node, the data must be fully instantiated.

Data Values

Using the **Values** column of the Types tab, you can read values automatically from the data, or you can specify measurement levels and values in a separate dialog box.

The options available from the Values drop-down list provide instructions for auto-typing, as shown in the following table.

Table 23. Instructions for auto-typing

Option	Function
<Read>	Data is read when the node is executed.
<Read+>	Data is read and appended to the current data (if any exist).
<Pass>	No data is read.
<Current>	Keep current data values.
Specify...	A separate dialog box opens for you to specify values and measurement level options.

Executing a Type node or clicking **Read Values** auto-types and reads values from your data source based on your selection. These values can also be specified manually using the Specify option or by double-clicking a cell in the **Field** column.

After you make changes for fields in the Type node, you can reset value information using the following buttons on the dialog box toolbar:

- Using the **Clear Values** button, you can clear changes to field values made in this node (non-inherited values) and reread values from upstream operations. This option is useful for resetting changes that you may have made for specific fields upstream.
- Using the **Clear All Values** button, you can reset values for **all** fields read into the node. This option effectively sets the *Values* column to **Read** for all fields. This option is useful to reset values for all fields and reread values and measurement levels from upstream operations.

Grey text in the Values column

Within either a Type node, or a Source node, if the data in the **Values** column is shown in black text, it indicates that the values of that field have been read and are stored within that node. If no black text is present in this field, the values of that field have not been read and are determined further upstream.

There are occasions when you can see the data as gray text. This occurs when SPSS Modeler can identify or infer the valid values of a field without actually reading and storing the data. This is likely to occur if you use one of the following nodes:

- User Input node. Because the data is defined within the node, the range of values for a field is always known, even if the values have not been stored in the node.
- Statistics File source node. If there is metadata present for the data types it enables SPSS Modeler to infer the possible range of values without reading or storing the data.

In either node, the values are shown in gray text until you click **Read Values**.

Warning: If you do not instantiate the data in your stream, and your data values are shown in gray, any checking of type values that you set in the **Check** column is not applied.

Using the Values Dialog Box

Clicking the **Values** or **Missing** column of the Types tab displays a drop-down list of predefined values. Choosing the **Specify...** option on this list opens a separate dialog box where you can set options for reading, specifying, labeling, and handling values for the selected field.

Many of the controls are common to all types of data. These common controls are discussed here.

Measurement Displays the currently selected measurement level. You can change the setting to reflect the way that you intend to use data. For instance, if a field called `day_of_week` contains numbers that represent individual days, you might want to change this to nominal data in order to create a distribution node that examines each category individually.

Storage Displays the storage type if known. Storage types are unaffected by the measurement level that you choose. To alter the storage type, you can use the Data tab in Fixed File and Variable File source nodes, or a conversion function in a Filler node.

Model Field For fields generated as a result of scoring a model nugget, model field details can also be viewed. These include the name of the target field as well as the role of the field in modeling (whether a predicted value, probability, propensity, and so on).

Values Select a method to determine values for the selected field. Selections that you make here override any selections that you made earlier from the **Values** column of the Type node dialog box. Choices for reading values include the following:

- **Read from data** Select to read values when the node is executed. This option is the same as **<Read>**.
- **Pass** Select not to read data for the current field. This option is the same as **<Pass>**.
- **Specify values and labels** Options here are used to specify values and labels for the selected field. Used with value checking, use this option to specify values that are based on your knowledge of the current field. This option activates unique controls for each type of field. Options for values and labels are covered individually in subsequent topics.

  Note: You cannot specify values or labels for a field whose measurement level is Typeless or <Default>.
- **Extend values from data** Select to append the current data with the values that you enter here. For example, if `field_1` has a range from (0,10), and you enter a range of values from (8,16), the range is extended by adding the 16, without removing the original minimum. The new range would be (0,16). Choosing this option automatically sets the auto-typing option to **<Read+>**.

Max list length Only available for data with a measurement level of either Geospatial or Collection. Set the maximum length of the list by specifying the number of elements the list can contain.

Max string length Only available for typeless data; use this field when you are generating SQL to create a table. Enter the value of the largest string in your data; this generates a column in the table that is big enough for the string. If the string length value is not available, a default string size is used that may not be appropriate for the data (for example, if the value is too small, errors can occur when writing data to the table; too large a value could adversely affect performance.

Check values Select a method of coercing values to conform to the specified continuous, flag, or nominal values. This option corresponds to the **Check** column in the Type node dialog box, and settings made here override those in the dialog box. Used with the **Specify values and labels** option, value checking allows you to conform values in the data with expected values. For example, if you specify values as 1, 0 and then use the **Discard** option, you can discard all records with values other than 1 or 0.

Define blanks Select to activate the following controls that you use to declare missing values or blanks in your data.

- **Missing values** Use this table to define specific values (such as 99 or 0) as blanks. The value should be appropriate for the storage type of the field.
- **Range** Used to specify a range of missing values, for example, ages 1–17 or greater than 65. If a bound value is left blank, then the range is unbounded; for example, if a lower bound of 100 is specified with no upper bound, then all values greater than or equal to 100 is defined as missing. The bound values are inclusive; for example, a range with a lower bound of 5 and an upper bound of 10 includes 5 and 10 in the range definition. A missing value range can be defined for any storage type, including date/time and string (in which case the alphabetic sort order is used to determine whether a value is within the range).
- **Null/White space** You can also specify system nulls (displayed in the data as `$null$`) and white space (string values with no visible characters) as blanks.

Note: The Type node also treats empty strings as white space for purposes of analysis, although they are stored differently internally and may be handled differently in certain cases.

Note: To code blanks as undefined or `$null$`, use the Filler node.

Description Use this text box to specify a field label. These labels appear in various locations, such as in graphs, tables, output, and model browsers, depending on selections you make in the **Stream Properties** dialog box.

Specifying Values and Labels for Continuous Data

The *Continuous* measurement level is used for numeric fields. There are three storage types for continuous data:

- Real
- Integer
- Date/Time

The same dialog box is used to edit all continuous fields; the storage type is displayed for reference only.

Specifying Values

The following controls are unique to continuous fields and are used to specify a range of values:

Lower. Specify a lower limit for the value range.

Upper. Specify an upper limit for the value range.

Specifying Labels

You can specify labels for any value of a range field. Click the **Labels** button to open a separate dialog box for specifying value labels.

Values and Labels Subdialog Box

Clicking **Labels** in the Values dialog box for a range field opens a new dialog box in which you can specify labels for any value in the range.

You can use the *Values* and *Labels* columns in this table to define value and label pairs. Currently defined pairs are shown here. You can add new label pairs by clicking in an empty cell and entering a value and its corresponding label. *Note*: Adding value/value-label pairs to this table will not cause any new values to be added to the field. Instead, it simply creates metadata for the field value.

The labels that you specify in the Type node are displayed in many places (as ToolTips, output labels, and so on), depending on selections that you make in the stream properties dialog box.

Specifying Values and Labels for Nominal and Ordinal Data

Nominal (set) and ordinal (ordered set) measurement levels indicate that the data values are used discretely as a member of the set. The storage types for a set can be string, integer, real number, or date/time.

The following controls are unique to nominal and ordinal fields and are used to specify values and labels:

Values. The *Values* column in the table allows you to specify values based on your knowledge of the current field. Using this table, you can enter expected values for the field and check the dataset's conformity to these values using the Check Values drop-down list. Using the arrow and delete buttons, you can modify existing values as well as reorder or delete values.

Labels. The *Labels* column enables you to specify labels for each value in the set. These labels appear in a variety of locations, such as graphs, tables, output, and model browsers, depending on selections that you make in the stream properties dialog box.

Specifying Values for a Flag

Flag fields are used to display data that have two distinct values. The storage types for flags can be string, integer, real number, or date/time.

True. Specify a flag value for the field when the condition is met.

False. Specify a flag value for the field when the condition is not met.

Labels. Specify labels for each value in the flag field. These labels appear in a variety of locations, such as graphs, tables, output, and model browsers, depending on selections that you make in the stream properties dialog box.

Specifying Values for Collection Data

Collection fields are used to display non-geospatial data that is in a list.

The only item that you can set for the Collection **Measurement** level is the **List measure**. By default this measure is set to Typeless but you can select another value from to set the measurement level of the elements within the list. You can choose one of the following options:

- Typeless
- Continuous
- Nominal
- Ordinal
- Flag

Specifying values for geospatial data

Geospatial fields are used to display geospatial data that is in a list.

For the Geospatial **Measurement** level, you can set the following options to set the measurement level of the elements within the list:

Type Select the measurement sublevel of the geospatial field. The available sublevels are determined by the depth of the list field; the defaults are: Point (zero depth), LineString (depth of one), and Polygon (depth of one).

For more information about sublevels, see “Geospatial measurement sublevels” on page 124.

For more information about list depths, see “List storage and associated measurement levels” on page 8.

Coordinate system This option is only available if you changed the measurement level to geospatial from a non-geospatial level. To apply a coordinate system to your geospatial data, select this check box. By default, the coordinate system set in the **Tools** > **Stream Properties** > **Options** > **Geospatial** pane is shown. To use a different coordinate system, click the **Change** button to display the **Select Coordinate System** dialog box and choose the system that you require.

For more information about coordinate systems, see the Setting Geospatial Options for Streams topic in the Working with Streams section of the SPSS Modeler User's Guide.

Defining Missing Values

The **Missing** column of the Types tab indicates whether missing value handling has been defined for a field. The possible settings are:

On (*). Indicates that missing values handling is defined for this field. This could be done by means of a downstream Filler node, or through an explicit specification using the Specify option (see below).

Off. The field has no missing value handling defined.

Specify. Choose this option to display a dialog where you can declare explicit values to be considered as missing values for this field.

Checking Type Values

Turning on the Check option for each field examines all values in that field to determine whether they comply with the current type settings or the values that you have specified in the Specify Values dialog box. This is useful for cleaning up datasets and reducing the size of a dataset within a single operation.

The setting of the *Check* column in the Type node dialog box determines what happens when a value outside of the type limits is discovered. To change the Check settings for a field, use the drop-down list for that field in the *Check* column. To set the Check settings for all fields, click in the *Field* column and press Ctrl-A. Then use the drop-down list for any field in the *Check* column.

The following Check settings are available:

None. Values will be passed through without checking. This is the default setting.

Nullify. Change values outside of the limits to the system null (`$null$`).

Coerce. Fields whose measurement levels are fully instantiated will be checked for values that fall outside the specified ranges. Unspecified values will be converted to a legal value for that measurement level using the following rules:

- For flags, any value other than the true and false value is converted to the false value.
- For sets (nominal or ordinal), any unknown value is converted to the first member of the set's values.
- Numbers greater than the upper limit of a range are replaced by the upper limit.
- Numbers less than the lower limit of a range are replaced by the lower limit.
- Null values in a range are given the midpoint value for that range.

Discard. When illegal values are found, the entire record is discarded.

Warn. The number of illegal items is counted and reported in the stream properties dialog box when all of the data have been read.

Abort. The first illegal value encountered terminates the running of the stream. The error is reported in the stream properties dialog box.

Setting the field role

The role of a field specifies how it is used in model building—for example, whether a field is an input or target (the thing being predicted).

Note: The Partition, Frequency and Record ID roles can each be applied to a single field only.

The following roles are available:

Input. The field will be used as an input to machine learning (a predictor field).

Target. The field will be used as an output or target for machine learning (one of the fields that the model will try to predict).

Both. The field will be used as both an input and an output by the Apriori node. All other modeling nodes will ignore the field.

None. The field will be ignored by machine learning. Fields whose measurement level has been set to **Typeless** are automatically set to **None** in the **Role** column.

Partition. Indicates a field used to partition the data into separate samples for training, testing, and (optional) validation purposes. The field must be an instantiated set type with two or three possible values (as defined in the Field Values dialog box). The first value represents the training sample, the second represents the testing sample, and the third (if present) represents the validation sample. Any additional values are ignored, and flag fields cannot be used. Note that to use the partition in an analysis, partitioning must be enabled on the Model Options tab in the appropriate model-building or analysis node. Records with null values for the partition field are excluded from the analysis when partitioning is enabled. If multiple partition fields have been defined in the stream, a single partition field must be specified on the Fields tab in each applicable modeling node. If a suitable field doesn't already exist in

your data, you can create one using a Partition node or Derive node. See the topic "Partition Node" on page 155 for more information.

Split. (Nominal, ordinal and flag fields only) Specifies that a model is to be built for each possible value of the field.

Frequency. (Numeric fields only) Setting this role enables the field value to be used as a frequency weighting factor for the record. This feature is supported by C&R Tree, CHAID, QUEST and Linear models only; all other nodes ignore this role. Frequency weighting is enabled by means of the **Use frequency weight** option on the Fields tab of those modeling nodes that support the feature.

Record ID. The field will be used as the unique record identifier. This feature is ignored by most nodes; however it is supported by Linear models, and is required for the IBM Netezza in-database mining nodes.

Copying Type Attributes

You can easily copy the attributes of a type, such as values, checking options, and missing values from one field to another:

1. Right-click on the field whose attributes you want to copy.
2. From the context menu, choose **Copy**.
3. Right-click on the field(s) whose attributes you want to change.
4. From the context menu, choose **Paste Special**. *Note*: You can select multiple fields using the Ctrl-click method or by using the **Select Fields** option from the context menu.

A new dialog box opens, from which you can select the specific attributes that you want to paste. If you are pasting into multiple fields, the options that you select here will apply to all target fields.

Paste the following attributes. Select from the list below to paste attributes from one field to another.

- **Type.** Select to paste the measurement level.
- **Values.** Select to paste the field values.
- **Missing.** Select to paste missing value settings.
- **Check.** Select to paste value checking options.
- **Role.** Select to paste the role of a field.

Field Format Settings Tab

The Format tab on the Table and Type nodes shows a list of current or unused fields and formatting options for each field. Following is a description of each column in the field formatting table:

Field. This shows the name of the selected field.

Format. By double-clicking a cell in this column, you can specify formatting for fields on an individual basis using the dialog box that opens. See the topic "Setting Field Format options" on page 133 for more information. Formatting specified here overrides formatting specified in the overall stream properties.

Note: The Statistics Export and Statistics Output nodes export *.sav* files that include per-field formatting in their metadata. If a per-field format is specified that is not supported by the IBM SPSS Statistics *.sav* file format, then the node will use the IBM SPSS Statistics default format.

Justify. Use this column to specify how the values should be justified within the table column. The default setting is **Auto**, which left-justifies symbolic values and right-justifies numeric values. You can override the default by selecting **Left**, **Right**, or **Center**.

Column Width. By default, column widths are automatically calculated based on the values of the field. To override the automatic width calculation, click a table cell and use the drop-down list to select a new width. To enter a custom width not listed here, open the Field Formats subdialog box by double-clicking a table cell in the Field or Format column. Alternatively, you can right-click on a cell and choose **Set Format**.

View Current Fields. By default, the dialog box shows the list of currently active fields. To view the list of unused fields, select **View unused fields settings**.

Context Menu. The context menu for this tab provides various selection and setting update options. Right-click in a column to display this menu.

- **Select All.** Selects all fields.
- **Select None.** Clears the selection.
- **Select Fields.** Selects fields based on type or storage characteristics. Options are **Select Categorical**, **Select Continuous** (numeric), **Select Typeless**, **Select Strings**, **Select Numbers**, or **Select Date/Time**. See the topic "Measurement levels" on page 122 for more information.
- **Set Format.** Opens a subdialog box for specifying date, time, and decimal options on a per-field basis.
- **Set Justify.** Sets the justification for the selected field(s). Options are **Auto**, **Center**, **Left**, or **Right**.
- **Set Column Width.** Sets the field width for selected fields. Specify **Auto** to read width from the data. Or you can set field width to 5, 10, 20, 30, 50, 100, or 200.

Setting Field Format options

Field formatting is specified on a subdialog box available from the Format tab on the Type and Table nodes. If you have selected more than one field before opening this dialog box, then settings from the first field in the selection are used for all. Clicking **OK** after making specifications here will apply these settings to all fields selected on the Format tab.

The following options are available on a per-field basis. Many of these settings can also be specified in the stream properties dialog box. Any settings made at the field level override the default specified for the stream.

Date format. Select a date format to be used for date storage fields or when strings are interpreted as dates by CLEM date functions.

Time format. Select a time format to be used for time storage fields or when strings are interpreted as times by CLEM time functions.

Number display format. You can choose from standard (**####.###**), scientific (**#.###E+##**), or currency display formats (**$###.##**).

Decimal symbol. Select either a comma (,) or period (.) as a decimal separator.

Grouping symbol. For number display formats, select the symbol used to group values (For example, the comma in *3,000.00*). Options include none, period, comma, space, and locale-defined (in which case the default for the current locale is used).

Decimal places (standard, scientific, currency, export). For number display formats, specifies the number of decimal places to be used when displaying real numbers. This option is specified separately for each display format. Note that the **Export decimal places** setting only applies to flat file exports, and it overrides the stream properties. The stream default for flat file export is whatever is specified for the **Standard decimal places** setting in the stream properties. The number of decimal places exported by the XML Export node is always 6.

Justify. Specifies how the values should be justified within the column. The default setting is **Auto**, which left-justifies symbolic values and right-justifies numeric values. You can override the default by selecting left, right, or center.

Column width. By default, column widths are automatically calculated based on the values of the field. You can specify a custom width in intervals of five using the arrows to the right of the list box.

Filtering or renaming fields

You can rename or exclude fields at any point in a stream. For example, as a medical researcher, you may not be concerned about the potassium level (field-level data) of patients (record-level data); therefore, you can filter out the K (potassium) field. This can be done using a separate Filter node or using the Filter tab on a source or output node. The functionality is the same regardless of which node it is accessed from.

- From source nodes, such as Variable File, Fixed File, Statistics File, XML, or Extension Import, you can rename or filter fields as the data are read into IBM SPSS Modeler.
- Using a Filter node, you can rename or filter fields at any point in the stream.
- From Statistics Export, Statistics Transform, Statistics Model, and Statistics Output nodes, you can filter or rename fields to conform to IBM SPSS Statistics naming standards.
- You can use the Filter tab in any of the above nodes to define or edit multiple response sets.
- Finally, you can use a Filter node to map fields from one source node to another.

Setting filtering options

The table used on the Filter tab shows the name of each field as it comes into the node as well as the name of each field as it leaves. You can use the options in this table to rename or filter out fields that are duplicates or are unnecessary for downstream operations.

- **Field.** Displays the input fields from currently connected data sources.
- **Filter.** Displays the filter status of all input fields. Filtered fields include a red X in this column, indicating that this field will not be passed downstream. Click in the *Filter* column for a selected field to turn filtering on and off. You can also select options for multiple fields simultaneously using the Shift-click method of selection.
- **Field.** Displays the fields as they leave the Filter node. Duplicate names are displayed in red. You can edit field names by clicking in this column and entering a new name. Or, remove fields by clicking in the *Filter* column to disable duplicate fields.

All columns in the table can be sorted by clicking on the column header.

View current fields. Select to view fields for datasets actively connected to the Filter node. This option is selected by default and is the most common method of using Filter nodes.

View unused field settings. Select to view fields for datasets that were once but are no longer connected to the Filter node. This option is useful when copying Filter nodes from one stream to another or when saving and reloading Filter nodes.

Filter Button Menu

Click the Filter button in the upper left corner of the dialog box to access a menu that provides a number of shortcuts and other options.

You can choose to:

- Remove all fields.
- Include all fields.
- Toggle all fields.
- Remove duplicates. Note that selecting this option removes all occurrences of the duplicate name, including the first one.
- Rename fields and multiple response sets to conform with other applications. See the topic “Renaming or Filtering Fields for IBM SPSS Statistics ” on page 328 for more information.
- Truncate field names.
- Anonymize field and multiple response set names.
- Use input field names.
- Edit Multiple Response Sets. See the topic “Editing Multiple Response Sets” on page 135 for more information.
- Set the default filter state.

You can also use the arrow toggle buttons at the top of the dialog box to specify whether you want to include or discard fields by default. This is useful for large datasets where only a few fields are to be

included downstream. For example, you can select only the fields you want to keep and specify that all others should be discarded (rather than individually selecting all of the fields to discard).

Truncating Field Names

From the Filter button menu (upper left corner of the Filter tab), you can choose to truncate field names.

Maximum length. Specify a number of characters to limit the length of field names.

Number of digits. If field names, when truncated, are no longer unique, they will be further truncated and differentiated by adding digits to the name. You can specify the number of digits used. Use the arrow buttons to adjust the number.

For example, the following table illustrates how field names in a medical dataset are truncated using the default settings (maximum length=8 and number of digits=2).

Table 24. Field name truncation

Field Names	Truncated Field Names
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

Anonymizing Field Names

You can anonymize field names from any node that includes a Filter tab by clicking the Filter button menu in the upper left corner and choosing **Anonymize Field Names**. Anonymized field names consist of a string prefix plus a unique numeric-based value.

Anonymize names of. Choose **Selected fields only** to anonymize only the names of fields already selected on the Filter tab. The default is **All fields**, which anonymizes all field names.

Field names prefix. The default prefix for anonymized field names is **anon_**; choose **Custom** and type your own prefix if you want a different one.

Anonymize multiple response sets. Anonymizes the names of multiple response sets in the same manner as fields. See the topic “Editing Multiple Response Sets” on page 135 for more information.

To restore the original field names, choose **Use Input Field Names** from the filter button menu.

Editing Multiple Response Sets

You can add or edit multiple response sets from any node that includes a Filter tab by clicking the Filter button menu in the upper left corner and choosing **Edit Multiple Response Sets**.

Multiple response sets are used to record data that can have more than one value for each case—for example, when asking survey respondents which museums they have visited or which magazines they read. Multiple response sets can be imported into IBM SPSS Modeler using a Data Collection source node or a Statistics File source node and can be defined in IBM SPSS Modeler using a Filter node.

Click **New** to create a new multiple response set, or click **Edit** to modify an existing set.

Name and label. Specifies the name and description for the set.

Type. Multiple response questions can be handled in one of two ways:

- **Multiple dichotomy set.** A separate flag field is created for each possible response; thus, if there are 10 magazines, there are 10 flag fields, each of which can have values, such as 0 or 1 for *true* or *false*. The counted value allows you to specify which value is counted as true. This method is useful when you want to allow respondents to choose all options that apply.

- **Multiple category set.** A nominal field is created for each response up to the maximum number of answers from a given respondent. Each nominal field has values representing the possible answers, such as *1* for *Time*, *2* for *Newsweek*, and *3* for *PC Week*. This method is most useful when you want to limit the number of answers—for example, when asking respondents to choose the three magazines they read most frequently.

Fields in set. Use the icons on the right to add or remove fields.

Comments

- All fields included in a multiple response set must have the same storage.
- Sets are distinct from the fields they include. For example, deleting a set will not cause the fields it includes to be deleted—merely the links between those fields. The set is still visible upstream from the point of deletion but is not visible downstream.
- If fields are renamed using a Filter node (directly on the tab or by choosing the Rename for IBM SPSS Statistics, **Truncate**, or **Anonymize** options on the Filter menu), any references to these fields used in multiple response sets will also be updated. However, any fields in a multiple response set that are dropped by the Filter node will not be removed from the multiple response set. Such fields, although no longer visible in the stream, are still referenced by the multiple response set; this could be a consideration when exporting, for example.

Derive node

One of the most powerful features in IBM SPSS Modeler is the ability to modify data values and derive new fields from existing data. During lengthy data mining projects, it is common to perform several derivations, such as extracting a customer ID from a string of Web log data or creating a customer lifetime value based on transaction and demographic data. All of these transformations can be performed, using a variety of field operations nodes.

Several nodes provide the ability to derive new fields:

The Derive node modifies data values or creates new fields from one or more existing fields. It creates fields of type formula, flag, nominal, state, count, and conditional.

The Reclassify node transforms one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis.

The Binning node automatically creates new nominal (set) fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing groups of income as deviations from the mean. Once you have created bins for the new field, you can generate a Derive node based on the cut points.

The Set to Flag node derives multiple flag fields based on the categorical values defined for one or more nominal fields.

The Restructure node converts a nominal or flag field into a group of fields that can be populated with the values of yet another field. For example, given a field named *payment type*, with values of *credit*, *cash*, and *debit*, three new fields would be created (*credit*, *cash*, *debit*), each of which might contain the value of the actual payment made.

The History node creates new fields containing data from fields in previous records. History nodes are most often used for sequential data, such as time series data. Before using a History node, you may want to sort the data using a Sort node.

Using the Derive node

Using the Derive node, you can create six types of new fields from one or more existing fields:

- **Formula.** The new field is the result of an arbitrary CLEM expression.
- **Flag.** The new field is a flag, representing a specified condition.
- **Nominal.** The new field is nominal, meaning that its members are a group of specified values.
- **State.** The new field is one of two states. Switching between these states is triggered by a specified condition.
- **Count.** The new field is based on the number of times that a condition has been true.
- **Conditional.** The new field is the value of one of two expressions, depending on the value of a condition.

Each of these nodes contains a set of special options in the Derive node dialog box. These options are discussed in subsequent topics.

Note that use of the following may change row order:

- Executing in a database via SQL pushback
- Executing via remote IBM SPSS Analytic Server
- Using functions that run in embedded IBM SPSS Analytic Server
- Deriving a list (for example, see “Deriving a list or geospatial field” on page 139)
- Calling any of the spatial functions

Setting Basic Options for the Derive Node

At the top of the dialog box for Derive nodes are a number of options for selecting the type of Derive node that you need.

Mode. Select **Single** or **Multiple**, depending on whether you want to derive multiple fields. When **Multiple** is selected, the dialog box changes to include options for multiple Derive fields.

Derive field. For simple Derive nodes, specify the name of the field that you want to derive and add to each record. The default name is Derive*N*, where *N* is the number of Derive nodes that you have created thus far during the current session.

Derive as. Select a type of Derive node, such as Formula or Nominal, from the drop-down list. For each type, a new field is created based on the conditions that you specify in the type-specific dialog box.

Selecting an option from the drop-down list will add a new set of controls to the main dialog box according to the properties of each Derive node type.

Field type. Select a measurement level, such as continuous, categorical, or flag, for the newly derived node. This option is common to all forms of Derive nodes.

Note: Deriving new fields often requires the use of special functions or mathematical expressions. To help you create these expressions, an Expression Builder is available from the dialog box for all types of Derive nodes and provides rule checking as well as a complete list of CLEM expressions.

Deriving Multiple Fields

Setting the mode to **Multiple** within a Derive node gives you the capability to derive multiple fields based on the same condition within the same node. This feature saves time when you want to make identical transformations on several fields in your dataset. For example, if you want to build a regression model predicting current salary based on beginning salary and previous experience, it might be beneficial to apply a log transformation to all three skewed variables. Rather than add a new Derive node for each

transformation, you can apply the same function to all fields at once. Simply select all fields from which to derive a new field and then type the derive expression using the `@FIELD` function within the field parentheses.

Note: The `@FIELD` function is an important tool for deriving multiple fields at the same time. It allows you to refer to the contents of the current field or fields without specifying the exact field name. For instance, a CLEM expression used to apply a log transformation to multiple fields is `log(@FIELD)`.

The following options are added to the dialog box when you select **Multiple** mode:

Derive from. Use the Field Chooser to select fields from which to derive new fields. One output field will be generated for each selected field. *Note*: Selected fields do not need to be the same storage type; however, the Derive operation will fail if the condition is not valid for *all* fields.

Field name extension. Type the extension that you would like added to the new field name(s). For example, for a new field containing the log of *Current Salary*, you could add the extension *log_* to the field name, producing *log_Current Salary*. Use the radio buttons to choose whether to add the extension as a prefix (at the beginning) or as a suffix (at the end) of the field name. The default name is Derive*N*, where *N* is the number of Derive nodes that you have created thus far during the current session.

As in the single-mode Derive node, you now need to create an expression to use for deriving a new field. Depending on the type of Derive operation selected, there are a number of options to create a condition. These options are discussed in subsequent topics. To create an expression, you can simply type in the formula field(s) or use the Expression Builder by clicking the calculator button. Remember to use the `@FIELD` function when referring to manipulations on multiple fields.

Selecting Multiple Fields

For all nodes that perform operations on multiple input fields, such as Derive (multiple mode), Aggregate, Sort, Multiplot, and Time Plot, you can easily select multiple fields using the Select Fields dialog box.

Sort by. You can sort available fields for viewing by selecting one of the following options:

- **Natural.** View the order of fields as they have been passed down the data stream into the current node.
- **Name.** Use alphabetical order to sort fields for viewing.
- **Type.** View fields sorted by their measurement level. This option is useful when selecting fields with a particular measurement level.

Select fields from the list one at a time or use the Shift-click and Ctrl-click methods to select multiple fields. You can also use the buttons below the list to select groups of fields based on their measurement level, or to select or deselect all fields in the table.

Setting Derive Formula Options

Derive Formula nodes create a new field for each record in a data set based on the results of a CLEM expression. This expression cannot be conditional. To derive values that are based on a conditional expression, use the flag or conditional type of Derive node.

Formula Specify a formula by using the CLEM language to derive a value for the new field.

Note: Because SPSS Modeler cannot know what submeasurement level is to be used for a derived list field, for the Collection and Geospatial measurement levels, you can click **Specify...** to open the **Values** dialog box and set the required submeasurement level. For more information, see “Setting derived list values” on page 138.

For Geospatial fields, the only relevant storage types are **Real** and **Integer** (the default setting is **Real**).

Setting derived list values

The **Values** dialog box is displayed when you select **Specify...** from the Derive node Formula **Field type** drop-down list. In this dialog box, you set the submeasurement level values to be used for the Formula **Field type** measurement levels of either Collection or Geospatial.

Measurement Select either **Collection** or **Geospatial**. If you select any other measurement level, the dialog box displays a message that there are no editable values.

Collection

The only item that you can set for the Collection **Measurement** level is the **List measure**. By default this measure is set to Typeless but you can select another value to set the measurement level of the elements within the list. You can choose one of the following options:

- Typeless
- Categorical
- Continuous
- Nominal
- Ordinal
- Flag

Geospatial

For the Geospatial **Measurement** level, you can select the following options to set the measurement level of the elements within the list:

Type Select the measurement sublevel of the geospatial field. The available sublevels are determined by the depth of the list field; the defaults are as follows:

- Point (zero depth)
- LineString (depth of one)
- Polygon (depth of one)
- MultiPoint (depth of one)
- MultiLineString (depth of two)
- MultiPolygon (depth of two)

For more information about sublevels, see the Geospatial Measurement Sublevels topic in the Type Node section of the SPSS Modeler Source, Process, and Output Nodes Guide.

For more information about list depths, see the List Storage and Associated Measurement Levels topic in the Source Node section of the SPSS Modeler Source, Process, and Output Nodes Guide.

Coordinate system This option is only available if you changed the measurement level to geospatial from a non-geospatial level. To apply a coordinate system to your geospatial data, select this check box. By default, the coordinate system set in the **Tools** > **Stream Properties** > **Options** > **Geospatial** pane is shown. To use a different coordinate system, click the **Change** button to display the **Select Coordinate System** dialog box and choose the system that matches your data.

For more information about coordinate systems, see the Setting Geospatial Options for Streams topic in the Working with Streams section of the SPSS Modeler User's Guide.

Deriving a list or geospatial field

There are occasions when data that should be recorded as a list item is imported into SPSS Modeler with the wrong attributes. For example, as separate geospatial fields, such as an x coordinate and a y coordinate, or a longitude and latitude, as individual rows in a `.csv` file. In this situation, you must combine the individual fields into a single list field; one way to do this is by using the Derive node.

Note: You must know which is the x (or longitude) field and which is the y (or latitude field) field when you combine geospatial data. You must combine your data so that the resulting list field has the order of elements: [x, y] or [Longitude, Latitude], which are the standard formats for geospatial coordinates.

The following steps show a simple example of deriving a list field.

1. In your stream, attach a Derive node to your Source node.
2. On the Settings tab of the Derive node, select **Formula** from the **Derive as** list.
3. In **Field Type** select either **Collection** for a non-geospatial list, or **Geospatial**. By default, SPSS Modeler uses a "best guess" approach to set the correct list detail; you can select **Specify...** to open the **Values** dialog box. This dialog can be used for a collection to enter further information about the data in your list, and for geospatial it can be used to set the type of the data and specify the coordinate system of the data.

   Note: For geospatial, the coordinate system you specify must exactly match the coordinate system of the data. If this is not the case, geospatial functionality will produce incorrect results.
4. In the **Formula** pane, enter the formula to combine your data into the correct list format. Alternatively, click the calculator button to open the Expression Builder.

   A simple example of a formula to derive a list is [x, y], where x and y are separate fields in the data source. The new derived field that is created is a list where the value for each record is the concatenated x and y values for that record.

   Note: Fields that are combined into a list in this way must have the same storage type.

For more information about lists, and depths of lists, see "List storage and associated measurement levels" on page 8.

Setting Derive Flag Options

Derive Flag nodes are used to indicate a specific condition, such as high blood pressure or customer account inactivity. A flag field is created for each record, and when the true condition is met, the flag value for true is added in the field.

True value. Specify a value to include in the flag field for records that match the condition specified below. The default is T.

False value. Specify a value to include in the flag field for records that do *not* match the condition specified below. The default is F.

True when. Specify a CLEM condition to evaluate certain values of each record and give the record a true value or a false value (defined above). Note that the true value will be given to records in the case of non-false numeric values.

Note: To return an empty string, you should type opening and closing quotes with nothing between them, such as `""`. Empty strings are often used, for example, as the false value in order to enable true values to stand out more clearly in a table. Similarly, quotes should be used if you want a string value that would otherwise be treated as a number

Example

In releases of IBM SPSS Modeler prior to 12.0, multiple responses were imported into a single field, with values separated by commas. For example:

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

To prepare this data for analysis, you could use the `hassubstring` function to generate a separate flag field for each response with an expression such as:

```
hassubstring(museums,"museum_of_design")
```

Setting Derive Nominal Options

Derive Nominal nodes are used to execute a set of CLEM conditions in order to determine which condition each record satisfies. As a condition is met for each record, a value (indicating which set of conditions was met) will be added to the new, derived field.

Default value. Specify a value to be used in the new field if none of the conditions are met.

Set field to. Specify a value to enter in the new field when a particular condition is met. Each value in the list has an associated condition that you specify in the adjacent column.

If this condition is true. Specify a condition for each member in the set field to list. Use the Expression Builder to select from available functions and fields. You can use the arrow and delete buttons to reorder or remove conditions.

A condition works by testing the values of a particular field in the dataset. As each condition is tested, the values specified above will be assigned to the new field to indicate which, if any, condition was met. If none of the conditions are met, the default value is used.

Setting Derive State Options

Derive State nodes are somewhat similar to Derive Flag nodes. A Flag node sets values depending on the fulfillment of a *single* condition for the current record, but a Derive State node can change the values of a field depending on how it fulfills *two independent* conditions. This means that the value will change (turn on or off) as each condition is met.

Initial state. Select whether to give each record of the new field the **On** or **Off** value initially. Note that this value can change as each condition is met.

"On" value. Specify the value for the new field when the On condition is met.

Switch "On" when. Specify a CLEM condition that will change the state to On when the condition is true. Click the calculator button to open the Expression Builder.

"Off" value. Specify the value for the new field when the Off condition is met.

Switch "Off" when. Specify a CLEM condition that will change the state to Off when the condition is false. Click the calculator button to open the Expression Builder.

Note: To specify an empty string, you should type opening and closing quotes with nothing between them, such as " ". Similarly, quotes should be used if you want a string value that would otherwise be treated as a number.

Setting Derive Count Options

A Derive Count node is used to apply a series of conditions to the values of a numeric field in the dataset. As each condition is met, the value of the derived count field is increased by a set increment. This type of Derive node is useful for time series data.

Initial value. Sets a value used on execution for the new field. The initial value must be a numeric constant. Use the arrow buttons to increase or decrease the value.

Increment when. Specify the CLEM condition that, when met, will change the derived value based on the number specified in Increment by. Click the calculator button to open the Expression Builder.

Increment by. Set the value used to increment the count. You can use either a numeric constant or the result of a CLEM expression.

Reset when. Specify a condition that, when met, will reset the derived value to the initial value. Click the calculator button to open the Expression Builder.

Setting Derive Conditional Options

Derive Conditional nodes use a series of If-Then-Else statements to derive the value of the new field.

If. Specify a CLEM condition that will be evaluated for each record upon execution. If the condition is true (or non-false, in the case of numbers), the new field is given the value specified below by the Then expression. Click the calculator button to open the Expression Builder.

Then. Specify a value or CLEM expression for the new field when the If statement above is true (or non-false). Click the calculator button to open the Expression Builder.

Else. Specify a value or CLEM expression for the new field when the If statement above is false. Click the calculator button to open the Expression Builder.

Recoding Values with the Derive Node

Derive nodes can also be used to recode values, for example by converting a string field with categorical values to a numeric nominal (set) field.

1. For Derive As, select the type of field (Nominal, Flag, etc.) as appropriate.
2. Specify the conditions for recoding values. For example, you could set the value to 1 if `Drug='drugA'`, 2 if `Drug='drugB'`, and so on.

Filler node

Filler nodes are used to replace field values and change storage. You can choose to replace values based on a specified CLEM condition, such as `@BLANK(FIELD)`. Alternatively, you can choose to replace all blanks or null values with a specific value. Filler nodes are often used in conjunction with the Type node to replace missing values. For example, you can fill blanks with the mean value of a field by specifying an expression such as `@GLOBAL_MEAN`. This expression will fill all blanks with the mean value as calculated by a Set Globals node.

Fill in fields. Using the Field Chooser (button to the right of the text field), select fields from the dataset whose values will be examined and replaced. The default behavior is to replace values depending on the Condition and Replace with expressions specified below. You can also select an alternative method of replacement using the Replace options below.

Note: When selecting multiple fields to replace with a user-defined value, it is important that the field types are similar (all numeric or all symbolic).

Replace. Select to replace the values of the selected field(s) using one of the following methods:

- **Based on condition.** This option activates the Condition field and Expression Builder for you to create an expression used as a condition for replacement with the value specified.
- **Always.** Replaces all values of the selected field. For example, you could use this option to convert the storage of income to a string using the following CLEM expression: `(to_string(income))`.
- **Blank values.** Replaces all user-specified blank values in the selected field. The standard condition `@BLANK(@FIELD)` is used to select blanks. *Note*: You can define blanks using the Types tab of the source node or with a Type node.
- **Null values.** Replaces all system null values in the selected field. The standard condition `@NULL(@FIELD)` is used to select nulls.
- **Blank and null values.** Replaces both blank values and system nulls in the selected field. This option is useful when you are unsure whether or not nulls have been defined as missing values.

Condition. This option is available when you have selected the **Based on condition** option. Use this text box to specify a CLEM expression for evaluating the selected fields. Click the calculator button to open the Expression Builder.

Replace with. Specify a CLEM expression to give a new value to the selected fields. You can also replace the value with a null value by typing `undef` in the text box. Click the calculator button to open the Expression Builder.

Note: When the field(s) selected are string, you should replace them with a string value. Using the default 0 or another numeric value as the replacement value for string fields will result in an error.

Note that use of the following may change row order:

- Executing in a database via SQL pushback
- Executing via remote IBM SPSS Analytic Server
- Using functions that run in embedded IBM SPSS Analytic Server
- Deriving a list (for example, see "Deriving a list or geospatial field" on page 139)
- Calling any of the spatial functions

Storage Conversion Using the Filler Node

Using the Replace condition of a Filler node, you can easily convert the field storage for single or multiple fields. For example, using the conversion function `to_integer`, you could convert *income* from a string to an integer using the following CLEM expression: `to_integer(income)`.

You can view available conversion functions and automatically create a CLEM expression using the Expression Builder. From the Functions drop-down list, select **Conversion** to view a list of storage conversion functions. The following conversion functions are available:

- `to_integer(ITEM)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ITEM)`
- `to_time(ITEM)`
- `to_timestamp(ITEM)`
- `to_date(ITEM)`
- `to_datetime(ITEM)`

Converting date and time values. Note that conversion functions (and any other functions that require a specific type of input such as a date or time value) depend on the current formats specified in the stream options dialog box. For example if you want to convert a string field with values *Jan 2003*, *Feb 2003*, etc. to date storage, select **MON YYYY** as the default date format for the stream.

Conversion functions are also available from the Derive node, for temporary conversion during a derive calculation. You can also use the Derive node to perform other manipulations such as recoding string fields with categorical values. See the topic "Recoding Values with the Derive Node" on page 142 for more information.

Reclassify Node

The Reclassify node enables the transformation from one set of categorical values to another. Reclassification is useful for collapsing categories or regrouping data for analysis. For example, you could reclassify the values for *Product* into three groups, such as *Kitchenware*, *Bath and Linens*, and *Appliances*. Often, this operation is performed directly from a Distribution node by grouping values and generating a Reclassify node. See the topic "Using a Distribution Node" on page 213 for more information.

Reclassification can be performed for one or more symbolic fields. You can also choose to substitute the new values for the existing field or generate a new field.

When to use a Reclassify node

Before using a Reclassify node, consider whether another Field Operations node is more appropriate for the task at hand:

- To transform numeric ranges into sets using an automatic method, such as ranks or percentiles, you should use a Binning node. See the topic "Binning Node" on page 147 for more information.
- To classify numeric ranges into sets manually, you should use a Derive node. For example, if you want to collapse salary values into specific salary range categories, you should use a Derive node to define each category manually.

- To create one or more flag fields based on the values of a categorical field, such as *Mortgage_type*, you should use a Set to Flag node.
- To convert a categorical field to numeric storage, you can use a Derive node. For example, you could convert *No* and *Yes* values to 0 and 1, respectively. See the topic "Recoding Values with the Derive Node" on page 142 for more information.

Setting Options for the Reclassify Node

There are three steps to using the Reclassify node:

1. First, select whether you want to reclassify multiple fields or a single field.
2. Next, choose whether to recode into the existing field or create a new field.
3. Then, use the dynamic options in the Reclassify node dialog box to map sets as desired.

   Mode. Select **Single** to reclassify the categories for one field. Select **Multiple** to activate options enabling the transformation of more than one field at a time.

   Reclassify into. Select **New field** to keep the original nominal field and derive an additional field containing the reclassified values. Select **Existing field** to overwrite the values in the original field with the new classifications. This is essentially a "fill" operation.

   Once you have specified mode and replacement options, you must select the transformation field and specify the new classification values using the dynamic options on the bottom half of the dialog box. These options vary depending on the mode you have selected above.

   Reclassify field(s). Use the Field Chooser button on the right to select one (Single mode) or more (Multiple mode) categorical fields.

   New field name. Specify a name for the new nominal field containing recoded values. This option is available only in Single mode when **New field** is selected above. When **Existing field** is selected, the original field name is retained. When working in Multiple mode, this option is replaced with controls for specifying an extension added to each new field. See the topic "Reclassifying Multiple Fields" on page 145 for more information.

   Reclassify values. This table enables a clear mapping from old set values to those you specify here.

   - **Original value.** This column lists existing values for the select field(s).
   - **New value.** Use this column to type new category values or select one from the drop-down list. When you automatically generate a Reclassify node using values from a Distribution chart, these values are included in the drop-down list. This allows you to quickly map existing values to a known set of values. For example, healthcare organizations sometimes group diagnoses differently based upon network or locale. After a merger or acquisition, all parties will be required to reclassify new or even existing data in a consistent fashion. Rather than manually typing each target value from a lengthy list, you can read the master list of values in to IBM SPSS Modeler, run a Distribution chart for the *Diagnosis* field, and generate a Reclassify (values) node for this field directly from the chart. This process will make all of the target Diagnosis values available from the New Values drop-down list.
4. Click **Get** to read original values for one or more fields selected above.
5. Click **Copy** to paste original values over to the *New value* column for fields that have not been mapped yet. The unmapped original values are added to the drop-down list.
6. Click **Clear new** to erase all specifications in the *New value* column. *Note*: This option does not erase the values from the drop-down list.
7. Click **Auto** to automatically generate consecutive integers for each of the original values. Only integer values (no real values, such as 1.5, 2.5, and so on) can be generated.

For example, you can automatically generate consecutive product ID numbers for product names or course numbers for university class offerings. This functionality corresponds to the Automatic Recode transformation for sets in IBM SPSS Statistics.

For unspecified values use. This option is used for filling unspecified values in the new field. You can either choose to keep the original value by selecting **Original value** or specify a default value.

Reclassifying Multiple Fields

To map category values for more than one field at a time, set the mode to **Multiple**. This enables new settings in the Reclassify dialog box, which are described below.

Reclassify fields. Use the Field Chooser button on the right to select the fields that you want to transform. Using the Field Chooser, you can select all fields at once or fields of a similar type, such as nominal or flag.

Field name extension. When recoding multiple fields simultaneously, it is more efficient to specify a common extension added to all new fields rather than individual field names. Specify an extension such as `_recode` and select whether to append or prepend this extension to the original field names.

Storage and Measurement Level for Reclassified Fields

The Reclassify node always creates a nominal field from the recode operation. In some cases, this may change the measurement level of the field when using the **Existing field** reclassification mode.

The new field's storage (how data are *stored* rather than how they are *used*) is calculated based on the following Settings tab options:

- If unspecified values are set to use a default value, the storage type is determined by examining both the new values as well as the default value and determining the appropriate storage. For example, if all values can be parsed as integers, the field will have the integer storage type.
- If unspecified values are set to use the original values, the storage type is based on the storage of the original field. If all of the values can be parsed as the storage of the original field, then that storage is preserved; otherwise, the storage is determined by finding the most appropriate storage type encompassing both old and new values. For example, reclassifying an integer set { 1, 2, 3, 4, 5 } with the reclassification 4 => 0, 5 => 0 generates a new integer set { 1, 2, 3, 0 }, whereas with the reclassification 4 => “Over 3”, 5 => “Over 3” will generate the string set { “1”, “2”, “3”, “Over 3” }.

Note: If the original type was uninstantiated, the new type will be also be uninstantiated.

Anonymize Node

The Anonymize node enables you to disguise field names, field values, or both when working with data that are to be included in a model downstream of the node. In this way, the generated model can be freely distributed (for example, to Technical Support) with no danger that unauthorized users will be able to view confidential data, such as employee records or patients' medical records.

Depending on where you place the Anonymize node in the stream, you may need to make changes to other nodes. For example, if you insert an Anonymize node upstream from a Select node, the selection criteria in the Select node will need to be changed if they are acting on values that have now become anonymized.

The method to be used for anonymizing depends on various factors. For field names and all field values except Continuous measurement levels, the data are replaced by a string of the form:

```
prefix_Sn
```

where *prefix_* is either a user-specified string or the default string `anon_`, and *n* is an integer value that starts at 0 and is incremented for each unique value (for example, `anon_S0`, `anon_S1`, etc.).

Field values of type Continuous must be transformed because numeric ranges deal with integer or real values rather than strings. As such, they can be anonymized only by transforming the range into a

different range, thus disguising the original data. Transformation of a value *x* in the range is performed in the following way:

```
A*(x + B)
```

where:

A is a scale factor, which must be greater than 0.

B is a translation offset to be added to the values.

Example

In the case of a field *AGE* where the scale factor *A* is set to 7 and the translation offset *B* is set to 3, the values for *AGE* are transformed into:

```
7*(AGE + 3)
```

Setting Options for the Anonymize Node

Here you can choose which fields are to have their values disguised further downstream.

Note that the data fields must be instantiated upstream from the Anonymize node before anonymize operations can be performed. You can instantiate the data by clicking the **Read Values** button on a Type node or on the Types tab of a source node.

Field. Lists the fields in the current dataset. If any field names have already been anonymized, the anonymized names are shown here.

Measurement. The measurement level of the field.

Anonymize Values. Select one or more fields, click this column, and choose **Yes** to anonymize the field value using the default prefix **anon_**; choose **Specify** to display a dialog box in which you can either enter your own prefix or, in the case of field values of type *Continuous*, specify whether the transformation of field values is to use random or user-specified values. Note that *Continuous* and non-*Continuous* field types cannot be specified in the same operation; you must do this separately for each type of field.

View current fields. Select to view fields for datasets actively connected to the Anonymize node. This option is selected by default.

View unused field settings. Select to view fields for datasets that were once but are no longer connected to the node. This option is useful when copying nodes from one stream to another or when saving and reloading nodes.

Specifying How Field Values Will Be Anonymized

The Replace Values dialog box lets you choose whether to use the default prefix for anonymized field values or to use a custom prefix. Clicking **OK** in this dialog box changes the setting of Anonymize Values on the Settings tab to **Yes** for the selected field or fields.

Field values prefix. The default prefix for anonymized field values is **anon_**; choose **Custom** and enter your own prefix if you want a different one.

The Transform Values dialog box is displayed only for fields of type Continuous and allows you to specify whether the transformation of field values is to use random or user-specified values.

Random. Choose this option to use random values for the transformation. **Set random seed** is selected by default; specify a value in the **Seed** field, or use the default value.

Fixed. Choose this option to specify your own values for the transformation.

- **Scale by.** The number by which field values will be multiplied in the transformation. Minimum value is 1; maximum is normally 10, but this may be lowered to avoid overflow.
- **Translate by.** The number that will be added to field values in the transformation. Minimum value is 0; maximum is normally 1000, but this may be lowered to avoid overflow.

Anonymizing Field Values

Fields that have been selected for anonymization on the Settings tab have their values anonymized:

- When you run the stream containing the Anonymize node
- When you preview the values

To preview the values, click the **Anonymize Values** button on the Anonymized Values tab. Next, select a field name from the drop-down list.

If the measurement level is Continuous, the display shows the:

- Minimum and maximum values of the original range
- Equation used to transform the values

If the measurement level is anything other than Continuous, the screen displays the original and anonymized values for that field.

If the display appears with a yellow background, this indicates that either the setting for the selected field has changed since the last time the values were anonymized or that changes have been made to the data upstream of the Anonymize node such that the anonymized values may no longer be correct. The current set of values is displayed; click the **Anonymize Values** button again to generate a new set of values according to the current setting.

Anonymize Values. Creates anonymized values for the selected field and displays them in the table. If you are using random seeding for a field of type Continuous, clicking this button repeatedly creates a different set of values each time.

Clear Values. Clears the original and anonymized values from the table.

Binning Node

The Binning node enables you to automatically create new nominal fields based on the values of one or more existing continuous (numeric range) fields. For example, you can transform a continuous income field into a new categorical field containing income groups of equal width, or as deviations from the mean. Alternatively, you can select a categorical "supervisor" field in order to preserve the strength of the original association between the two fields.

Binning can be useful for a number of reasons, including:

- **Algorithm requirements.** Certain algorithms, such as Naive Bayes, Logistic Regression, require categorical inputs.
- **Performance.** Algorithms such as multinomial logistic may perform better if the number of distinct values of input fields is reduced. For example, use the median or mean value for each bin rather than using the original values.
- **Data Privacy.** Sensitive personal information, such as salaries, may be reported in ranges rather than actual salary figures in order to protect privacy.

A number of binning methods are available. Once you have created bins for the new field, you can generate a Derive node based on the cut points.

When to use a Binning node

Before using a Binning node, consider whether another technique is more appropriate for the task at hand:

- To manually specify cut points for categories, such as specific predefined salary ranges, use a Derive node. See the topic “Derive node” on page 136 for more information.
- To create new categories for existing sets, use a Reclassify node. See the topic “Reclassify Node” on page 143 for more information.

Missing Value Handling

The Binning node handles missing values in the following ways:

- **User-specified blanks.** Missing values specified as blanks are included during the transformation. For example, if you designated –99 to indicate a blank value using the Type node, this value will be included in the binning process. To ignore blanks during binning, you should use a Filler node to replace the blank values with the system null value.
- **System-missing values ($null$).** Null values are ignored during the binning transformation and remain nulls after the transformation.

The Settings tab provides options for available techniques. The View tab displays cut points established for data previously run through the node.

Setting Options for the Binning Node

Using the Binning node, you can automatically generate bins (categories) using the following techniques:

- Fixed-width binning
- Tiles (equal count or sum)
- Mean and standard deviation
- Ranks
- Optimized relative to a categorical "supervisor" field

The lower part of the dialog box changes dynamically depending on the binning method you select.

Bin fields. Continuous (numeric range) fields pending transformation are displayed here. The Binning node enables you to bin multiple fields simultaneously. Add or remove fields using the buttons on the right.

Binning method. Select the method used to determine cut points for new field bins (categories). Subsequent topics describe the options available in each case.

Bin thresholds. Specify how the bin thresholds are computed.

- **Always recompute.** Cut points and bin allocations are always recomputed when the node is run.
- **Read from Bin Values tab if available.** Cut points and bin allocations are computed only as necessary (for example, when new data has been added).

The following topics discuss options for the available methods of binning.

Fixed-Width Bins

When you choose **Fixed-width** as the binning method, a new set of options is displayed in the dialog box.

Name extension. Specify an extension to use for the generated field(s). *_BIN* is the default extension. You may also specify whether the extension is added to the start (**Prefix**) or end (**Suffix**) of the field name. For example, you could generate a new field called *income_BIN*.

Bin width. Specify a value (integer or real) used to calculate the “width” of the bin. For example, you can use the default value, 10, to bin the field *Age*. Since *Age* has a range from 18–65, the generated bins would be as shown in the following table.

Table 25. Bins for Age with range 18–65

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
>=13 to <23	>=23 to <33	>=33 to <43	>=43 to <53	>=53 to <63	>=63 to <73

The start of bin intervals is calculated using the lowest scanned value minus half of the bin width (as specified). For example, in the bins shown above, 13 is used to start the intervals according to the following calculation: 18 *[lowest data value]* – 5 *[0.5 × (Bin width of 10)]* = 13.

No. of bins. Use this option to specify an integer used to determine the number of fixed-width bins (categories) for the new field(s).

Once you have executed the Binning node in a stream, you can view the bin thresholds generated by clicking the **Preview** tab in the Binning node dialog box. See the topic "Previewing the Generated Bins" on page 152 for more information.

Tiles (Equal Count or Sum)

The tile binning method creates nominal fields that can be used to split scanned records into percentile groups (or quartiles, deciles, and so on) so that each group contains the same number of records, or the sum of the values in each group is equal. Records are ranked in ascending order based on the value of the specified bin field, so that records with the lowest values for the selected bin variable are assigned a rank of 1, the next set of records are ranked 2, and so on. The threshold values for each bin are generated automatically based on the data and tiling method used.

Tile name extension. Specify an extension used for field(s) generated using standard p-tiles. The default extension is *_TILE* plus *N*, where *N* is the tile number. You may also specify whether the extension is added to the start (**Prefix**) or end (**Suffix**) of the field name. For example, you could generate a new field called *income_BIN4*.

Custom tile extension. Specify an extension used for a custom tile range. The default is *_TILEN*. Note that *N* in this case will *not* be replaced by the custom number.

Available p-tiles are:

- **Quartile.** Generate 4 bins, each containing 25% of the cases.
- **Quintile.** Generate 5 bins, each containing 20% of the cases.
- **Decile.** Generate 10 bins, each containing 10% of the cases.
- **Vingtile.** Generate 20 bins, each containing 5% of the cases.
- **Percentile.** Generate 100 bins, each containing 1% of the cases.
- **Custom N.** Select to specify the number of bins. For example, a value of 3 would produce 3 banded categories (2 cut points), each containing 33.3% of the cases.

Note that if there are fewer discrete values in the data than the number of tiles specified, all tiles will not be used. In such cases, the new distribution is likely to reflect the original distribution of your data.

Tiling method. Specifies the method used to assign records to bins.

- **Record count.** Seeks to assign an equal number of records to each bin.
- **Sum of values.** Seeks to assign records to bins such that the sum of the values in each bin is equal. When targeting sales efforts, for example, this method can be used to assign prospects to decile groups based on value per record, with the highest value prospects in the top bin. For example, a pharmaceutical company might rank physicians into decile groups based on the number of prescriptions they write. While each decile would contain approximately the same number of scripts, the number of individuals contributing those scripts would not be the same, with the individuals who write the most scripts concentrated in decile 10. Note that this approach assumes that all values are greater than zero, and may yield unexpected results if this is not the case.

Ties. A tie condition results when values on either side of a cut point are identical. For example, if you are assigning deciles and more than 10% of records have the same value for the bin field, then all of them cannot fit into the same bin without forcing the threshold one way or another. Ties can be moved up to the next bin or kept in the current one but must be resolved so that all records with identical values fall into the same bin, even if this causes some bins to have more records than expected. The thresholds of subsequent bins may also be adjusted as a result, causing values to be assigned differently for the same set of numbers based on the method used to resolve ties.

- **Add to next.** Select to move the tie values up to the next bin.
- **Keep in current.** Keeps tie values in the current (lower) bin. This method may result in fewer total bins being created.
- **Assign randomly.** Select to allocate the tie values randomly to a bin. This attempts to keep the number of records in each bin at an equal amount.

Example: Tiling by Record Count

The following table illustrates how simplified field values are ranked as quartiles when tiling by record count. Note the results vary depending on the selected ties option.

Table 26. Tiling by record count example

Values	Add to Next	Keep in Current
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

The number of items per bin is calculated as:

```
total number of value / number of tiles
```

In the simplified example above, the desired number of items per bin is 1.25 (5 values / 4 quartiles). The value 13 (being value number 2) straddles the 1.25 desired count threshold and is therefore treated differently depending on the selected ties option. In **Add to Next** mode, it is added into bin 2. In **Keep in Current** mode, it is left in bin 1, pushing the range of values for bin 4 outside that of existing data values. As a result, only three bins are created, and the thresholds for each bin are adjusted accordingly, as shown in the following table.

Table 27. Binning example result

Bin	Lower	Upper
1	>=10	<15
2	>=15	<20
3	>=20	<=20

Note: The speed of binning by tiles may benefit from enabling parallel processing.

Rank Cases

When you choose **Ranks** as the binning method, a new set of options is displayed in the dialog box.

Ranking creates new fields containing ranks, fractional ranks, and percentile values for numeric fields depending on the options specified below.

Rank order. Select **Ascending** (lowest value is marked 1) or **Descending** (highest value is marked 1).

Rank. Select to rank cases in ascending or descending order as specified above. The range of values in the new field will be 1–*N*, where *N* is the number of discrete values in the original field. Tied values are given the average of their rank.

Fractional rank. Select to rank cases where the value of the new field equals rank divided by the sum of the weights of the nonmissing cases. Fractional ranks fall in the range of 0–1.

Percentage fractional rank. Each rank is divided by the number of records with valid values and multiplied by 100. Percentage fractional ranks fall in the range of 1–100.

Extension. For all rank options, you can create custom extensions and specify whether the extension is added to the start (**Prefix**) or end (**Suffix**) of the field name. For example, you could generate a new field called *income_P_RANK*.

Mean/Standard Deviation

When you choose **Mean/standard deviation** as the binning method, a new set of options is displayed in the dialog box.

This method generates one or more new fields with banded categories based on the values of the mean and standard deviation of the distribution of the specified field(s). Select the number of deviations to use below.

Name extension. Specify an extension to use for the generated field(s). *_SDBIN* is the default extension. You may also specify whether the extension is added to the start (**Prefix**) or end (**Suffix**) of the field name. For example, you could generate a new field called *income_SDBIN*.

- **+/– 1 standard deviation.** Select to generate three bins.
- **+/– 2 standard deviations.** Select to generate five bins.
- **+/– 3 standard deviations.** Select to generate seven bins.

For example, selecting +/–1 standard deviation results in the three bins as calculated and shown in the following table.

Table 28. Standard deviation bin example

Bin 1	Bin 2	Bin 3
`x < (Mean – Std. Dev)`	`(Mean – Std. Dev) <= x <= (Mean + Std. Dev)`	`x > (Mean + Std. Dev)`

In a normal distribution, 68% of the cases fall within one standard deviation of the mean, 95% within two standard deviations, and 99% within three standard deviations. Note, however, that creating banded categories based on standard deviations may result in some bins being defined outside the actual data range and even outside the range of possible data values (for example, a negative salary range).

Optimal Binning

If the field you want to bin is strongly associated with another categorical field, you can select the categorical field as a "supervisor" field in order to create the bins in such a way as to preserve the strength of the original association between the two fields.

For example, suppose you have used cluster analysis to group states based on delinquency rates for home loans, with the highest rates in the first cluster. In this case, you might choose *Percent past due* and *Percent of foreclosures* as the Bin fields and the cluster membership field generated by the model as the supervisor field.

Name extension Specify an extension to use for the generated field(s) and whether to add it at the start (**Prefix**) or end (**Suffix**) of the field name. For example, you could generate a new field called `pastdue_OPTIMAL` and another called `inforeclosure_OPTIMAL`.

Supervisor field A categorical field used to construct the bins.

Pre-bin fields to improve performance with large datasets Indicates if preprocessing should be used to streamline optimal binning. This groups scale values into a large number of bins using a simple unsupervised binning method, represents values within each bin by the mean, and adjusts the case weight accordingly before proceeding with supervised binning. In practical terms, this method trades a degree of precision for speed and is recommended for large datasets. You can also specify the maximum number of bins that any variable should end up in after preprocessing when this option is used.

Merge bins that have relatively small case counts with a larger neighbor. If enabled, indicates that a bin is merged if the ratio of its size (number of cases) to that of a neighboring bin is smaller than the specified threshold; note that larger thresholds may result in more merging.

Cut Point Settings

The Cut Point Settings dialog box enables you to specify advanced options for the optimal binning algorithm. These options tell the algorithm how to calculate the bins using the target field.

Bin end points. You can specify whether the lower or upper end points should be inclusive (lower <= *x*) or exclusive (lower < *x*).

First and last bins. For both the first and last bin, you can specify whether the bins should be unbounded (extending toward positive or negative infinity) or bounded by the lowest or highest data points.

Previewing the Generated Bins

The Bin Values tab in the Binning node allows you to view the thresholds for generated bins. Using the Generate menu, you can also generate a Derive node that can be used to apply these thresholds from one dataset to another.

Binned field. Use the drop-down list to select a field for viewing. Field names shown use the original field name for clarity.

Tile. Use the drop-down list to select a tile, such as 10 or 100, for viewing. This option is available only when bins have been generated using the tile method (equal count or sum).

Bin thresholds. Threshold values are shown here for each generated bin, along with the number of records that fall into each bin. For the optimal binning method only, the number of records in each bin is shown as a percentage of the whole. Note that thresholds are not applicable when the rank binning method is used.

Read Values. Reads binned values from the dataset. Note that thresholds will also be overwritten when new data are run through the stream.

Generating a Derive Node

You can use the Generate menu to create a Derive node based on the current thresholds. This is useful for applying established bin thresholds from one set of data to another. Furthermore, once these split points are known, a Derive operation is more efficient (meaning faster) than a Binning operation when working with large datasets.

RFM Analysis Node

The Recency, Frequency, Monetary (RFM) Analysis node enables you to determine quantitatively which customers are likely to be the best ones by examining how recently they last purchased from you (recency), how often they purchased (frequency), and how much they spent over all transactions (monetary).

The reasoning behind RFM analysis is that customers who purchase a product or service once are more likely to purchase again. The categorized customer data is separated into a number of bins, with the binning criteria adjusted as you require. In each of the bins, customers are assigned a score; these scores are then combined to provide an overall RFM score. This score is a representation of the customer's membership in the bins created for each of the RFM parameters. This binned data may be sufficient for your needs, for example, by identifying the most frequent, high-value customers; alternatively, it can be passed on in a stream for further modeling and analysis.

Note, however, that although the ability to analyze and rank RFM scores is a useful tool, you must be aware of certain factors when using it. There may be a temptation to target customers with the highest rankings; however, over-solicitation of these customers could lead to resentment and an actual fall in repeat business. It is also worth remembering that customers with low scores should not be neglected but instead may be cultivated to become better customers. Conversely, high scores alone do not necessarily reflect a good sales prospect, depending on the market. For example, a customer in bin 5 for recency, meaning that they have purchased very recently, may not actually be the best target customer for someone selling expensive, longer-life products such as cars or televisions.

Note: Depending on how your data are stored, you may need to precede the RFM Analysis node with an RFM Aggregate node to transform the data into a usable format. For example, input data must be in customer format, with one row per customer; if the customers' data are in transactional form, use an RFM Aggregate node upstream to derive the recency, frequency, and monetary fields. See the topic "RFM Aggregate Node" on page 73 for more information.

The RFM Aggregate and RFM Analysis nodes in IBM SPSS Modeler are set up to use independent binning; that is, they rank and bin data on each measure of recency, frequency, and monetary value, without regard to their values or the other two measures.

RFM Analysis Node Settings

Recency. Using the Field Chooser (button to the right of the text box), select the recency field. This may be a date, timestamp, or simple number. Note that when a date or timestamp represents the date of the most recent transaction, the highest value is considered the most recent; where a number is specified, it represents the time elapsed since the most recent transaction and the lowest value is considered as the most recent.

Note: If the RFM Analysis node is preceded in the stream by an RFM Aggregate node, the Recency, Frequency, and Monetary fields generated by the RFM Aggregate node should be selected as inputs in the RFM Analysis node.

Frequency. Using the Field Chooser, select the frequency field to be used.

Monetary. Using the Field Chooser, select the monetary field to be used.

Number of bins. For each of the three output types, select the number of bins to be created. The default is 5.

Note: The minimum number of bins is 2, and the maximum is 9.

Weight. By default, the highest importance when calculating scores is given to the recency data, followed by frequency, and then monetary. If required, you can amend the weighting affecting one or several of these to change which is given the highest importance.

The RFM score is calculated as follows: (Recency score x Recency weight) + (Frequency score x Frequency weight) + (Monetary score x Monetary weight).

Ties. Specify how identical (tied) scores are to be binned. The options are:

- **Add to next.** Select to move the tie values up to the next bin.
- **Keep in current.** Keeps tie values in the current (lower) bin. This method may result in fewer total bins being created. (This is the default value.)

Bin thresholds. Specify whether the RFM scores and bin allocations are always recomputed when the node is executed or that they are computed only as necessary (for example, when new data has been added). If you select **Read from Bin Values tab if available** you can edit the upper and lower cut points for the different bins on the Bin Values tab.

When executed, the RFM Analysis node bins the raw recency, frequency, and monetary fields and adds the following new fields to the dataset:

- Recency score. A rank (bin value) for Recency
- Frequency score. A rank (bin value) for Frequency
- Monetary score. A rank (bin value) for Monetary
- RFM score. The weighted sum of the recency, frequency, and monetary scores.

Add outliers to end bins. If you select this check box, records that lie below the lower bin are added to the lower bin, and records above the highest bin are added to the highest bin—otherwise, they are given a null value. This box is available only if you select **Read from Bin Values tab if available**.

RFM Analysis Node Binning

The Bin Values tab allows you to view, and in certain cases amend, the thresholds for generated bins.

Note: You can only amend values on this tab if you select **Read from Bin Values tab if available** on the Settings tab.

Binned field. Use the drop-down list to select a field for dividing into bins. The available values are those selected on the Settings tab.

Bin values table. Threshold values are shown here for each generated bin. If you select **Read from Bin Values tab if available** on the Settings tab, you can amend the upper and lower cut points for each bin by double-clicking on the relevant cell.

Read Values. Reads binned values from the dataset and populates the bin values table. Note that if you select **Always recompute** on the Settings tab, the bin thresholds will be overwritten when new data are run through the stream.

Ensemble Node

The Ensemble node combines two or more model nuggets to obtain more accurate predictions than can be gained from any of the individual models. By combining predictions from multiple models, limitations in individual models may be avoided, resulting in a higher overall accuracy. Models combined in this manner typically perform at least as well as the best of the individual models and often better.

This combining of nodes happens automatically in the Auto Classifier, Auto Numeric and Auto Cluster automated modeling nodes.

After using an Ensemble node, you can use an Analysis node or Evaluation node to compare the accuracy of the combined results with each of the input models. To do this, make sure the **Filter out fields generated by ensembled models** option is not selected on the Settings tab in the Ensemble node.

Output Fields

Each Ensemble node generates a field containing the combined scores. The name is based on the specified target field and prefixed with *$XF_*, *$XS_*, or *$XR_*, depending on the field measurement level--flag, nominal (set), or continuous (range), respectively. For example, if the target is a flag field named *response*, the output field would be *$XF_response*.

Confidence or propensity fields. For flag and nominal fields, additional confidence or propensity fields are created based on the ensemble method, as detailed in the following table.

Table 29. Ensemble method field creation

Ensemble method	Field name
Voting Confidence-weighted voting Raw-propensity-weighted voting Adjusted-propensity-weighted voting Highest confidence wins	*$XFC_<field>*
Average raw propensity	*$XFRP_<field>*
Average adjusted raw propensity	*$XFAP_<field>*

Ensemble Node Settings

Target field for ensemble. Select a single field that is used as the target by two or more upstream models. The upstream models can use flag, nominal, or continuous targets, but at least two of the models must share the same target in order to combine scores.

Filter out fields generated by ensembled models. Removes from the output all of the additional fields generated by the individual models that feed into the Ensemble node. Select this check box if you are interested only in the combined score from all of the input models. Ensure that this option is deselected if, for example, you want to use an Analysis node or Evaluation node to compare the accuracy of the combined score with that of each of the individual input models.

Available settings depend on the measurement level of the field selected as the target.

Continuous Targets

For a continuous target, scores will be averaged. This is the only available method for combining scores.

When averaging scores or estimates, the Ensemble node uses a standard error calculation to work out the difference between the measured or estimated values and the true values, and to show how close those estimates matched. Standard error calculations are generated by default for new models; however, you can deselect the check box for existing models, for example, if they are to be regenerated.

Categorical Targets

For categorical targets, a number of methods are supported, including **voting**, which works by tallying the number of times each possible predicted value is chosen and selecting the value with the highest total. For example, if three out of five models predict *yes* and the other two predict *no*, then *yes* wins by a vote of 3 to 2. Alternatively, votes can be **weighted** based on the confidence or propensity value for each prediction. The weights are then summed, and the value with the highest total is again selected. The confidence for the final prediction is the sum of the weights for the winning value divided by the number of models included in the ensemble.

All categorical fields. For both flag and nominal fields, the following methods are supported:

- Voting
- Confidence-weighted voting
- Highest confidence wins

Flag fields only. For flag fields only, a number of methods based on propensity are also available:

- Raw propensity-weighted voting
- Adjusted propensity-weighted voting
- Average raw propensity
- Average adjusted propensity

Voting ties. For voting methods, you can specify how ties are resolved.

- **Random selection.** One of the tied values is chosen at random.
- **Highest confidence.** The tied value that was predicted with the highest confidence wins. Note that this is not necessarily the same as the highest confidence of all predicted values.
- **Raw or adjusted propensity (flag fields only).** The tied value that was predicted with the largest absolute propensity, where the absolute propensity is calculated as:

```
abs(0.5 - propensity) *
2
```

Or, in the case of adjusted propensity:

```
abs(0.5 - adjusted propensity) * 2
```

Partition Node

Partition nodes are used to generate a partition field that splits the data into separate subsets or samples for the training, testing, and validation stages of model building. By using one sample to generate the

model and a separate sample to test it, you can get a good indication of how well the model will generalize to larger datasets that are similar to the current data.

The Partition node generates a nominal field with the role set to **Partition**. Alternatively, if an appropriate field already exists in your data, it can be designated as a partition using a Type node. In this case, no separate Partition node is required. Any instantiated nominal field with two or three values can be used as a partition, but flag fields cannot be used. See the topic "Setting the field role" on page 131 for more information.

Multiple partition fields can be defined in a stream, but if so, a single partition field must be selected on the Fields tab in each modeling node that uses partitioning. (If only one partition is present, it is automatically used whenever partitioning is enabled.)

Enabling partitioning. To use the partition in an analysis, partitioning must be enabled on the Model Options tab in the appropriate model-building or analysis node. Deselecting this option makes it possible to disable partitioning without removing the field.

To create a partition field based on some other criterion such as a date range or location, you can also use a Derive node. See the topic "Derive node" on page 136 for more information.

Example. When building an RFM stream to identify recent customers who have positively responded to previous marketing campaigns, the marketing department of a sales company uses a Partition node to split the data into training and test partitions.

Partition Node Options

Partition field. Specifies the name of the field created by the node.

Partitions. You can partition the data into two samples (train and test) or three (train, test, and validation).

- **Train and test.** Partitions the data into two samples, allowing you to train the model with one sample and test with another.
- **Train, test, and validation.** Partitions the data into three samples, allowing you to train the model with one sample, test and refine the model using a second sample, and validate your results with a third. This reduces the size of each partition accordingly, however, and may be most suitable when working with a very large dataset.

Partition size. Specifies the relative size of each partition. If the sum of the partition sizes is less than 100%, then the records not included in a partition will be discarded. For example, if a user has 10 million records and has specified partition sizes of 5% training and 10% testing, after running the node, there should be roughly 500,000 training and one million testing records, with the remainder having been discarded.

Values. Specifies the values used to represent each partition sample in the data.

- **Use system-defined values ("1," "2," and "3").** Uses an integer to represent each partition; for example, all records that fall into the training sample have a value of 1 for the partition field. This ensures the data will be portable between locales and that if the partition field is reinstantiated elsewhere (for example, reading the data back from a database), the sort order is preserved (so that 1 will still represent the training partition). However, the values do require some interpretation.
- **Append labels to system-defined values.** Combines the integer with a label; for example, training partition records have a value of `1_Training`. This makes it possible for someone looking at the data to identify which value is which, and it preserves sort order. However, values are specific to a given locale.
- **Use labels as values.** Uses the label with no integer; for example, **Training**. This allows you to specify the values by editing the labels. However, it makes the data locale-specific, and reinstantiation of a partition column will put the values in their natural sort order, which may not correspond to their "semantic" order.

Seed. Only available when **Repeatable partition assignment** is selected. When sampling or partitioning records based on a random percentage, this option allows you to duplicate the same results in another

session. By specifying the starting value used by the random number generator, you can ensure the same records are assigned each time the node is executed. Enter the desired seed value, or click the **Generate** button to automatically generate a random value. If this option is not selected, a different sample will be generated each time the node is executed.

Note: When using the **Seed** option with records read from a database, a Sort node may be required prior to sampling in order to ensure the same result each time the node is executed. This is because the random seed depends on the order of records, which is not guaranteed to stay the same in a relational database. See the topic "Sort Node" on page 74 for more information.

Use unique field to assign partitions. Only available when **Repeatable partition assignment** is selected. (For Tier 1 databases only) Check this box to use SQL pushback to assign records to partitions. From the drop-down list, choose a field with unique values (such as an ID field) to ensure that records are assigned in a random but repeatable way.

Database tiers are explained in the description of the Database source node. See the topic "Database source node" on page 14 for more information.

Generating select nodes

Using the Generate menu in the Partition node, you can automatically generate a Select node for each partition. For example, you could select all records in the training partition to obtain further evaluation or analyses using only this partition.

Set to Flag Node

The Set to Flag node is used to derive flag fields based on the categorical values defined for one or more nominal fields. For example, your dataset might contain a nominal field, *BP* (blood pressure), with the values *High*, *Normal*, and *Low*. For easier data manipulation, you might create a flag field for high blood pressure, which indicates whether or not the patient has high blood pressure.

Setting Options for the Set to Flag Node

Set fields. Lists all data fields with a measurement level of *Nominal* (set). Select one from the list to display the values in the set. You can choose from these values to create a flag field. Note that data must be fully instantiated using an upstream source or Type node before you can see the available nominal fields (and their values). See the topic "Type Node" on page 121 for more information.

Field name extension. Select to enable controls for specifying an extension that will be added as a suffix or prefix to the new flag field. By default, new field names are automatically created by combining the original field name with the field value into a label, such as *Fieldname_fieldvalue*.

Available set values. Values in the set selected above are displayed here. Select one or more values for which you want to generate flags. For example, if the values in a field called *blood_pressure* are *High*, *Medium*, and *Low*, you can select *High* and add it to the list on the right. This will create a field with a flag for records with a value indicating high blood pressure.

Create flag fields. The newly created flag fields are listed here. You can specify options for naming the new field using the field name extension controls.

True value. Specify the true value used by the node when setting a flag. By default, this value is **T**.

False value. Specify the false value used by the node when setting a flag. By default, this value is **F**.

Aggregate keys. Select to group records together based on key fields specified below. When **Aggregate keys** is selected, all flag fields in a group will be "turned on" if *any* record was set to true. Use the Field Chooser to specify which key fields will be used to aggregate records.

Restructure Node

The Restructure node can be used to generate multiple fields based on the values of a nominal or flag field. The newly generated fields can contain values from another field or numeric flags (0 and 1). The

functionality of this node is similar to that of the Set to Flag node. However, it offers more flexibility. It allows you to create fields of any type (including numeric flags), using the values from another field. You can then perform aggregation or other manipulations with other nodes downstream. (The Set to Flag node lets you aggregate fields in one step, which may be convenient if you are creating flag fields.)

For example, the following dataset contains a nominal field, *Account*, with the values *Savings* and *Draft*. The opening balance and current balance are recorded for each account, and some customers have multiple accounts of each type. Let's say you want to know whether each customer has a particular account type, and if so, how much money is in each account type. You use the Restructure node to generate a field for each of the *Account* values, and you select *Current_Balance* as the value. Each new field is populated with the current balance for the given record.

Table 30. Sample data before restructuring

CustID	Account	Open_Bal	Current_Bal
12701	Draft	1000	1005.32
12702	Savings	100	144.51
12703	Savings	300	321.20
12703	Savings	150	204.51
12703	Draft	1200	586.32

Table 31. Sample data after restructuring

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Draft	1000	1005.32	1005.32	$null$
12702	Savings	100	144.51	$null$	144.51
12703	Savings	300	321.20	$null$	321.20
12703	Savings	150	204.51	$null$	204.51
12703	Draft	1200	586.32	586.32	$null$

Using the Restructure Node with the Aggregate Node

In many cases, you may want to pair the Restructure node with an Aggregate node. In the previous example, one customer (with the ID 12703) has three accounts. You can use an Aggregate node to calculate the total balance for each account type. The key field is *CustID*, and the aggregate fields are the new restructured fields, *Account_Draft_Current_Bal* and *Account_Savings_Current_Bal*. The following table shows the results.

Table 32. Sample data after restructuring and aggregation

CustID	Record_Count	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	1	1005.32	$null$
12702	1	$null$	144.51
12703	3	586.32	525.71

Setting Options for the Restructure Node

Available fields. Lists all data fields with a measurement level of *Nominal* (set) or *Flag*. Select one from the list to display the values in the set or flag, then choose from these values to create the restructured fields. Note that data must be fully instantiated using an upstream source or Type node before you can see the available fields (and their values). See the topic "Type Node" on page 121 for more information.

Available values. Values in the set selected above are displayed here. Select one or more values for which you want to generate restructured fields. For example, if the values in a field called *Blood Pressure* are *High*, *Medium*, and *Low*, you can select *High* and add it to the list on the right. This will create a field with a specified value (see below) for records with a value of *High*.

Create restructured fields. The newly created restructured fields are listed here. By default, new field names are automatically created by combining the original field name with the field value into a label, such as *Fieldname_fieldvalue*.

Include field name. Deselect to remove the original field name as a prefix from the new field names.

Use values from other fields. Specify one or more fields whose values will be used to populate the restructured fields. Use the Field Chooser to select one or more fields. For each field chosen, one new field is created. The value field name is appended to the restructured field name—for example, *BP_High_Age* or *BP_Low_Age*. Each new field inherits the type of the original value field.

Create numeric value flags. Select to populate the new fields with numeric value flags (0 for false and 1 for true), rather than using a value from another field.

Transpose Node

By default, columns are fields and rows are records or observations. If necessary, you can use a Transpose node to swap the data in rows and columns so that fields become records and records become fields. For example, if you have time series data where each series is a row rather than a column, you can transpose the data prior to analysis.

Setting options for the Transpose node

In the **Transpose method** drop-down, select the method you want the Transpose node to perform: **Both fields and records**, **Records to fields**, or **Fields to records**. The settings for each of the three methods are described in the following sections.

Restriction: The **Records to fields** and **Fields to records** methods are only supported on Windows 64-bit, Linux 64-bit, and Mac.

Both fields and records

New field names can be generated automatically based on a specified prefix, or read from an existing field in the data.

Use prefix. This option generates new field names automatically based on the specified prefix (`Field1`, `Field2`, and so on). You can customize the prefix as needed. With this option, you must specify the number of fields to be created, regardless of the number of rows in the original data. For example, if **Number of new fields** is set to 100, all data beyond the first 100 rows will be discarded. If there are fewer than 100 rows in the original data, some fields will be null. (You can increase the number of fields as needed, but the purpose of this setting is to avoid transposing a million records into a million fields, which would produce an unmanageable result.)

For example, suppose you have data with series in rows and a separate field (column) for each month. You can transpose this so that each series is in a separate field, with a row for each month.

Read from field. Reads field names from an existing field. With this option, the number of new fields is determined by the data, up to the specified maximum. Each value of the selected field becomes a new field in the output data. The selected field can have any storage type (integer, string, date, and so on), but in order to avoid duplicate field names, each value of the selected field must be unique (in other words,

the number of values should match the number of rows). If duplicate field names are encountered, a warning is displayed.

- **Read Values.** If the selected field has not been instantiated, select this option to populate the list of new field names. If the field has already been instantiated, then this step is not necessary.
- **Maximum number of values to read.** When reading fields names from the data, an upper limit is specified in order to avoid creating an inordinately large number of fields. (As noted above, transposing one million records into one million fields would produce an unmanageable result.)

For example, if the first column in your data specifies the name for each series, you can use these values as fields names in the transposed data.

Transpose. By default, only continuous (numeric range) fields are transposed (either integer or real storage). Optionally, you can choose a subset of numeric fields or transpose string fields instead. However, all transposed fields must be of the same storage type—either numeric or string but not both—since mixing the input fields would generate mixed values within each output column, which violates the rule that all values of a field must have the same storage. Other storage types (date, time, timestamp) cannot be transposed.

- **All numeric.** Transposes all numeric fields (integer or real storage). The number of rows in the output matches the number of numeric fields in the original data.
- **All string.** Transposes all string fields.
- **Custom.** Allows you to select a subset of numeric fields. The number of rows in the output matches the number of fields selected. This option is only available for numeric fields.

Row ID name. Specifies the name of the row ID field created by the node. The values of this field are determined by the names of the fields in the original data.

Tip: When transposing time series data from rows to columns, if your original data includes a row, such as date, month, or year, that labels the period for each measurement, be sure to read these labels into IBM SPSS Modeler as field names (as demonstrated in the above examples, which show the month or date as field names in the original data, respectively) rather than including the label in the first row of data. This will avoid mixing labels and values in each column (which would force numbers to be read as strings, since storage types cannot be mixed within a column).

Records to fields

Fields. The Fields list contains all fields entering the Transpose node.

Index. Use the Index section to select the fields you want to use as index fields.

Fields. Use the Fields section to select the fields you want to use as fields.

Value. Use the Value section to select the fields you want to use as value fields.

Aggregate function. When there are more than one records for an index, you must aggregate the records into one. Use the **Aggregate function** drop-down to specify how to aggregate the records using one of the following functions. Note that aggregation impacts all fields.

- **Mean.** Returns the mean values for each key field combination. The mean is a measure of central tendency, and is the arithmetic average (the sum divided by the number of cases).
- **Sum.** Returns summed values for each key field combination. The sum is the total of the values, across all cases with nonmissing values.
- **Min.** Returns minimum values for each key field combination.
- **Max.** Returns maximum values for each key field combination.
- **Median.** Returns the median values for each key field combination. The median is a measure of central tendency that is not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values). Also known as the 50th percentile or 2nd quartile.
- **Count.** Returns the count of non-null values for each key field combination.

Fields to records

Fields. The Fields list contains all fields entering the Transpose node.

Index. Use the Index section to select the fields you want to use as index fields.

Value. Use the Value section to select the fields you want to use as value fields. If you don't select any value fields, then all unassigned numeric fields will be used as values. But if not numeric fields are available, then all unassigned string fields will be used.

History Node

History nodes are most often used for sequential data, such as time series data. They are used to create new fields containing data from fields in previous records. When using a History node, you may want to have data that is presorted by a particular field. You can use a Sort node to do this.

Setting Options for the History Node

Selected fields. Using the Field Chooser (button to the right of the text box), select the fields for which you want a history. Each selected field is used to create new fields for all records in the dataset.

Offset. Specify the latest record prior to the current record from which you want to extract historical field values. For example, if Offset is set to 3, as each record passes through this node, the field values for the third record previous will be included in the current record. Use the Span settings to specify how far back records will be extracted from. Use the arrows to adjust the offset value.

Span. Specify how many prior records from which you want to extract values. For example, if Offset is set to 3 and Span is set to 5, each record that passes through the node will have five fields added to it for each field specified in the Selected Fields list. This means that when the node is processing record 10, fields will be added from record 7 through record 3. Use the arrows to adjust the span value.

Where history is unavailable. Select one of the following options for handling records that have no history values. This usually refers to the first several records at the top of the dataset, for which there are no previous records to use as a history.

- **Discard records.** Select to discard records where no history value is available for the field selected.
- **Leave history undefined.** Select to keep records where no history value is available. The history field will be filled with an undefined value, displayed as `$null$`.
- **Fill values with.** Specify a value or string to be used for records where no history value is available. The default replacement value is *undef*, the system null. Null values are displayed using the string `$null$`.

When selecting a replacement value, keep in mind the following rules in order for proper execution to occur:

- Selected fields should be of the same storage type.
- If all of the selected fields have numeric storage, the replacement value must be parsed as an integer.
- If all of the selected fields have real storage, the replacement value must be parsed as a real.
- If all of the selected fields have symbolic storage, the replacement value must be parsed as a string.
- If all of the selected fields have date/time storage, the replacement value must be parsed as a date/time field.

If any of the above conditions are not met, you will receive an error when executing the History node.

Field Reorder Node

The Field Reorder node enables you to define the natural order used to display fields downstream. This order affects the display of fields in a variety of places, such as tables, lists, and the Field Chooser. This operation is useful, for example, when working with wide datasets to make fields of interest more visible.

Setting Field Reorder Options

There are two ways to reorder fields: custom ordering and automatic sorting.

Custom Ordering

Select **Custom Order** to enable a table of field names and types where you can view all fields and use arrow buttons to create a custom order.

To reorder fields:

1. Select a field in the table. Use the Ctrl-click method to select multiple fields.
2. Use the simple arrow buttons to move the field(s) up or down one row.
3. Use the line-arrow buttons to move the field(s) to the bottom or top of the list.
4. Specify the order of fields not included here by moving up or down the divider row, indicated as `[other fields]`.

More information on [other fields]

Other fields. The purpose of the **[other fields]** divider row is to break the table into two halves.

- Fields appearing above the divider row will be ordered (as they appear in the table) at the top of all natural orders used to display the fields downstream of this node.
- Fields appearing below the divider row will be ordered (as they appear in the table) at the bottom of all natural orders used to display the fields downstream of this node.

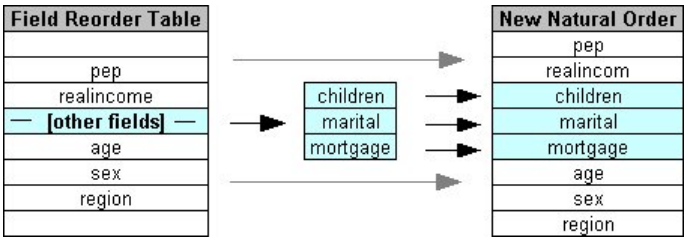


Figure 6. Diagram illustrating how "other fields" are incorporated into the new field order

- All other fields not appearing in the field reorder table will appear between these "top" and "bottom" fields as indicated by the placement of the divider row.

Additional custom sorting options include:

- Sort fields in ascending or descending order by clicking on the arrows above each column header (**Type**, **Name**, and **Storage**). When sorting by column, fields not specified here (indicated by the `[other fields]` row) are sorted last in their natural order.
- Click **Clear Unused** to delete all unused fields from the Field Reorder node. Unused fields are displayed in the table with a red font. This indicates that the field has been deleted in upstream operations.
- Specify ordering for any new fields (displayed with a lightning icon to indicate a new or unspecified field). When you click **OK** or **Apply**, the icon disappears.

Note: If fields are added upstream after a custom order has been applied, the new fields will be appended at the bottom of the custom list.

Automatic Sorting

Select **Automatic Sort** to specify a parameter for sorting. The dialog box options dynamically change to provide options for automatic sorting.

Sort By. Select one of three ways to sort fields read into the Reorder node. The arrow buttons indicate whether the order will be ascending or descending. Select one to make a change.

- Name
- Type
- Storage

Fields added upstream of the Field Reorder node after auto-sort has been applied will automatically be placed in their proper position based on the sort type selected.

Time Intervals Node

The original Time Intervals node, that was available in SPSS Modeler version 17.1 and earlier, was not compatible with Analytic Server (AS) and was deprecated in SPSS Modeler release 18.0.

The replacement Time Intervals node contains a number of changes from the original Time Intervals node. This new node can be used either with Analytic Server or in SPSS Modeler by itself.

You use the Time Intervals node to specify intervals and derive a new time field for estimating or forecasting. A full range of time intervals is supported, from seconds to years.

Use the node to derive a new time field; the new field has the same storage type as the input time field you chose. The node generates the following items:

- The field specified on the Fields tab as the **Time Field**, along with the chosen prefix/suffix. By default the prefix is `$TI_`.
- The fields specified on the Fields tab as the **Dimension fields**.
- The fields specified on the Fields tab as the **Fields to aggregate**.

A number of extra fields can also be generated, depending on the selected interval or period (such as the minute or second within which a measurement falls).

Time Interval - field options

Use the Fields tab in the Time Intervals node to select the data from which the new time interval is derived.

Fields Displays all of the input fields to the node with their measurement type icons. All time fields have the 'continuous' measurement type. Select the field to be used as the input.

Time Field Shows the input field from which the new time interval is derived; only a single continuous field is allowed. The field is used by the Time Intervals node as the aggregation key for converting the interval. The new field has the same storage type as the input time field chosen. If you select an integer field, it is considered to be a time index.

Dimensions fields Optionally, you can add fields here to create an individual time series that is based on the field values. As a simple example, with geospatial data, you can use a point field as a dimension. In this example, the data output from the Time Intervals node is sorted into time series for each point value in the point field.

Dimensions are ideal when you use flattened multi-dimensional data, similar to that generated by the TM1 node, or for supporting more complex data types like geospatial. Essentially, you can consider the use of the **Dimensions fields** as the equivalent of a **`Group By`** clause in a SQL query , or similar to **Key fields** in the Aggregate node; however, the **Dimensions fields** is more sophisticated in nature due to its ability to handle more complicated data structures than just traditional row and column data.

Fields to aggregate Select the fields to be aggregated as part of changing the period of the time field. Only the fields that you select here are available on the Build tab for the **Custom settings for specified fields** table. Any fields not included are filtered out of the data that leaves the node. This means that any fields remaining in the **Fields** list are filtered out of the data.

Time Interval - build options

Use the Build tab to specify options for changing the time interval and how fields in the data are aggregated, based on their measurement type.

When you aggregate data, any existing date, time, or timestamp fields are superseded by the generated fields and are dropped from the output. Other fields are aggregated based on the options you specify in this tab.

Time Interval Select the interval and periodicity for building the series.

Default settings Select the default aggregation to be applied to data of different types. The default is applied based on the measurement level; for example, continuous fields are aggregated by the sum, while nominal fields use the mode. You can set the default for 3 different measurement levels:

- **Continuous** Available functions for continuous fields include **Sum**, **Mean**, **Min**, **Max**, **Median**, **1st Quartile**, and **3rd Quartile**.
- **Nominal** Options include **Mode**, **Min**, and **Max**.
- **Flag** Options are either **True if any true** or **False if any false**.

Custom settings for specified fields You can specify exceptions to the default aggregation settings for individual fields. Use the icons on the right to add or remove fields from the table, or click the cell in the appropriate column to change the aggregation function for that field. Typeless fields are excluded from the list and cannot be added to the table.

New field name extension Specify the **Prefix** or **Suffix** applied to all fields generated by the node.

Reprojection Node

With geospatial or map data, two of the most common ways that are used to identify coordinates are projected coordinate and geographic coordinate systems. Within IBM SPSS Modeler, items such as the Expression Builder spatial functions, the Spatio-Temporal Prediction (STP) Node, and the Map Visualization Node use the projected coordinate system and therefore any data that you import that is recorded with a geographic coordinate system must be reprojected. Where possible, geospatial fields (any fields with a geospatial measurement level) are automatically reprojected when used (not when they are imported). Where any fields cannot be reprojected automatically you use the Reproject node to change their coordinate system. Reprojecting in this way means that you can correct situations where an error occurs due to use of an incorrect coordinate system.

Example situations where you might have to reproject to change the coordinate system are shown in the following list:

- *Append* If you try to append two data sets with different coordinate systems for a geospatial field SPSS Modeler displays the following error message:`Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

  <Field1> and *<Field2>* are the names of the geospatial fields that caused the error.
- *If/else expression* If you use an expression that contains an if/else statement with geospatial fields or return types in both parts of the expression, but with different coordinate systems, SPSS Modeler displays the following error message: `The conditional expression contains incompatible coordinate systems: <arg1> and <arg2>.`

  <arg1> and *<arg2>* are the then or else arguments that return a geospatial type with different coordinate systems.
- *Constructing a list of geospatial fields* To create a list field that consists of numerous geospatial fields, all geospatial field arguments that are supplied to the list expression must be in the same coordinate system. If they are not, then the following error message is displayed: `Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

For more information about coordinate systems, see the Setting Geospatial Options for Streams topic in the Working with Streams section of the SPSS Modeler User's Guide.

Setting options for the Reproject node

Fields

Geo fields
By default this list is empty. You can move geospatial fields into this list from the **Fields to be reprojected** list to ensure that those fields are not reprojected.

Fields to be reprojected
By default this list contains all geospatial fields that are input to this node. All fields in this list are reprojected to the coordinate system that you set in the **Coordinate System** area.

Coordinate System

Stream default
Select this option to use the default coordinate system.

Specify
If you select this option you can use the **Change** button to display the **Select Coordinate System** dialog box and choose the coordinate system to be used for reprojection.

For more information about coordinate systems, see the Setting Geospatial Options for Streams topic in the Working with Streams section of the SPSS Modeler User's Guide.

Chapter 5. Graph Nodes

Common Graph Nodes Features

Several phases of the data mining process use graphs and charts to explore data brought into IBM SPSS Modeler. For example, you can connect a Plot or Distribution node to a data source to gain insight into data types and distributions. You can then perform record and field manipulations to prepare the data for downstream modeling operations. Another common use of graphs is to check the distribution and relationships between newly derived fields.

The Graphs palette contains the following nodes:

The Graphboard node offers many different types of graphs in one single node. Using this node, you can choose the data fields you want to explore and then select a graph from those available for the selected data. The node automatically filters out any graph types that would not work with the field choices.

The Plot node shows the relationship between numeric fields. You can create a plot by using points (a scatterplot) or lines.

The Distribution node shows the occurrence of symbolic (categorical) values, such as mortgage type or gender. Typically, you might use the Distribution node to show imbalances in the data, which you could then rectify using a Balance node before creating a model.

The Histogram node shows the occurrence of values for numeric fields. It is often used to explore the data before manipulations and model building. Similar to the Distribution node, the Histogram node frequently reveals imbalances in the data.

The Collection node shows the distribution of values for one numeric field relative to the values of another. (It creates graphs that are similar to histograms.) It is useful for illustrating a variable or field whose values change over time. Using 3-D graphing, you can also include a symbolic axis displaying distributions by category.

The Multiplot node creates a plot that displays multiple *Y* fields over a single *X* field. The *Y* fields are plotted as colored lines; each is equivalent to a Plot node with Style set to **Line** and X Mode set to **Sort**. Multiplots are useful when you want to explore the fluctuation of several variables over time.

The Web node illustrates the strength of the relationship between values of two or more symbolic (categorical) fields. The graph uses lines of various widths to indicate connection strength. You might use a Web node, for example, to explore the relationship between the purchase of a set of items at an e-commerce site.

The Time Plot node displays one or more sets of time series data. Typically, you would first use a Time Intervals node to create a *TimeLabel* field, which would be used to label the *x* axis.

The Evaluation node helps to evaluate and compare predictive models. The evaluation chart shows how well models predict particular outcomes. It sorts records based on the predicted value and confidence of the prediction. It splits the records into groups of equal size (**quantiles**) and then plots the value of the business criterion for each quantile from highest to lowest. Multiple models are shown as separate lines in the plot.

The Map Visualization node can accept multiple input connections and display geospatial data on a map as a series of layers. Each layer is a single geospatial field; for example, the base layer might be a map of a country, then above that you might have one layer for roads, one layer for rivers, and one layer for towns.

The E-Plot (Beta) node shows the relationship between numeric fields. It is similar to the Plot node, but its options differ and its output uses a new graphing interface specific to this node. Use the beta-level node to play around with new graphing features.

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a tool for visualizing high-dimensional data. It converts affinities of data points to probabilities. This t-SNE node in SPSS Modeler is implemented in Python and requires the `scikit-learn`© Python library.

When you have added a graph node to a stream, you can double-click the node to open a dialog box for specifying options. Most graphs contain a number of unique options presented on one or more tabs. There are also several tab options common to all graphs. The following sections contain more information about these common options.

When you have configured the options for a graph node, you can run it from within the dialog box or as part of a stream. In the generated graph window, you can generate Derive (Set and Flag) and Select nodes based on a selection or region of data, effectively "subsetting" the data. For example, you might use this powerful feature to identify and exclude outliers.

Aesthetics, Overlays, Panels, and Animation

Overlays and Aesthetics

Aesthetics (and overlays) add dimensionality to a visualization. The effect of an aesthetic (grouping, clustering, or stacking) depends on the visualization type, the type of field (variable), and the graphic element type and statistic. For example, a categorical field for color may be used to group points in a scatterplot or to create the stacks in a stacked bar chart. A continuous numeric range for color may also be used to indicate the range's values for each point in a scatterplot.

You should experiment with the aesthetics and overlays to find one that fulfills your needs. The following descriptions may help you pick the right one.

Note: Not all aesthetics or overlays are available for all visualization types.

- **Color.** When color is defined by a categorical field, it splits the visualization based on the individual categories, one color for each category. When color is a continuous numeric range, it varies the color based on the value of the range field. If the graphic element (for example, a bar or box) represents more than one record/case and a range field is used for color, the color varies based on the *mean* of the range field.
- **Shape.** Shape is defined by a categorical field that splits the visualization into elements of different shapes, one for each category.
- **Transparency.** When transparency is defined by a categorical field, it splits the visualization based on the individual categories, one transparency level for each category. When transparency is a continuous

numeric range, it varies the transparency based on the value of the range field. If the graphic element (for example, a bar or box) represents more than one record/case and a range field is used for transparency, the color varies based on the *mean* of the range field. At the largest value, the graphic elements are fully transparent. At the smallest value, they are fully opaque.

- **Data Label.** Data labels are defined by any type of field whose values are used to create labels that are attached to the graphic elements.
- **Size.** When size is defined by a categorical field, it splits the visualization based on the individual categories, one size for each category. When size is a continuous numeric range, it varies the size based on the value of the range field. If the graphic element (for example, a bar or box) represents more than one record/case and a range field is used for size, the size varies based on the *mean* of the range field.

Paneling and Animation

Paneling. Paneling, also known as faceting, creates a table of graphs. One graph is generated for each category in the paneling fields, but all panels appear simultaneously. Paneling is useful for checking whether the visualization is subject to the conditions of the paneling fields. For example, you may panel a histogram by gender to determine whether the frequency distributions are equal across males and females. That is, you can check whether salary is subject to gender differences. Select a categorical field for paneling.

Animation. Animation resembles paneling in that multiple graphs are created from the values of the animation field, but these graphs are not shown together. Rather, you use the controls in Explore mode to animate the output and flip through a sequence of individual graphs. Furthermore, unlike paneling, animation does not require a categorical field. You can specify a continuous field whose values are split up into ranges automatically. You can vary the size of the range with the animation controls in explore mode. Not all visualizations offer animation.

Using the Output Tab

For all graph types, you can specify the following options for the filename and display of generated graphs.

Note: Distribution node graphs have additional settings.

Output name. Specifies the name of the graph produced when the node is run. **Auto** chooses a name based on the node that generates the output. Optionally, you can select **Custom** to specify a different name.

Output to screen. Select to generate and display the graph in a new window.

Output to file. Select to save the output as a file.

- **Output Graph.** Select to produce output in a graph format. Available only in Distribution nodes.
- **Output Table.** Select to produce output in a table format. Available only in Distribution nodes.
- **Filename.** Specify a filename used for the generated graph or table. Use the ellipsis button **(...)** to specify a specific file and location.
- **File type.** Specify the file type in the drop-down list. For all graph nodes, except the Distribution node with an **Output Table** option, the available graph file types are as follows.

  - Bitmap (.bmp)

  - PNG (.png)

  - Output object (.cou)

  - JPEG (.jpg)

  - HTML (.html)

  - ViZml document (.xml) for use in other IBM SPSS Statistics applications.

  For the **Output Table** option in the Distribution node, the available file types are as follows.

  - Tab delimited data (.tab)

- Comma delimited data (.csv)
- HTML (.html)
- Output object (.cou)

Paginate output. When saving output as HTML, this option is enabled to enable you to control the size of each HTML page. (Applies only to the Distribution node.)

Lines per page. When **Paginate output** is selected, this option is enabled to enable you to determine the length of each HTML page. The default setting is 400 rows. (Applies only to the Distribution node.)

Using the Annotations Tab

Used for all nodes, this tab offers options to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

3-D Graphs

Plots and collection graphs in IBM SPSS Modeler have the ability to display information on a third axis. This provides you with additional flexibility in visualizing your data to select subsets or deriving new fields for modeling.

Once you have created a 3-D graph, you can click on it and drag your mouse to rotate it and view it from any angle.

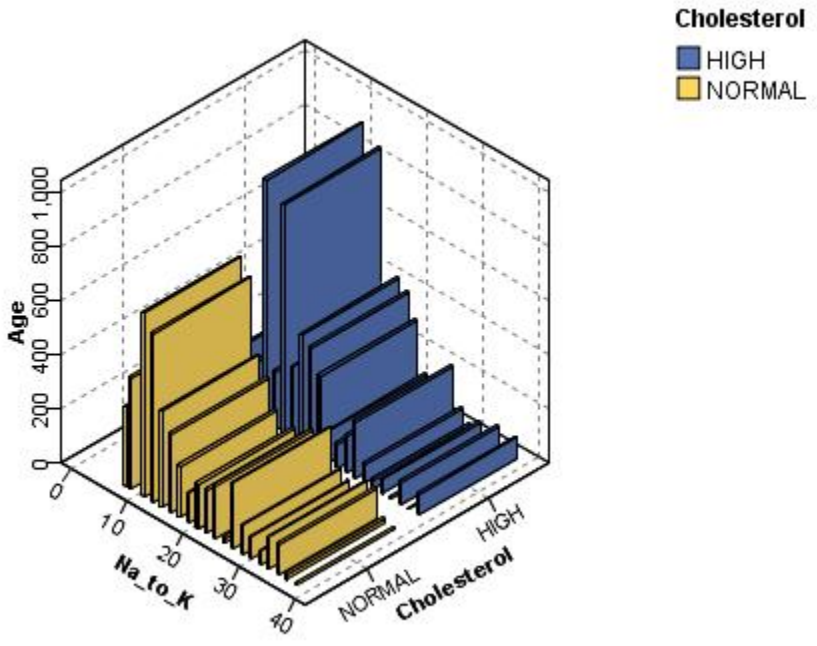


Figure 7. Collection graph with x, y, and z axes

There are two ways of creating 3-D graphs in IBM SPSS Modeler: plotting information on a third axis (true 3-D graphs) and displaying graphs with 3-D effects. Both methods are available for plots and collections.

To Plot Information on a Third Axis

1. In the graph node dialog box, click the **Plot** tab.
2. Click the 3-D button to enable options for the *z* axis.

3. Use the Field Chooser button to select a field for the *z* axis. In some cases, only symbolic fields are allowed here. The Field Chooser will display the appropriate fields.

To Add 3-D Effects to a Graph

1. Once you have created a graph, click the **Graph** tab in the output window.
2. Click the 3-D button to switch the view to a three-dimensional graph.

Graphboard node

The Graphboard node enables you to choose from many different graphs outputs (bar charts, pie charts, histograms, scatterplots, heatmaps, etc.) in one single node. You begin, in the first tab, by choosing the data fields you want to explore, and then the node presents you with a choice of graph types that work for your data. The node automatically filters out any graph types that would not work with the field choices. You can define detailed, or more advanced graph options in the Detailed tab.

Note: You must connect the Graphboard node to a stream with data in order to edit the node or select graph types.

There are two buttons that enable you to control which visualization templates (and stylesheets, and maps) are available:

Manage. Manage visualization templates, stylesheets, and maps on your computer. You can import, export, rename, and delete visualization templates, stylesheets, and maps on your local machine. See the topic “Managing Templates, Stylesheets, and Map Files” on page 195 for more information.

Location. Change the location in which visualization templates, stylesheets, and maps are stored. The current location is listed to the right of the button. See the topic “Setting the Location of Templates, Stylesheets, and Maps” on page 195 for more information.

Graphboard Basic Tab

If you aren't sure which visualization type would best represent your data, use the Basic tab. When you select your data, you are then presented with a subset of visualization types that are appropriate for the data. For examples, see “ Graphboard Examples” on page 184 .

1. Select one or more fields (variables) from the list. Use Ctrl+Click to select multiple fields.

   Note that the measurement level of the field determines the type of visualizations that are available. You can change the measurement level by right-clicking the field in the list and choosing an option. For more information about the available measurement level types, see “Field (Variable) Types” on page 173 .
2. Select a visualization type. For descriptions of the available types, see “Available Built-in Graphboard Visualization Types” on page 176 .
3. For certain visualizations, you can choose a summary statistic. Different subsets of statistics are available depending on whether the statistic is count-based or calculated from a continuous field. The available statistics also depend on the template itself. A full list of statistics that may be available follows the next step.
4. If you want to define more options, such as optional aesthetics and panel fields, click **Detailed**. See the topic “ Graphboard Detailed Tab” on page 174 for more information.

Summary Statistics Calculated from a Continuous Field

- *Mean*. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- *Median*. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).
- *Mode*. The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.

- *Minimum*. The smallest value of a numeric variable.
- *Maximum*. The largest value of a numeric variable.
- *Range*. The difference between the minimum and maximum values.
- *Mid Range*. The middle of the range, that is, the value whose difference from the minimum is equal to its difference from the maximum.
- *Sum*. The sum or total of the values, across all cases with nonmissing values.
- *Cumulative Sum*. The cumulative sum of the values. Each graphic element shows the sum for one subgroup plus the total sum of all previous groups.
- *Percent Sum*. The percentage within each subgroup based on a summed field compared to the sum across all groups.
- *Cumulative Percent Sum*. The cumulative percentage within each subgroup based on a summed field compared to the sum across all groups. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.
- *Variance*. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- *Standard Deviation*. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- *Standard Error*. A measure of how much the value of a test statistic varies from sample to sample. It is the standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample means.
- *Kurtosis*. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution.
- *Skewness*. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

The following region statistics may result in more than one graphic element per subgroup. When using the interval, area, or edge graphic elements, a region statistic results in one graphic element showing the range. All other graphic elements result in two separate elements, one showing the start of the range and one showing the end of the range.

- **Region: Range**. The range of values between the minimum and maximum values.
- **Region: 95% Confidence Interval of Mean**. A range of values that has a 95% chance of including the population mean.
- **Region: 95% Confidence Interval of Individual**. A range of values that has a 95% chance of including the predicted value given the individual case.
- **Region: 1 Standard Deviation above/below Mean**. A range of values between 1 *standard deviation* above and below the *mean*.
- **Region: 1 Standard Error above/below Mean**. A range of values between 1 *standard error* above and below the *mean*.

Count-Based Summary Statistics

- **Count**. The number of rows/cases.
- **Cumulative Count**. The cumulative number of rows/cases. Each graphic element shows the count for one subgroup plus the total count of all previous groups.
- **Percent of Count**. The percentage of rows/cases in each subgroup compared to the total number of rows/cases.

- **Cumulative Percent of Count**. The cumulative percentage of rows/cases in each subgroup compared to the total number of rows/cases. Each graphic element shows the percentage for one subgroup plus the total percentage of all previous groups.

Field (Variable) Types

Icons appear next to fields in field lists to indicate the field type and data type. Icons also identify multiple response sets.

Table 33. Measurement level icons

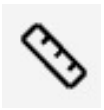
Measurement Level	Numeric	String	Date	Time
Continuous		n/a		
Ordered Set				
Set				

Table 34. Multiple response set icons

Multiple response set type	Icon
Multiple response set, multiple categories	
Multiple response set, multiple dichotomies	

Measurement Level

A field's measurement level is important when you create a visualization. Following is a description of the measurement levels. You can temporarily change the measurement level by right-clicking a field in a field list and choosing an option. In most cases, you need to consider only the two broadest classifications of fields, categorical and continuous:

Categorical. Data with a limited number of distinct values or categories (for example, gender or religion). Categorical fields can be string (alphanumeric) or numeric fields that use numeric codes to represent categories (for example, 0 = *male* and 1 = *female*). Also referred to as qualitative data. Sets, ordered sets, and flags are all categorical fields.

- *Set*. A field/variable whose values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation. Also known as a nominal variable.
- *Ordered Set*. A field/variable whose values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordered sets include attitude scores representing degree of satisfaction or confidence and preference rating scores. Also known as an ordinal variable.
- *Flag*. A field/variable with two distinct values, such as Yes and No or 1 and 2. Also known as a dichotomous or binary variable.

Continuous. Data measured on an interval or ratio scale, where the data values indicate both the order of values and the distance between values. For example, a salary of $72,195 is higher than a salary of

$52,398, and the distance between the two values is $19,797. Also referred to as quantitative, scale, or numeric range data.

Categorical fields define categories in the visualization, typically to draw separate graphic elements or to group graphic elements. Continuous fields are often summarized within categories of categorical fields. For example, a default visualization of income for gender categories would display the mean income for males and the mean income for females. The raw values for continuous fields can also be plotted, as in a scatterplot. For example, a scatterplot may show the current salary and beginning salary for each case. A categorical field could be used to group the cases by gender.

Data Types

Measurement level isn't the only property of a field that determines its type. A field is also stored as a specific data type. Possible data types are strings (non-numeric data such as letters), numeric values (real numbers), and dates. Unlike the measurement level, a field's data type cannot be changed temporarily. You must change the way the data are stored in the original data set.

Multiple Response Sets

Some data files support a special kind of "field" called a **multiple response set**. Multiple response sets aren't really "fields" in the normal sense. Multiple response sets use multiple fields to record responses to questions where the respondent can give more than one answer. Multiple response sets are treated like categorical fields, and most of the things you can do with categorical fields, you can also do with multiple response sets.

Multiple response sets can be multiple dichotomy sets or multiple category sets.

Multiple dichotomy sets. A multiple dichotomy set typically consists of multiple dichotomous fields: fields with only two possible values of a yes/no, present/absent, checked/not checked nature. Although the fields may not be strictly dichotomous, all of the fields in the set are coded the same way.

For example, a survey provides five possible responses to the question, "Which of the following sources do you rely on for news?" The respondent can indicate multiple choices by checking a box next to each choice. The five responses become five fields in the data file, coded 0 for *No* (not checked) and 1 for *Yes* (checked).

Multiple category sets. A multiple category set consists of multiple fields, all coded the same way, often with many possible response categories. For example, a survey item states, "Name up to three nationalities that best describe your ethnic heritage." There may be hundreds of possible responses, but for coding purposes the list is limited to the 40 most common nationalities, with everything else relegated to an "other" category. In the data file, the three choices become three fields, each with 41 categories (40 coded nationalities and one "other" category).

Graphboard Detailed Tab

Use the Detailed tab when you know what type of visualization you want to create or when you want to add optional aesthetics, panels, and/or animation to a visualization. For examples, see “ Graphboard Examples” on page 184 .

1. If you selected a visualization type on the Basic tab, it will be displayed. Otherwise, choose one from the drop-down list. For information about the visualization types, see “Available Built-in Graphboard Visualization Types” on page 176 .
2. To the immediate right of the visualization's thumbnail image are controls for specifying the fields (variables) required for the visualization type. You must specify all of these fields.
3. For certain visualizations, you can select a summary statistic. In some cases (such as with bar charts), you can use one of these summary options for the transparency aesthetic. For descriptions of the summary statistics, see “ Graphboard Basic Tab” on page 171 .
4. You can select one or more of the optional aesthetics. These can add dimensionality by allowing you to include other fields in the visualization. For example, you may use a field to vary the size of points in a scatterplot. For more information about optional aesthetics, see “Aesthetics, Overlays, Panels, and Animation” on page 168 . Please note that the transparency aesthetic is not supported through scripting.

5. If you are creating a map visualization, the **Map Files** group shows the map file or files that will be used. If there is a default map file, this file is displayed. To change the map file, click **Select a Map File** to display the Select Maps dialog box. You can also specify the default map file in this dialog box. See the topic "Selecting map files for map visualizations" on page 175 for more information.
6. You can select one or more of the paneling or animation options. For more information about paneling and animation options, see "Aesthetics, Overlays, Panels, and Animation" on page 168 .

Selecting map files for map visualizations

If you select a map visualization template, you need a map file that defines the geographic information for drawing the map. If there is a default map file, this will be used for the map visualization. To choose a different map file, click **Select a Map File** on the Detailed tab to display the Select Maps dialog box.

The Select Maps dialog box allows you to choose a primary map file and a reference map file. The map files define the geographic information for drawing the map. Your application is installed with a set of standard map files. If you have other ESRI shapefiles that you want to use, you first need to convert the shapefiles to SMZ files. See the topic "Converting and Distributing Map Shapefiles" on page 196 for more information. After converting the map, click **Manage...** on the Template Chooser dialog box to import the map into the Manage system so that it is available in the Select Maps dialog box.

Following are some points to consider when specifying map files:

- All map templates require at least one map file.
- The map file typically links a map key attribute to the data key.
- If the template does not require a map key that links to a data key, it requires a reference map file and fields that specify coordinates (such as longitude and latitude) for drawing elements on the reference map.
- Overlay map templates require two maps: a primary map file and a reference map file. The reference map is drawn first so that it is behind the primary map file.

For information about map terminology such as attributes and features, see "Key Concepts for Maps" on page 197.

Map File. You can select any map file that is in the Manage system. These include pre-installed map files and map files that you imported. For more information about managing map files, see "Managing Templates, Stylesheets, and Map Files" on page 195.

Map Key. Specify the attribute that you want to use as the key that links the map file to the data key.

Save this map file and settings as the default. Select this checkbox if you want to use the selected map file as the default. If you have specified a default map file, you don't need to specify a map file every time you create a map visualization.

Data Key. This control lists the same value as appears on the Template Chooser Detailed tab. It is provided here for convenience if you need to change the key because of the specific map file that you choose.

Display all map features in the visualization. When this option is checked, all features in the map are rendered in the visualization, even if there is no matching data key value. If you want to see only the features for which you have data, clear this option. Features identified by the map keys shown in the **Unmatched Map Keys** list will not be rendered in the visualization.

Compare Map and Data Values. The map key and data key link to each other to create the map visualization. The map key and the data key should draw from the same domain (for example, countries and regions). Click **Compare** to test whether the data key and map key values match. The displayed icon informs you of the state of the comparison. These icons are described below. If a comparison has been performed and there are data key values with no matching map key values, the data key values appear in the **Unmatched Data Keys** list. In the **Unmatched Map Keys** list, you can also see which map key values have no matching data key values. If **Display all map features in the visualization** is not checked, features identified by these map key values will not be rendered.

Table 35. Comparison icons	
Icon	**Description**
	No comparison has been performed. This is the default state before you click **Compare**. You should proceed with caution because you don't know if the data key and map key values match.
	A comparison has been performed, and the data key and map key values match completely. For every value of the data key, there is a matching feature identified by the map key.
	A comparison has been performed, and some data key and map key values do not match. For some data key values, there is no matching feature identified by the map key. You should proceed with caution. If you proceed, the map visualization will not include all data values.
	A comparison has been performed, and no data key and map key values match. You should choose a different data key or map key because no map will be rendered if you proceed.

Available Built-in Graphboard Visualization Types

You can create several different visualization types. All of the following built-in types are available on both the Basic and Detailed tabs. Some of the descriptions for the templates (especially the map templates) identify the fields (variables) specified on the Detailed tab using **special text**.

Table 36. Available graph types

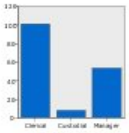
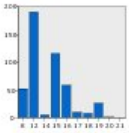
Chart icon	Description	Chart icon	Description
	**Bar** Calculates a summary statistic for a continuous numeric field and displays the results for each category of a categorical field as bars. *Requires*: A categorical field and a continuous field.		**Bar of Counts** Displays the proportion of rows/cases in each category of a categorical field as bars. You can also use the Distribution graph node to produce this graph. That node offers some additional options. See the topic "Distribution Node" on page 213 for more information. *Requires*: A single categorical field.

Table 36. Available graph types (continued)

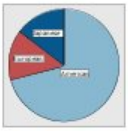
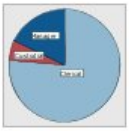
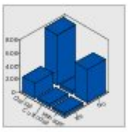
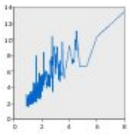
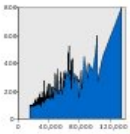
Chart icon	Description	Chart icon	Description
	**Pie** Calculates the sum of a continuous numeric field and displays the proportion of that sum distributed in each category of a categorical field as slices of a pie. *Requires*: A categorical field and a continuous field.		**Pie of Counts** Displays the proportion of rows/cases in each category of a categorical field as slices of a pie. *Requires*: A single categorical field.
	**3-D Bar** Calculates a summary statistic for a continuous numeric field and displays the results for the crossing of categories between two categorical fields. *Requires*: A pair of categorical fields and a continuous field.		**3-D Pie** This is the same as Pie except with an additional 3-D effect. *Requires*: A categorical field and a continuous field.
	**Line** Calculates a summary statistic for a field for each value of another field and draws a line connecting the values. You can also use the Plot graph node to produce a line plot graph. That node offers some additional options. See the topic "Plot Node" on page 202 for more information. *Requires*: A pair of fields of any type.		**Area** Calculates a summary statistic for a field for each value of another field and draws an area connecting the values. The difference between a line and area is minimal in that an area resembles a line with the space colored below it. However, if you use a color aesthetic, this results in a simple split of the line and a stacking of the area. *Requires*: A pair of fields of any type.

Table 36. Available graph types (continued)

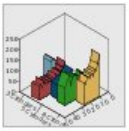
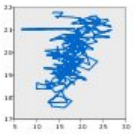
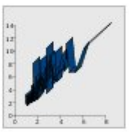
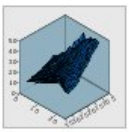
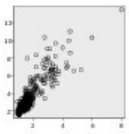
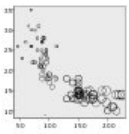
Chart icon	Description	Chart icon	Description
	**3-D Area** Displays the values of one field plotted against the values of another and split by a categorical field. An area element is drawn for each category. *Requires*: A categorical field and a pair of fields of any type.		**Path** Displays the values of one field plotted against the values of another, with a line connecting the values in the order they appear in the original dataset. The ordering is the primary difference between a path and a line. *Requires*: A pair of fields of any type.
	**Ribbon** Calculates a summary statistic for a field for each value of another field and draws a ribbon connecting the values. A ribbon is essentially a line with 3-D effects. It is not a true 3-D graph. *Requires*: A pair of fields of any type.		**Surface** Displays the values of three fields plotted against the values of one another, with a surface connecting the values. *Requires* : Three fields of any type.
	**Scatterplot** Displays the values of one field plotted against the values of another. This graph can highlight the relationship between the fields (if there is one). You can also use the Plot graph node to produce a scatterplot. That node offers some additional options. See the topic "Plot Node" on page 202 for more information. *Requires*: A pair of fields of any type.		**Bubble plot** Like a basic scatterplot, displays the values of one field plotted against the values of another. The difference is that the values of a third field are used to vary the size of the individual points. *Requires*: Three fields of any type.

Table 36. Available graph types (continued)

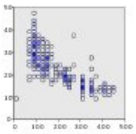
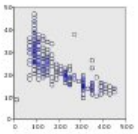
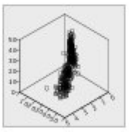
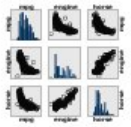
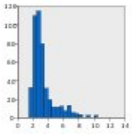
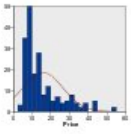
Chart icon	Description	Chart icon	Description
	**Binned Scatterplot** Like a basic scatterplot, displays the values of one field plotted against the values of another. The difference is that similar values are binned into groups and that the color or size aesthetic is used to indicate the number of cases in each bin. *Requires* : A pair of continuous fields.		**Hex Binned Scatterplot** See the description of Binned Scatterplot. The difference is the shape of the underlying bins, which are shaped like hexagons rather than circles. The resulting hex binned scatterplot looks similar to the binned scatterplot. However, the number of values in each bin will differ between the graphs because of the shape of the underlying bins. *Requires*: A pair of continuous fields.
	**3-D Scatterplot** Displays the values of three fields plotted against each other. This graph can highlight the relationship between the fields (if there is one). You can also use the Plot graph node to produce a 3-D scatterplot. That node offers some additional options. See the topic "Plot Node" on page 202 for more information. *Requires*: Three fields of any type.		**Scatterplot Matrix (SPLOM)** Displays the values of one field plotted against the values of another for every field. A SPLOM is like a table of scatterplots. The SPLOM also includes a histogram of each field. *Requires*: Two or more continuous fields.
	**Histogram** Displays the frequency distribution of a field. A histogram can help you determine the distribution type and see whether the distribution is skewed. You can also use the Histogram graph node to produce this graph. That node offers some additional options. See the topic "Histogram Plot Tab" on page 215 for more information. *Requires*: A single field of any type.		**Histogram with Normal Distribution** Displays the frequency distribution of a continuous field with a superimposed curve of the normal distribution. *Requires*: A single continuous field.

Table 36. Available graph types (continued)

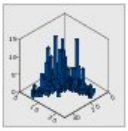
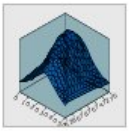
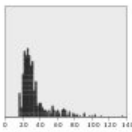
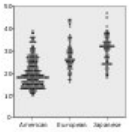
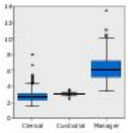
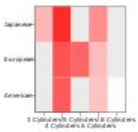
Chart icon	Description	Chart icon	Description
	**3-D Histogram** Displays the frequency distribution of a pair of continuous fields. *Requires* : A pair of continuous fields.		**3-D Density** Displays the frequency distribution of a pair of continuous fields. This is similar to a 3-D histogram, with the only difference being that a surface is used instead of bars to show the distribution. *Requires*: A pair of continuous fields.
	**Dot plot** Displays the individual cases/ rows and stacks them at the distinct data points on the *x* axis. This graph is similar to a histogram in that it shows the distribution of the data, but it displays each case/row rather than an aggregated count for a specific bin (range of values). *Requires* : A single field of any type.		**2-D Dot Plot** Displays the individual cases/ rows and stacks them at the distinct data points on the *y* axis for each category of a categorical field. *Requires*: A categorical field and a continuous field.
	**Boxplot** Calculates the five statistics (minimum, first quartile, median, third quartile, and maximum) for a continuous field for each category of a categorical field. The results are displayed as boxplot/ schema elements. The boxplots can help you see how the distribution of continuous data varies among the categories. *Requires*: A categorical field and a continuous field.		**Heat map** Calculates the mean for a continuous field for the crossing of categories between two categorical fields. *Requires* : A pair of categorical fields and a continuous field.

Table 36. Available graph types (continued)

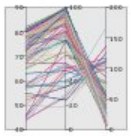
Chart icon	Description	Chart icon	Description
	**Parallel** Creates parallel axes for each field and draws a line through the field value for every row/ case in the data. *Requires*: Two or more continuous fields.		**Choropleth of Counts** Calculates the count for each category of a categorical field (**Data Key**) and draws a map that uses color saturation to represent the counts in the map features that correspond to the categories. *Requires* : A categorical field. A map file whose key matches the **Data Key** categories.
	Choropleth of Means/ Medians/Sums Calculates the mean, median, or sum of a continuous field (**Color**) for each category of a categorical field (**Data Key**) and draws a map that uses color saturation to represent the calculated statistics in the map features that correspond to the categories. *Requires*: A categorical field and a continuous field. A map file whose key matches the **Data Key** categories.		**Choropleth of Values** Draws a map that uses color to represent values of a categorical field (**Color**) for the map features that correspond to values defined by another categorical field (**Data Key**). If there are multiple categorical values of the Color field for each feature, the modal value is used. *Requires*: A pair of categorical fields. A map file whose key matches the **Data Key** categories.
	Coordinates on a Choropleth of Counts This is similar to Choropleth of Counts except that there are two additional continuous fields (**Longitude** and **Latitude**) that identify coordinates for drawing points on the choropleth map. *Requires*: A categorical field and a pair of continuous fields. A map file whose key matches the **Data Key** categories.		**Coordinates on a Choropleth of Means/Medians/Sums** This is similar to Choropleth of Means/Medians/Sums except that there are two additional continuous fields (**Longitude** and **Latitude**) that identify coordinates for drawing points on the choropleth map. *Requires* : A categorical field and three continuous fields. A map file whose key matches the **Data Key** categories.

Table 36. Available graph types (continued)

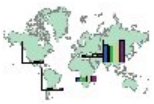
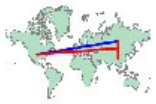
Chart icon	Description	Chart icon	Description
	Coordinates on a Choropleth of Values This is similar to Choropleth of Values except that there are two additional continuous fields (**Longitude** and **Latitude**) that identify coordinates for drawing points on the choropleth map. *Requires*: A pair of categorical fields and a pair of continuous fields. A map file whose key matches the **Data Key** categories.		**Bars of Counts on a Map** Calculates the proportion of rows/cases in each category of a categorical field (**Categories**) for each map feature (**Data Key**) and draws a map and the bar charts at the center of each map feature. *Requires*: A pair of categorical fields. A map file whose key matches the **Data Key** categories.
	**Bars on a Map** Calculates a summary statistic for a continuous field (**Values**) and displays the results for each category of a categorical field (**Categories**) for each map feature (**Data Key**) as bar charts positioned in the center of each map feature. *Requires*: A pair of categorical fields and a continuous field. A map file whose key matches the **Data Key** categories.		**Pie of Counts of a Map** Displays the proportion of rows/cases in each category of a categorical field (**Categories**) for each map feature (**Data Key**) and draws a map and the proportions as slices of a pie chart at the center of each map feature. *Requires*: A pair of categorical fields. A map file whose key matches the **Data Key** categories.
	Pie on a Map Calculates the sum of a continuous field (**Values**) in each category of a categorical field (**Categories**) for each map feature (**Data Key**) and draws a map and the sums as slices of a pie chart at the center of each map feature. *Requires*: A pair of categorical fields and a continuous field. A map file whose key matches the **Data Key** categories.		**Line Chart on a Map** Calculates a summary statistic for a continuous field (**Y**) for each value of another field (**X**) for each map feature (**Data Key**) and draws a map and line charts connecting the values at the center of each map feature. *Requires*: A categorical field and a pair of fields of any type. A map file whose key matches the **Data Key** categories.

Table 36. Available graph types (continued)

Chart icon	Description	Chart icon	Description
	Coordinates on a Reference Map Draws a map and points using continuous fields (**Longitude** and **Latitude**) that identify coordinates for the points. *Requires*: A pair of range fields. A map file.		**Arrows on a Reference Map** Draws a map and arrows using continuous fields that identify the start points (**Start Long** and **Start Lat**) and end points (**End Long** and **End Lat**) for each arrow. Each record/case in the data results in an arrow in the map. *Requires* : Four continuous fields. A map file.
	Point Overlay Map Draws a reference map and superimposes another point map over it with point features colored by a categorical field (**Color**). *Requires*: A pair of categorical fields. A point map file whose key matches the **Data Key** categories. A reference map file.		**Polygon Overlay Map** Draws a reference map and superimposes another polygon map over it with polygon features colored by a categorical field (**Color**). *Requires*: A pair of categorical fields. A polygon map file whose key matches the **Data Key** categories. A reference map file.
	**Line Overlay Map** Draws a reference map and superimposes another line map over it with line features colored by a categorical field (**Color**). *Requires*: A pair of categorical fields. A line map file whose key matches the **Data Key** categories. A reference map file.		

Creating Map Visualizations

For many visualizations, you have to make only two choices: fields (variables) of interest and a template to visualize those fields. No additional choices or actions are required. Map visualizations require at least one additional step: select a map file that defines the geographic information for the map visualization.

The basic steps for creating a simple map are the following:

1. Select the fields of interest on the Basic tab. For information about the type and number of fields required for different map visualizations, see “Available Built-in Graphboard Visualization Types” on page 176 .
2. Select a map template.

3. Click the Detailed tab.
4. Check that the **Data Key** and other required dropdown lists are set to the correct fields.
5. In the Map Files group, click **Select a Map File**.
6. Use the Select Maps dialog box to choose the map file and map key. The map key's values must match the values for of the field specified by **Data Key**. You can use the **Compare** button to compare these values. If you select an overlay map template, you will also need to choose a reference map. The reference map is not keyed to the data. It is used as the background for the main map. For more information about the Select Maps dialog box, see “Selecting map files for map visualizations” on page 175 .
7. Click **OK** to close the Select Maps dialog box.
8. In the Graphboard Template Chooser, click **Run** to create the map visualization.

Graphboard Examples

This section includes several different examples to demonstrate the available options. The examples also provide information for interpreting the resulting visualizations.

These examples use the stream named *graphboard.str*, which references the data files named *employee_data.sav*, *customer_subset.sav*, and *worldsales.sav*. These files are available from the *Demos* folder of any IBM SPSS Modeler Client installation. This can be accessed from the IBM SPSS Modeler program group on the Windows Start menu. The *graphboard.str* file is in the *streams* folder.

It is recommended that you read the examples in the order presented. Subsequent examples build on previous ones.

Example: Bar Chart with a Summary Statistic

We will create a bar chart that summarizes a continuous numeric field/variable for each category of a set/categorical variable. Specifically, we will create a bar chart that shows the mean salary for men and women.

This and several of the following examples use *Employee data*, which is a hypothetical dataset containing information about a company's employees.

1. Add a Statistics File source node that points to *employee_data.sav*.
2. Add a Graphboard node and open it for editing.
3. On the Basic tab, select *Gender* and *Current Salary* . (Use Ctrl+Click to select multiple fields/variables.)
4. Select **Bar**.
5. From the Summary drop-down list, select **Mean**.
6. Click **Run**.
7. On the resulting display, click the "Display field and value labels" toolbar button (the second of the group of two in the center of the toolbar).

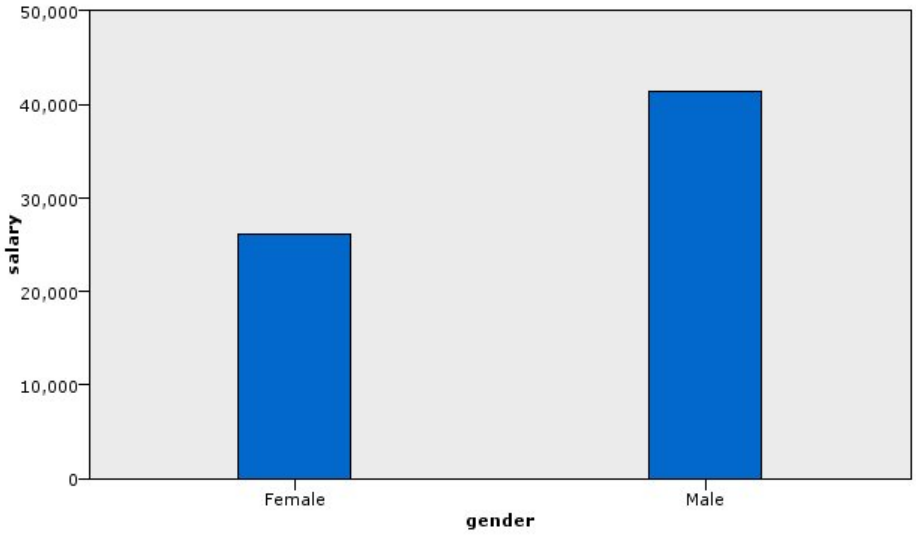


Figure 8. Bar chart with a summary statistic

We can observe the following:

- Based on the height of the bars, it is clear that the mean salary for males is greater than the mean salary for females.

Example: Stacked Bar Chart with a Summary Statistic

We will now create a stacked bar chart to see whether the difference in mean salary between males and females is subject to job type. Perhaps females, on average, make more than males for certain job types.

Note: This example uses *Employee data*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Employment Category* and *Current Salary* . (Use Ctrl+Click to select multiple fields/variables.)
3. Select **Bar**.
4. From the Summary list, select **Mean**.
5. Click the Detailed tab. Note that your selections from the previous tab are reflected here.
6. In the Optional Aesthetics group, choose *gender* from the Color drop-down list.
7. Click **Run**.

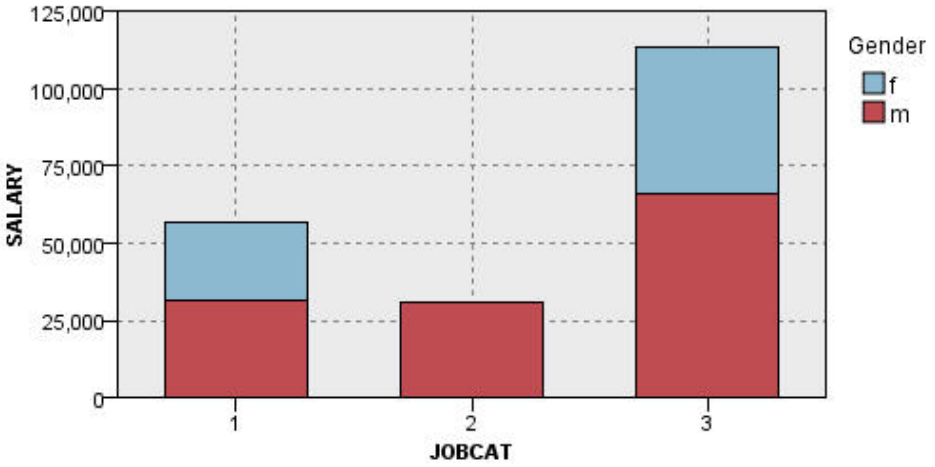


Figure 9. Stacked bar chart

We can observe the following:

- The difference in mean salaries for each job type does not appear to be as great as it was in the bar chart that compared the mean salaries for all males and females. Perhaps there are varying number of males and females in each group. You could check this by creating a bar chart of counts.
- Regardless of job type, the mean salary for males is always greater than the mean salary for females.

Example: Paneled Histogram

We will create a histogram that is paneled by gender so we can compare the frequency distributions of salary for men and women. The frequency distribution shows how many cases/rows lie within specific salary ranges. The paneled histogram can help us further analyze the difference in salaries between genders.

Note: This example uses *Employee data*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Current Salary* .
3. Select **Histogram**.
4. Click the Detailed tab.
5. In the Panels and Animation group, choose *gender* from the Panel Across drop-down list.
6. Click **Run**.

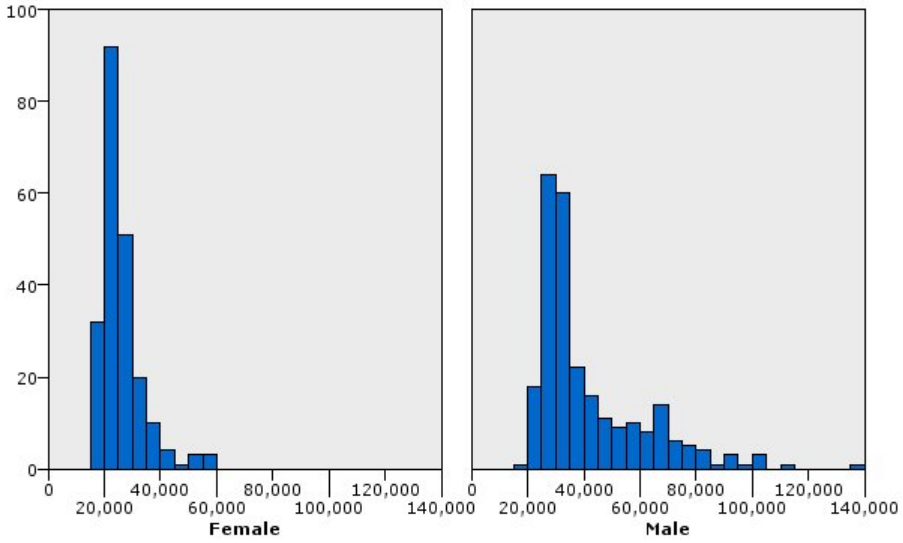


Figure 10. Paneled histogram

We can observe the following:

- Neither frequency distribution is a normal distribution. That is, the histograms do not resemble bell curves, as they would if the data were normally distributed.
- The taller bars are on the left side of each graph. Therefore, for both men and women, more make lower rather than higher salaries.
- The frequency distributions of salary among men and women are not equal. Notice the shape of the histograms. There are more men who make higher salaries than woman who make higher salaries.

Example: Paneled Dot Plot

Like a histogram, a dot plot shows the distribution of a continuous numeric range. Unlike a histogram, which shows counts for binned ranges of data, a dot plot shows every row/case in the data. Therefore, a dot plot provides more granularity compared to the histogram. In fact, using a dot plot may be the preferred starting point when analyzing frequency distributions.

Note: This example uses *Employee data*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Current Salary* .
3. Select **Dot Plot**.
4. Click the Detailed tab.
5. In the Panels and Animation group, choose *gender* from the Panel Across drop-down list.
6. Click **Run**.
7. Maximize the resulting output window to view the plot more clearly.

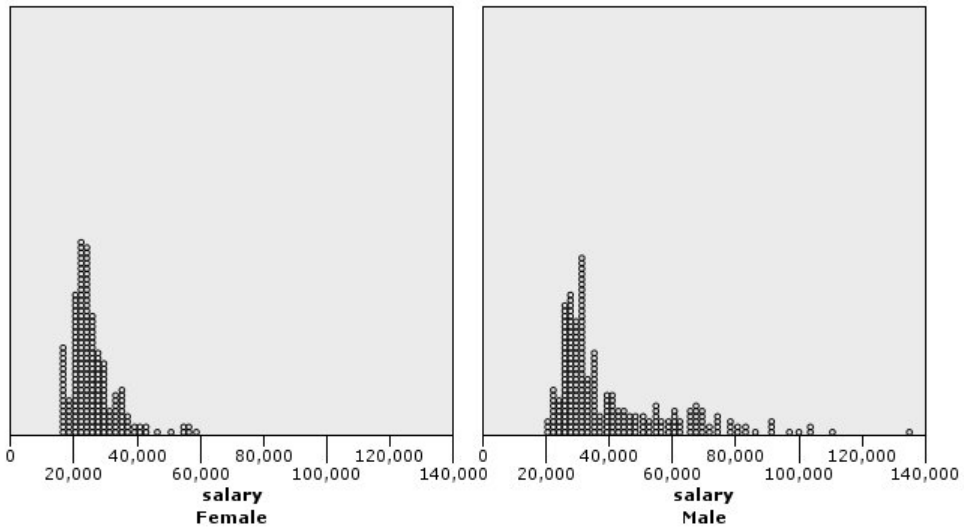


Figure 11. Paneled dot plot

Compared to the histogram (see “Example: Paneled Histogram” on page 186), we can observe the following:

- The peak at 20,000 that appeared in the histogram for females is less dramatic in the dot plot. There are many cases/rows concentrated around that value, but most of those values are closer to 25,000. This level of granularity is not apparent in the histogram.
- Although the histogram for males suggests that the mean salary for males gradually declines after 40,000, the dot plot shows that the distribution is fairly uniform after that value, to 80,000. At any one salary value in that range, there are three or more males who earn that particular salary.

Example: Boxplot

A boxplot is another useful visualization for viewing how the data are distributed. A boxplot contains several statistical measures that we will explore after creating the visualization.

Note: This example uses *Employee data*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Gender* and *Current Salary* . (Use Ctrl+Click to select multiple fields/variables.)
3. Select **Boxplot**.
4. Click **Run**.

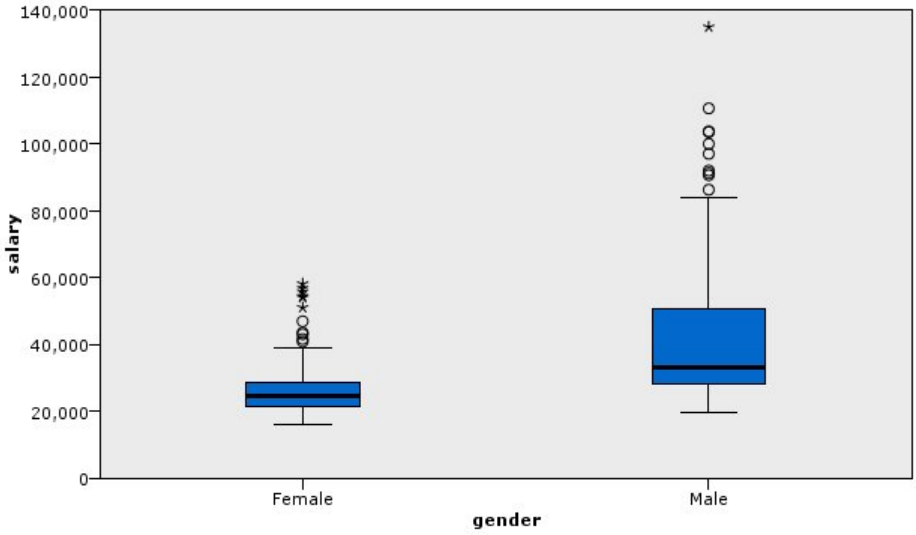


Figure 12. Boxplot

Let's explore the different parts of the boxplot:

- The dark line in the middle of the boxes is the median of *salary*. Half of the cases/rows have a value greater than the median, and half have a value lower. Like the mean, the median is a measure of central tendency. Unlike the mean, it is less influenced by cases/rows with extreme values. In this example, the median is lower than the mean (compare to "Example: Bar Chart with a Summary Statistic" on page 184). The difference between the mean and median indicates that there are a few cases/rows with extreme values that are elevating the mean. That is, there are a few employees who earn large salaries.
- The bottom of the box indicates the 25th percentile. Twenty-five percent of cases/rows have values below the 25th percentile. The top of the box represents the 75th percentile. Twenty-five percent of cases/rows have values above the 75th percentile. This means that 50% of the case/rows lie within the box. The box is much shorter for females than for males. This is one clue that *salary* varies less for females than for males. The top and bottom of the box are often called **hinges**.
- The T-bars that extend from the boxes are called **inner fences** or **whiskers**. These extend to 1.5 times the height of the box or, if no case/row has a value in that range, to the minimum or maximum values. If the data are distributed normally, approximately 95% or the data are expected to lie between the inner fences. In this example, the inner fences extend less for females compared to males, another indication that *salary* varies less for females than for males.
- The points are **outliers**. These are defined as values that do not fall in the inner fences. Outliers are extreme values. The asterisks or stars are **extreme outliers**. These represent cases/rows that have values more than three times the height of the boxes. There are several outliers for both females and males. Remember that the mean is greater than the median. The greater mean is caused by these outliers.

Example: Pie Chart

We will now use a different dataset to explore some other visualization types. The dataset is *customer_subset*, which is a hypothetical data file that contains information about customers.

We will first create a pie chart to check the proportion of customers in different geographic regions.

1. Add a Statistics File source node that points to *customer_subset.sav*.
2. Add a Graphboard node and open it for editing.
3. On the Basic tab, select *Geographic indicator* .
4. Select **Pie of Counts**.
5. Click **Run**.

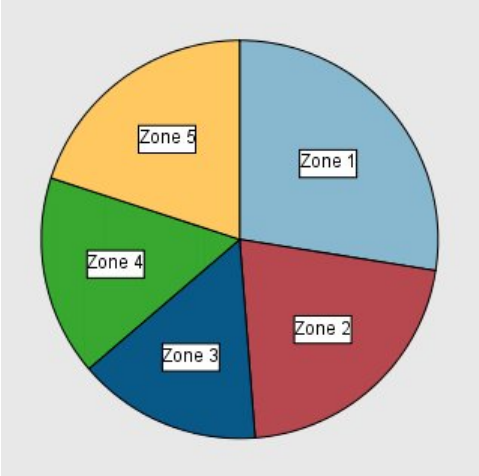


Figure 13. Pie chart

We can observe the following:

- Zone 1 has more customers than in each of the other zones.
- Customers are equally distributed among the other zones.

Example: Heat Map

We will now create a categorical heat map to check the mean income for customers in different geographic regions and age groups.

Note: This example uses *customer_subset*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Geographic indicator*, *Age category*, and *Household income in thousands* , in that order. (Use Ctrl+Click to select multiple fields/variables.)
3. Select **Heat Map**.
4. Click **Run**.
5. On the resulting output window, click the "Display field and value labels" toolbar button (the right-hand one of the two in the center of the toolbar).

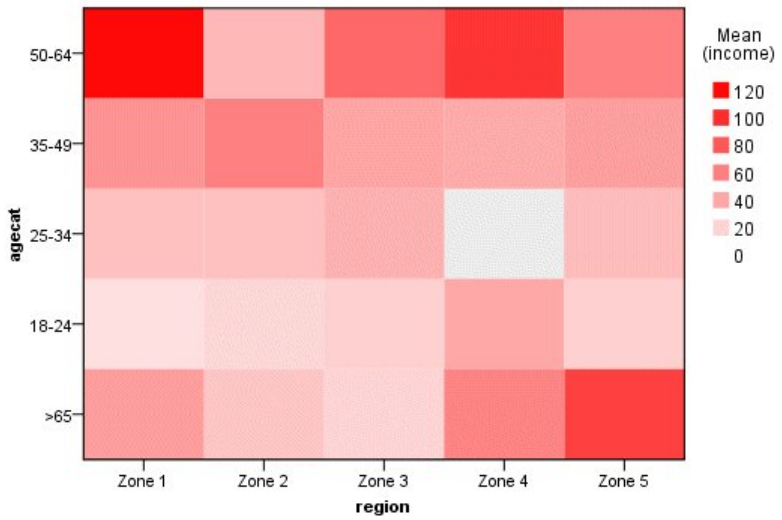


Figure 14. Categorical heat map

We can observe the following:

- A heat map is like a table that uses colors instead of numbers to represent the values for the cells. Bright, deep red indicates the highest value, while gray indicates a low value. The value of each cell is the mean of the continuous field/variable for each pair of categories.
- Except in Zone 2 and Zone 5, the group of customers whose age is between 50 and 64 have a greater mean household income than those in other groups.
- There are no customers between the ages of 25 and 34 in Zone 4.

Example: Scatterplot Matrix (SPLOM)

We will create a scatterplot matrix of several different variables so that we can determine whether there are any relationships among the variables in the dataset.

Note: This example uses *customer_subset*.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Age in years*, *Household income in thousands*, and *Credit card debt in thousands* . (Use Ctrl+Click to select multiple fields/variables.)
3. Select **SPLOM**.
4. Click **Run**.
5. Maximize the output window to view the matrix more clearly.

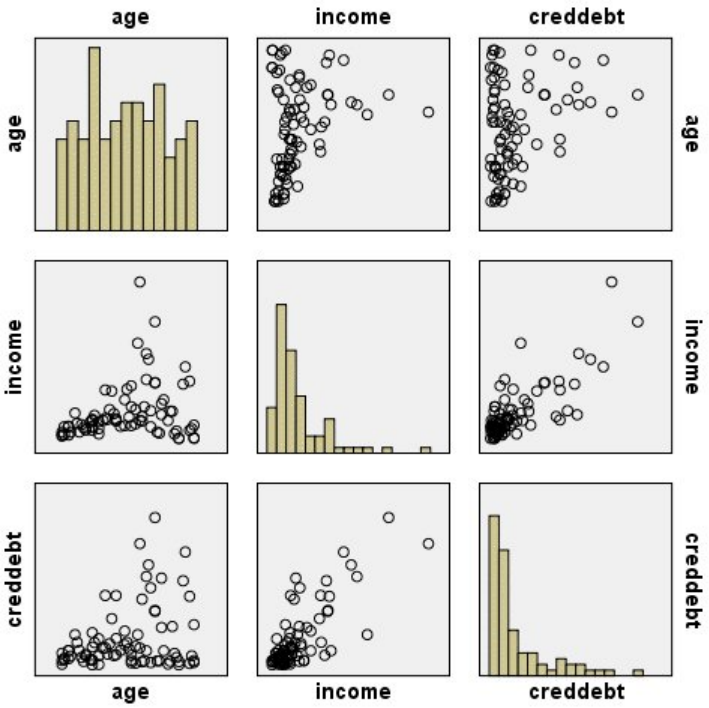


Figure 15. Scatterplot matrix (SPLOM)

We can observe the following:

- The histograms displayed on the diagonal show the distribution of each variable in the SPLOM. The histogram for *age* appears in the upper left cell, that for *income* in the center cell, and that for *creddebt* in the lower right cell. None of the variables appears normally distributed. That is, none of the histograms resembles a bell curve. Also, note that the histograms for *income* and *creddebt* are positively skewed.
- There does not seem to be any relationship between *age* and the other variables.
- There is a linear relationship between *income* and *creddebt*. That is, *creddebt* increases as *income* increases. You may want to create individual scatterplots of the these variables and the other related variables to explore the relationships further.

Example: Choropleth (Color Map) of Sums

We will now create a map visualization. Then, in the subsequent example, we will create a variation of this visualization. The dataset is *worldsales*, which is a hypothetical data file that contains sales revenue by continent and product.

1. Add a Graphboard node and open it for editing.
2. On the Basic tab, select *Continent* and *Revenue*. (Use Ctrl+Click to select multiple fields/variables.)
3. Select **Choropleth of Sums**.
4. Click the Detailed tab.
5. In the Optional Aesthetics group, choose *Continent* from the Data Label drop-down list.
6. In the Map Files group, click **Select a Map File**.
7. In the Select Maps dialog box, check that **Map** is set to *Continents* and **Map Key** is set to *CONTINENT*.
8. In the Compare Map and Data Values groups, click **Compare** to ensure the map keys match the data keys. In this example, all of data key values have matching map keys and features. We can also see that there is no data for Oceania.
9. In the Select Maps dialog box, click **OK**.
10. Click **Run**.

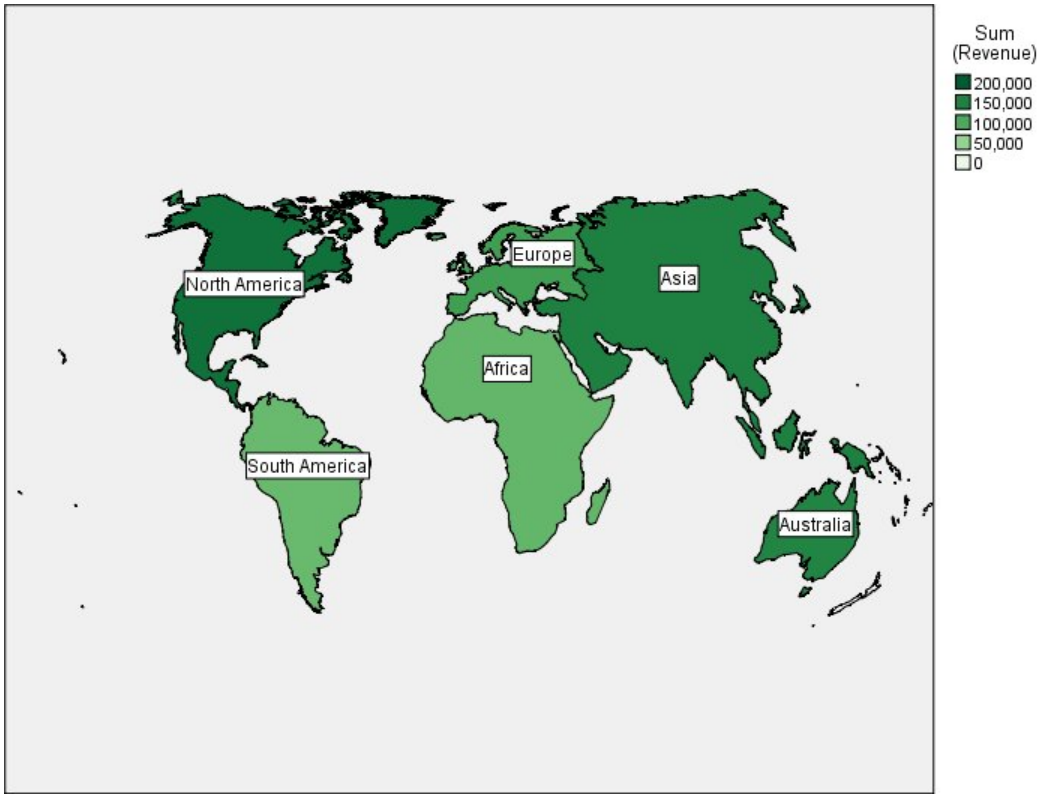


Figure 16. Choropleth of Sums

From that map visualization, we can easily see that revenue is highest in North America and lowest in South America and Africa. Each continent is labeled because we used *Continent* for the data label aesthetic.

Example: Bar Charts on a Map

This example shows how revenue breaks down by product in each continent.

Note: This example uses *worldsales*.

1. Add a Graphboard node and open it for editing.

2. On the Basic tab, select *Continent*, *Product*, and *Revenue*. (Use Ctrl+Click to select multiple fields/variables.)
3. Select **Bars on a Map**.
4. Click the Detailed tab.

   When using more than one field of a specific type, it is important to check that each field is assigned to the correct slot.
5. From the Categories drop-down list, choose *Product*.
6. From the Values drop-down list, choose *Revenue*.
7. From the Data Key drop-down list, choose *Continent*.
8. From the Summary drop-down list, choose *Sum*.
9. In the Map Files group, click **Select a Map File**.
10. In the Select Maps dialog box, check that **Map** is set to *Continents* and **Map Key** is set to *CONTINENT*.
11. In the Compare Map and Data Values groups, click **Compare** to ensure the map keys match the data keys. In this example, all of data key values have matching map keys and features. We can also see that there is no data for Oceania.
12. On the Select Maps dialog box, click **OK**.
13. Click **Run**.
14. Maximize the resulting output window to see the display more clearly.

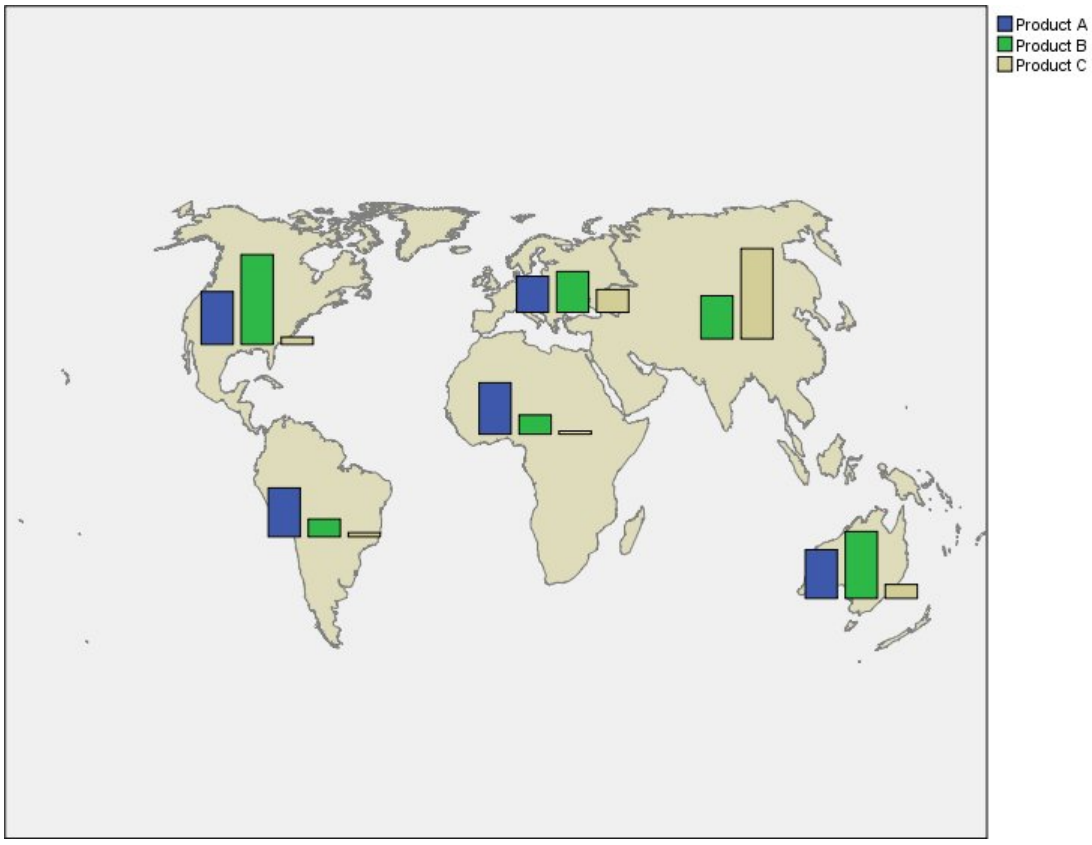


Figure 17. Bar Charts on a Map

We can observe the following:

- The distribution of total revenue across products is very similar in South America and Africa.
- *Product C* generates the least revenue everywhere except in Asia.
- There is no or minimal revenue from *Product A* in Asia.

Graphboard appearance tab

You can specify appearance options before graph creation.

General appearance options

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

Sampling. Specify a method for larger datasets. You can specify a maximum dataset size or use the default number of records. Performance is enhanced for large datasets when you select the **Sample** option. Alternatively, you can choose to plot all data points by selecting **Use all data**, but you should note that this may dramatically decrease the performance of the software.

Stylesheet appearance options

There are two buttons that enable you to control which visualization templates (and stylesheets, and maps) are available:

Manage. Manage visualization templates, stylesheets, and maps on your computer. You can import, export, rename, and delete visualization templates, stylesheets, and maps on your local machine. See the topic "Managing Templates, Stylesheets, and Map Files" on page 195 for more information.

Location. Change the location in which visualization templates, stylesheets, and maps are stored. The current location is listed to the right of the button. See the topic "Setting the Location of Templates, Stylesheets, and Maps" on page 195 for more information.

The following example shows where appearance options are placed on a graph. (Note that not all graphs use all these options.)

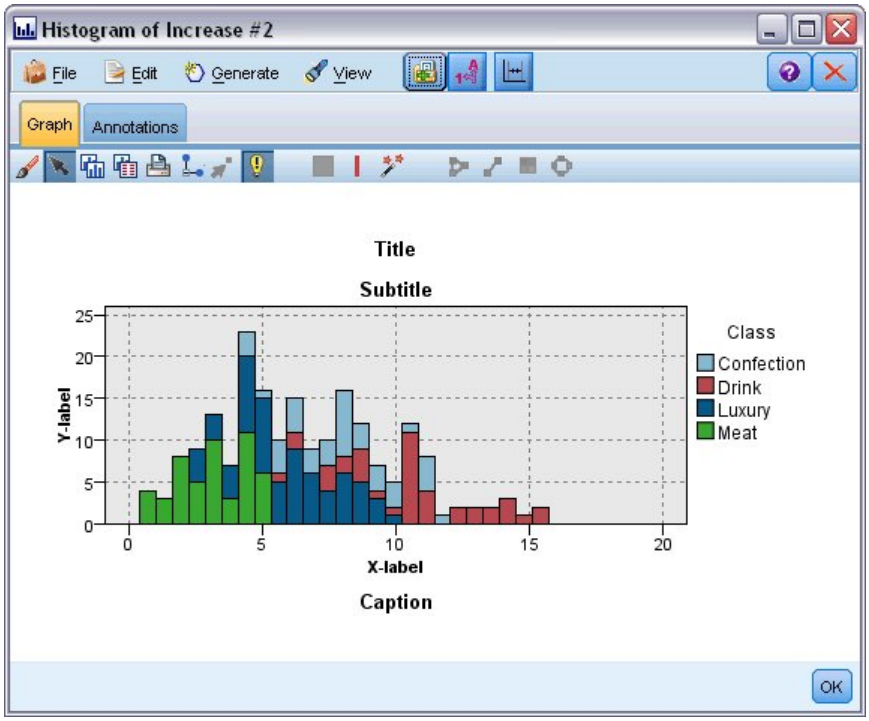


Figure 18. Position of various graph appearance options

Setting the Location of Templates, Stylesheets, and Maps

Visualization templates, visualization stylesheets, and map files are stored in a specific local folder or in the IBM SPSS Collaboration and Deployment Services Repository. When selecting templates, stylesheets, and maps, only the built-in ones in this location are displayed. By keeping all templates, stylesheets, and map files in one place, IBM SPSS applications can easily access them. For information about adding additional templates, stylesheets, and map files to this location, see "Managing Templates, Stylesheets, and Map Files" on page 195 .

How to Set the Location of Templates, Stylesheets, and Map Files

1. In a template or stylesheet dialog box, click **Location...** to display the Templates, Stylesheets, and Maps dialog box.
2. Select an option for the default location for templates, stylesheets, and map files:

   Local Machine. Templates, stylesheets, and map files are located in a specific folder on your local computer. On Windows XP, this folder is *C:\Documents and Settings\<user>\Application Data\SPSSInc\Graphboard*. The folder cannot be changed.

   IBM SPSS Collaboration and Deployment Services Repository . Templates, stylesheets, and map files are located in a user-specified folder in the IBM SPSS Collaboration and Deployment Services Repository. To identify the specific folder, click **Folder**. For more information, see "Using the IBM SPSS Collaboration and Deployment Services Repository as the Template, Stylesheet, and Map File Location" on page 195 .
3. Click **OK**.

Using the IBM SPSS Collaboration and Deployment Services Repository as the Template, Stylesheet, and Map File Location

Visualization templates and stylesheets can be stored in the IBM SPSS Collaboration and Deployment Services Repository. This location is a specific folder in the IBM SPSS Collaboration and Deployment Services Repository. If this is set as the default location, any templates, stylesheets, and map files in this location are available for selection.

How to Set a Folder in IBM SPSS Collaboration and Deployment Services Repository as the Location for Templates, Stylesheets, and Map Files

1. In a dialog box with a Location button, click **Location...**.
2. Select IBM SPSS Collaboration and Deployment Services Repository.
3. Click **Folder**.

   Note: If you are not already connected to the IBM SPSS Collaboration and Deployment Services Repository, you will be prompted for connection information.
4. In the Select Folder dialog box, select the folder in which templates, stylesheets, and map files are stored.
5. Optionally, you can select a label from **Retrieve Label**. Only templates, stylesheets, and map files with that label will be displayed.
6. If you are looking for a folder that contains a particular template or stylesheet, you may want to search for the template, stylesheet, or map file on the Search tab. The Select Folder dialog box automatically selects the folder in which the found template, stylesheet, or map file is located.
7. Click **Select Folder**.

Managing Templates, Stylesheets, and Map Files

You can manage the templates, stylesheets, and map files in the local location on your computer by using the Manage Templates, Stylesheets, and Maps dialog box. This dialog box allows you to import, export, rename, and delete visualization templates, stylesheets, and map files in the local location on your computer.

Click **Manage...** in one of the dialog boxes where you select templates, stylesheets, or maps.

Manage Templates, Stylesheets, and Maps Dialog Box

The Template tab lists all the local templates. The Stylesheet tab lists all the local stylesheets, as well as displaying example visualizations with sample data. You can select one of the stylesheets to apply its styles to the example visualizations. See the topic "Applying stylesheets" on page 265 for more information. The Map tab lists all the local map files. This tab also displays the map keys including sample values, a comment if one was provided when creating the map, and a preview of the map.

The following buttons operate on whichever tab is currently activated.

Import. Import a visualization template, stylesheet, or map file from the file system. Importing a template, stylesheet, or map file makes it available to the IBM SPSS application. If another user sent you a template, stylesheet, or map file, you import the file before using it in your application.

Export. Export a visualization template, stylesheet, or map file to the file system. Export a template, stylesheet, or map file when you want to send it to another user.

Rename. Rename the selected visualization template, stylesheet, or map file. You cannot change a name to one that's already used.

Export Map Key. Export the map keys as a comma-separated values (CSV) file. This button is enabled only on the Map tab.

Delete. Delete the selected visualization template(s), stylesheet(s), or map file(s). You can select multiple templates, stylesheets, or map files with Ctrl-click. There is no undo action for deleting so proceed with caution.

Converting and Distributing Map Shapefiles

The Graphboard Template Chooser allows you to create map visualizations from the combination of a visualization template and an SMZ file. SMZ files are similar to ESRI shapefiles (SHP file format) in that they contain the geographic information for drawing a map (for example, country borders), but they are optimized for map visualizations. The Graphboard Template Chooser is pre-installed with a select number of SMZ files. If you have an existing ESRI shapefile that you want to use for map visualizations, you first need to convert the shapefile to an SMZ file using the Map Conversion Utility. The Map Conversion Utility supports point, polyline, or polygon (shape types 1, 3, and 5) ESRI shapefiles containing a single layer.

In addition to converting ESRI shapefiles, the Map Conversion Utility allows you to modify the map's level of detail, change feature labels, merge features, and move features, among many other optional changes. You can also use the Map Conversion Utility to modify an existing SMZ file (including the pre-installed ones).

Editing pre-installed SMZ files

1. Export the SMZ file from the Manage system. See the topic "Managing Templates, Stylesheets, and Map Files" on page 195 for more information.
2. Use the Map Conversion Utility to open and edit the exported SMZ file. It is recommended that you save the file with a different name. See the topic "Using the Map Conversion Utility" on page 197 for more information.
3. Import the modified SMZ file into the Manage system. See the topic "Managing Templates, Stylesheets, and Map Files" on page 195 for more information.

Additional Resources for Map Files

Geospatial data in the SHP file format, which could be used to support your mapping needs, is available from many private and public sources. Check with local government web sites if you seek free data. Many of the templates in this product are based on publicly available data obtained from GeoCommons (http://www.geocommons.com) and the U.S. Census Bureau (http://www.census.gov).

IMPORTANT NOTICE: Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related

to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM program, unless indicated in a Notices file accompanying this IBM program, and use of the materials sites is at your own risk.

Key Concepts for Maps

Understanding some key concepts related to shapefiles will help you use the Map Conversion Utility effectively.

A **shapefile** provides the geographic information for drawing a map. There are three types of shapefiles that the Map Conversion Utility supports:

- **Point.** The shapefile identifies the locations of points, such as cities.
- **Polyline.** The shapefile identifies paths and their locations, such as rivers.
- **Polygon.** The shapefile identifies bounded regions and their locations, such as countries.

Most commonly, you will use a polygon shapefile. Choropleth maps are created from polygon shapefiles. Choropleth maps use color to represent a value within individual polygons (regions). Point and polyline shapefiles are typically overlaid on a polygon shapefile. An example is a point shapefile of U.S. cities overlaid on a polygon shapefile of U.S. states.

A shapefile is comprised of **features**. Features are the individual geographical entities. For example, features may be countries, states, cities, and so on. The shapefile also contains data about the features. These data are stored in **attributes**. Attributes are similar to fields or variables in a data file. There is at least one attribute that is the **map key** for the feature. The map key may be a label, such as country or state name. The map key is what you will link to a variable/field in a data file to create a map visualization.

Note that you will be able to retain only the key attribute or attributes in the SMZ file. The Map Conversion Utility does not support saving additional attributes. This means that you will need to create multiple SMZ files if you want to aggregate at different levels. For example, if you want to aggregate U.S. states and regions, you will need separate SMZ files: one that has a key identifying states and one that has a key identifying regions.

Using the Map Conversion Utility

How to start the Map Conversion Utility

From the menus choose:

Tools > **Map Conversion Utility**

There are four major screens (steps) in the Map Conversion Utility. One of the steps also involves sub-steps for more detailed control over editing the map file.

Step 1 - Choose Destination and Source Files

You first have to select a source map file and a destination for the converted map file. You will need both the *.shp* and *.dbf* files for the shapefile.

Select a .shp (ESRI) or .smz file for conversion. Browse to an existing map file on your computer. This is the file that you will convert to and save as an SMZ file. The *.dbf* file for the shapefile *must* be stored in the same location with a base file name that matches the *.shp* file. The *.dbf* file is required because it contains the attribute information for the *.shp* file.

Set a destination and file name for the converted map file. Enter a path and file name for the SMZ file that will be created from the original map source.

- **Import to the Template Chooser.** In addition to saving a file on the file system, you can optionally add the map to the Manage list of the Template Chooser. If you choose this option, the map will automatically be available in the Template Chooser for IBM SPSS products installed on your computer. If you don't import to the Template Chooser now, you will need to import it manually at a later time. For

more information about importing maps into the Template Chooser Manage system, see “Managing Templates, Stylesheets, and Map Files” on page 195 .

Step 2 - Choose Map Key

Now you will choose which map keys to include in the SMZ file. You can then change some options that will affect rendering of the map. The subsequent steps in the Map Conversion Utility include a preview of the map. The rendering options that you choose will be used to generate the map preview.

Choose primary map key. Select the attribute that is the primary key for identifying and labeling features in the map. For example, a world map's primary key may be the attribute identifying country names. The primary key will also link your data to the map features, so be sure that the values (labels) of the attribute you choose will match the values in your data. Example labels are displayed when you choose an attribute. If you need to change these labels, you will be able to do so in a later step.

Choose additional keys to include. In addition to the primary map key, check any other key attributes that you want to include in the generated SMZ file. For example, some attributes may contain translated labels. If you expect data coded in other languages, you may want to preserve these attributes. Note that you can choose only those additional keys that represent the same features as the primary key. For example if the primary key were the full names of U.S. states, you can select only those alternate keys that represent U.S. States, such as state abbreviations.

Automatically smooth the map. Shapefiles with polygons typically contain too many data points and too much detail for statistical map visualizations. The excess details can be distracting and negatively impact performance. You can reduce the level of detail and generalize the map with smoothing. The map will look crisper and render more quickly as a result. When the map is automatically smoothed, the maximum angle is 15 degrees and the percentage to keep is 99. For information about these settings, see “Smooth the map” on page 198 . Note that you have the opportunity to apply additional smoothing later in another step.

Remove boundaries between touching polygons in the same feature. Some features may include sub-features that have boundaries internal to the main features of interest. For example, a world map of continents may contain internal boundaries for the countries contained within each continent. If you choose this option, the internal boundaries will not appear in the map. In the world map of continents example, choosing this option would remove the country borders, while retaining the continent borders.

Step 3 - Edit the Map

Now that you have specified basic options for the map, you can edit more specific ones. These modifications are optional. This step of the Map Conversion Utility guides you through the associated tasks and displays a preview of the map so you can verify your changes. Some tasks may not be available depending on shapefile type (point, polyline, or polygon) and coordinate system.

Every task has the following common controls on the left side of the Map Conversion Utility.

Show the labels on the map. By default, feature labels are not shown in the preview. You can choose to display the labels. Although the labels can help identify features, they may interfere with direct selection on the preview map. Turn on this option when you need it, for example when you are editing the feature labels.

Color the map preview. By default, the preview map displays areas with a solid color. All features are the same color. You can choose to have an assortment of colors assigned to the individual map features. This option may help distinguish different features in the map. It will be especially helpful when you merge features and want to see how the new features are represented in the preview.

Every task also has the following common control on the right side of the Map Conversion Utility.

Undo. Click **Undo** to revert to the last state. You can undo a maximum of 100 changes.

Smooth the map

Shapefiles with polygons typically contain too many data points and too much detail for statistical map visualizations. The excess details can be distracting and negatively impact performance. You can reduce

the level of detail and generalize the map with smoothing. The map will look crisper and render more quickly as a result. This option is not available for point and polyline maps.

Max. angle. The maximum angle, which must be a value between 1 and 20, specifies the tolerance for smoothing sets of points that are almost linear. A larger value allows more tolerance for the linear smoothing and will subsequently drop more points, resulting in a more generalized map. To apply linear smoothing, the Map Conversion Utility checks the inner angle formed by every set of three points in the map. If 180 minus the angle is less than the specified value, the Map Conversion Utility will drop the middle point. In other words, the Map Conversion Utility is checking whether the line formed by the three points is almost straight. If it is, the Map Conversion Utility treats the line as a straight line between the endpoints and drops the middle point.

Percent to keep. The percentage to keep, which must be a value between 90 and 100, determines the amount of land area to keep when the map is smoothed. This option affects only those features that have multiple polygons, as would be the case if a feature included islands. If the total area of a feature minus a polygon is greater than the specified percentage of the original area, the Map Conversion Utility will drop the polygon from the map. The Map Conversion Utility will never remove all polygons for the feature. That is, there will always be at least one polygon for the feature, regardless of the amount of smoothing applied.

After you choose a maximum angle and percentage to keep, click **Apply**. The preview will update with the smoothing changes. If you need to smooth the map again, repeat until you obtain the level of smoothness you want. Note that there is a limit to the smoothing. If you repeatedly smooth, you will reach a point when no additional smoothing can be applied to the map.

Edit the feature labels

You can edit the feature labels as needed (perhaps to match your expected data) and also re-position the labels in the map. Even if you don't think you need to change the labels, you should review them before creating visualizations from the map. Because labels are not shown by default in the preview, you may also want to select **Show the labels on the map** to display them.

Keys. Select the key containing the feature labels that you want to review and/or edit.

Features. This list displays the feature labels contained in the selected key. To edit the label, double-click it in the list. If labels are shown on the map, you can also double-click the feature labels directly in the map preview. If you want to compare the labels to an actual data file, click **Compare**.

X/Y. These text boxes list the current center point of the selected feature's label on the map. The units are displayed in the map's coordinates. These may be local, Cartesian coordinates (for example, the State Plane Coordinate System) or geographic coordinates (where **X** is longitude and **Y** is latitude). Enter coordinates for the new position of the label. If labels are shown, you can also click and drag a label on the map to move it. The text boxes will update with the new position.

Compare. If you have a data file that contains data values that are intended to match the features labels for a particular key, click **Compare** to display the Compare to an External Data Source dialog box. In this dialog box you will be able to open the data file and compare its values directly with those in the map key's feature labels.

Compare to an External Data Source dialog box

The Compare to an External Data Source dialog box allows you to open a tab-separated values file (with a *.txt* extension), a comma-separated values file (with a *.csv* extension), or a data file formatted for IBM SPSS Statistics (with a *.sav* extension). When the file is open, you can select a field in the data file to compare to the feature labels in a specific map key. You can then correct any discrepancies in the map file.

Fields in the data file. Choose the field whose values you want to compare to the feature labels. If the first row in the *.txt* or *.csv* file contains descriptive labels for each field, check **Use first row as column labels**. Otherwise, each field will be identified by its position in the data file (for example, "Column 1", "Column 2", and so on).

Key to compare. Choose the map key whose feature labels you want to compare to the data file field values.

Compare. Click when you are ready to compare the values.

Comparison Results. By default, the Comparison Results table lists only the unmatched field values in the data file. The application tries to find a related feature label, usually by checking for inserted or missing spaces. Click the dropdown list in the *Map Label* column to match the feature label in the map file to the displayed field value. If there is no corresponding feature label in your map file, choose *Leave Unmatched*. If you want to see all field values, even those that already match a feature label, clear **Display only unmatched cases**. You may do this if you want to override one or more matches.

You can use each feature only once to match it to a field value. If you want to match multiple features to a single field value, you can merge the features and then match the new, merged feature to the field value. For more information about merging features, see "Merge features" on page 200 .

Merge features

Merging features is useful for creating larger regions in a map. For example, if you were converting a map of states, you could merge the states (the features in this example) into larger North, South, East, and West regions.

Keys. Select the map key containing the feature labels that will help you identify the features you want to merge.

Features. Click the first feature that you want to merge. Ctrl-click the other features that you want to merge. Note that the features will also be selected in the map preview. You can click and Ctrl-click the features directly in the map preview in addition to selecting them from the list.

After selecting the features that you want to merge, click **Merge** to display the Name the Merged Feature dialog box, where you will be able to apply a label to the new feature. You may want to check **Color the map preview** after merging features to ensure that the results are what you expect.

After merging features, you may also want to move the label for the new feature. You can do that in the *Edit the feature labels* task. See the topic "Edit the feature labels" on page 199 for more information.

Name the Merged Feature dialog box

The Name the Merged Feature dialog box allows you to assign labels to the new merged feature.

The Labels table displays information for each key in the map file and allows you to assign a label for each key.

New Label. Enter a new label for the merged feature to assign to the specific map key.

Key. The map key to which you are assigning the new label.

Old Labels. The labels for the features that will be merged into the new feature.

Remove boundaries between touching polygons. Check this option to remove the boundaries from the features that were merged together. For example, if you merged states into geographic regions, this option would remove the boundaries around the individual states.

Move features

You can move features in the map. This can be useful when you want to put features together, like a mainland and outlying islands.

Keys. Select the map key containing the feature labels that will help you identify the features you want to move.

Features. Click the feature that you want to move. Note that the feature will be selected in the map preview. You can also click the feature directly in the map preview.

X/Y. These text boxes list the current center point of the feature on the map. The units are displayed in the map's coordinates. These may be local, Cartesian coordinates (for example, the State Plane Coordinate System) or geographic coordinates (where **X** is longitude and **Y** is latitude). Enter coordinates

for the new position of the feature. You can also click and drag a feature on the map to move it. The text boxes will update with the new position.

Delete features

You can delete unwanted features from the map. This can be useful when you want to remove some clutter by deleting features that won't be of interest in the map visualization.

Keys. Select the map key containing the feature labels that will help you identify the features you want to delete.

Features. Click the feature that you want to delete. If you want to delete multiple features simultaneously, Ctrl-click the additional features. Note that the features will also be selected in the map preview. You can click and Ctrl-click the features directly in the map preview in addition to selecting them from the list.

Delete individual elements

In addition to deleting whole features, you can delete some of the individual elements that compose the features, such as lakes and small islands. This option is not available for point maps.

Elements. Click the elements that you want to delete. If you want to delete multiple elements simultaneously, Ctrl-click the additional elements. Note that the elements will also be selected in the map preview. You can click and Ctrl-click the elements directly in the map preview in addition to selecting them from the list. Because the list of element names is not descriptive (each element is assigned a number within the feature), you should check the selection in the map preview to ensure you have selected the elements you want.

Set the projection

The map projection specifies the way in which the three-dimensional Earth is represented in two dimensions. All projections cause distortions. However, some projections are more suitable depending on whether you are viewing a global map or a more local one. Also some projections preserve the shape of the original features. Projections that preserve shape are conformal projections. This option is available only for maps with geographic coordinates (longitude and latitude).

Unlike other options in the Map Conversion Utility, the projection can be changed after a map visualization is created.

Projection. Select a map projection. If you are creating a global or hemispherical map, use the *Local*, *Mercator*, or *Winkel Tripel* projections. For smaller areas, use the *Local*, *Lambert Conformal Conic*, or *Transverse Mercator* projections. All projections use the WGS83 ellipsoid for the datum.

- The **Local** projection is always used when the map was created with a local coordinate system, such as the State Plane Coordinate System. These coordinate systems are defined by Cartesian coordinates rather than geographic coordinates (longitude and latitude). In the Local projection, horizontal and vertical lines are equally spaced in a Cartesian coordinate system. The Local projection is not conformal.
- The **Mercator** projection is a conformal projection for global maps. Horizontal and vertical lines are straight and always perpendicular to each other. Note that the Mercator projection extends to infinity as it approaches the North and South Poles, so it cannot be used if your map includes the North or South Pole. Distortion is greatest when the map approaches these limits.
- The **Winkel Tripel** projection is a non-conformal projection for global maps. Although it is not conformal, it provides a good balance between shape and size. Except for the Equator and Prime Meridian, all lines are curved. If your global map includes the North or South Pole, this is a good projection choice.
- As its name suggests, the **Lambert Conformal Conic** projection is a conformal projection and is used for maps of continental or smaller land masses that are longer East and West compared to North and South.
- The **Transverse Mercator** is another conformal projection for maps of continental or smaller land masses. Use this projection for land masses that are longer North and South compared to East and West.

Step 4 - Finish

At this point you can add a comment to describe the map file and also create a sample data file from the map keys.

Map keys. If there are multiple keys in the map file, select a map key whose feature labels you want to display in the preview. If you create a data file from the map, these labels will be used for the data values.

Comment. Enter a comment that describes the map or provides additional information that may be relevant to your users, such as the sources for the original shapefiles. The comment will appear in the Graphboard Template Chooser's Manage system.

Create a data set from the feature labels. Check this option if you want to create a data file from the displayed feature labels. When you click **Browse...**, you will be able to specify a location and file name. If you add a *.txt* extension, the file will be saved as a tab-separate values file. If you add a *.csv* extension, the file will be saved as a comma-separated values file. If you add a *.sav* extension, the file will be saved in IBM SPSS Statistics format. SAV is the default when no extension is specified.

Distributing map files

In the first step of the Map Conversion Utility, you chose a location where to save the converted SMZ file. You may have also chosen to add the map to the Manage system for the Graphboard Template Chooser. If you chose to save to the Manage system, the map will be available to you in any IBM SPSS product that you run on the same computer.

To distribute the map to other users, you will need to send the SMZ to them. Those users can then use the Manage system to import the map. You can simply send the file whose location you specified in step 1. If you want to send a file that's in the Manage system, you first need to export it:

1. In the Template Chooser, click **Manage...**
2. Click the Map tab.
3. Select the map you want to distribute.
4. Click **Export...** and choose a location where you want to save the file.

Now you can send the physical map file to other users. Users will need to reverse this process and import the map into the Manage system.

Plot Node

Plot nodes show the relationship between numeric fields. You can create a plot using points (also known as a scatterplot), or you can use lines. You can create three types of line plots by specifying an X Mode in the dialog box.

X Mode = Sort

Setting X Mode to **Sort** causes data to be sorted by values for the field plotted on the *x* axis. This produces a single line running from left to right on the graph. Using a nominal field as an overlay produces multiple lines of different hues running from left to right on the graph.

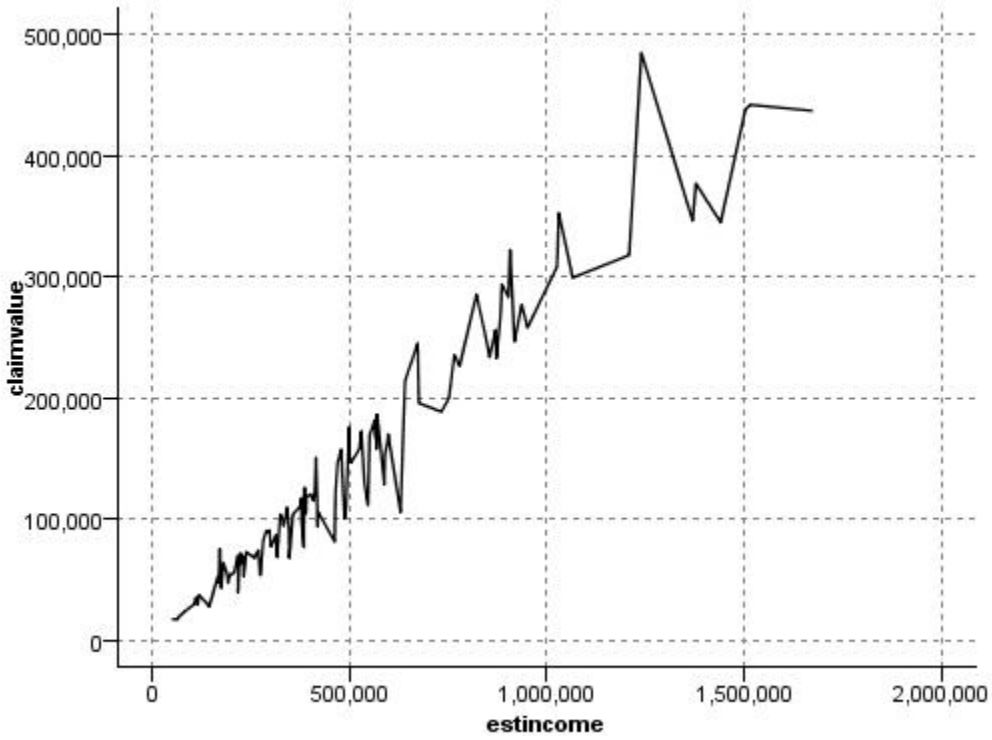


Figure 19. Line plot with X Mode set to Sort

X Mode = Overlay

Setting X Mode to **Overlay** creates multiple line plots on the same graph. Data are not sorted for an overlay plot; as long as the values on the *x* axis increase, data will be plotted on a single line. If the values decrease, a new line begins. For example, as *x* moves from 0 to 100, the *y* values will be plotted on a single line. When *x* falls below 100, a new line will be plotted in addition to the first one. The finished plot might have numerous plots useful for comparing several series of *y* values. This type of plot is useful for data with a periodic time component, such as electricity demand over successive 24-hour periods.

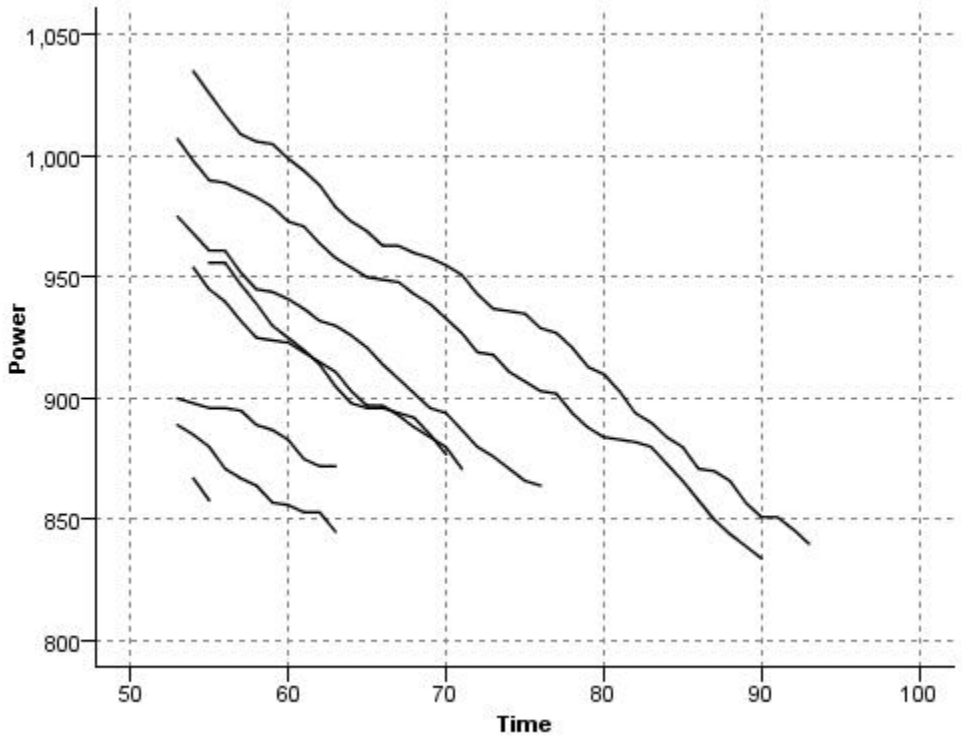


Figure 20. Line plot with X Mode set to Overlay

X Mode = As Read

Setting X Mode to **As Read** plots *x* and *y* values as they are read from the data source. This option is useful for data with a time series component where you are interested in trends or patterns that depend on the order of the data. You may need to sort the data before creating this type of plot. It may also be useful to compare two similar plots with X Mode set to **Sort** and **As Read** in order to determine how much of a pattern depends on the sorting.

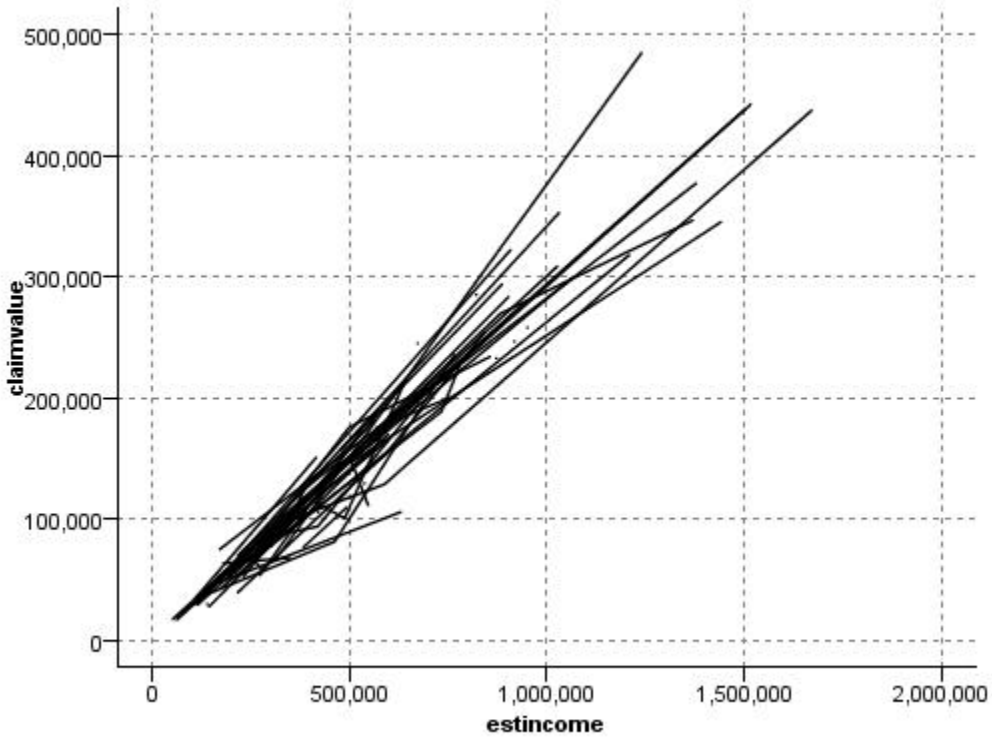


Figure 21. Line plot shown earlier as Sort, executed again with X Mode set to As Read

You can also use the Graphboard node to produce scatterplots and line plots. However, you have more options to choose from in this node. See the topic "Available Built-in Graphboard Visualization Types" on page 176 for more information.

Plot Node Tab

Plots show values of a *Y* field against values of an *X* field. Often, these fields correspond to a dependent variable and an independent variable, respectively.

X field. From the list, select the field to display on the horizontal *x* axis.

Y field. From the list, select the field to display on the vertical *y* axis.

Z field. When you click the 3-D chart button, you can then choose a field from the list to display on the *z* axis.

Overlay. There are several ways to illustrate categories for data values. For example, you can use *maincrop* as a color overlay to indicate the *estincome* and *claimvalue* values for the main crop grown by claim applicants. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Overlay type. Specifies whether an overlay function or smoother is displayed. The smoother and overlay functions are always calculated as a function of y.

- **None.** No overlay is displayed.
- **Smoother.** Displays a smoothed fit line computed using locally weighted iterative robust least squares regression (LOESS). This method effectively computes a series of regressions, each focused on a small area within the plot. This produces a series of "local" regression lines that are then joined to create a smooth curve.

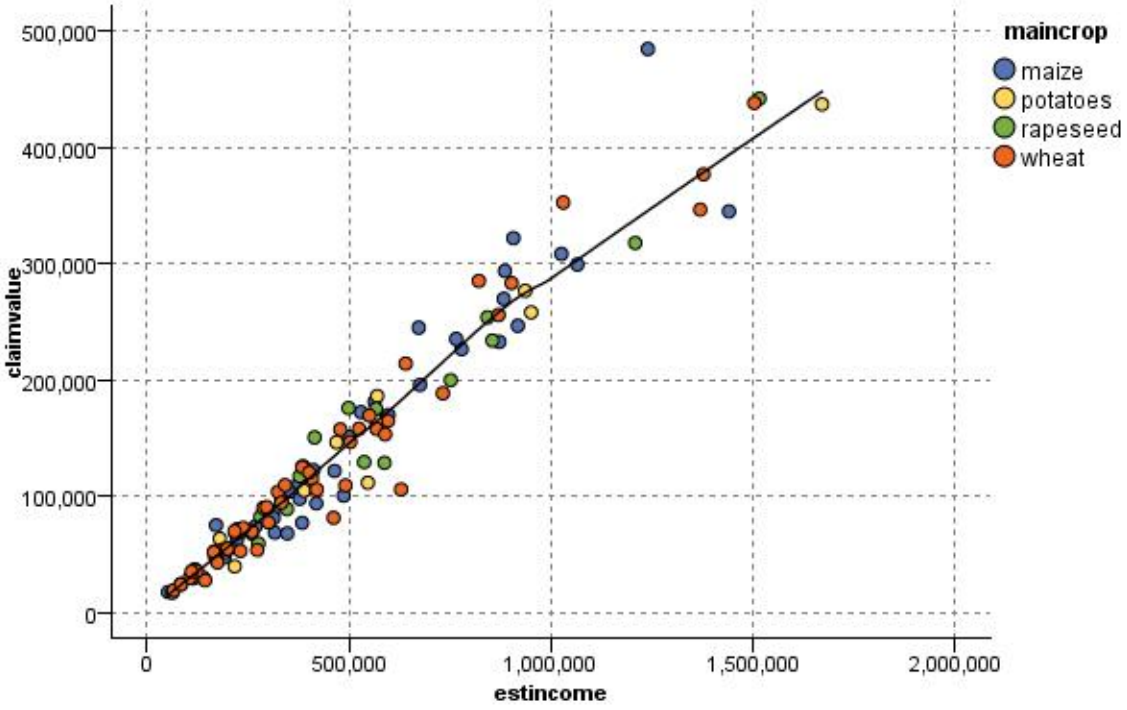


Figure 22. Plot with a LOESS smoother overlay

- **Function.** Select to specify a known function to compare to actual values. For example, to compare actual versus predicted values, you can plot the function y = x as an overlay. Specify a function for **y =** in the text box. The default function is y = x, but you can specify any sort of function, such as a quadratic function or an arbitrary expression, in terms of x.

  Note: Overlay functions are not available for a panel or animation graph.

Once you have set options for a plot, you can run the plot directly from the dialog box by clicking **Run**. You may, however, want to use the Options tab for additional specifications, such as binning, X Mode, and style.

Plot Options Tab

Style. Select either **Point** or **Line** for the plot style. Selecting **Line** activates the **X Mode** control. Selecting **Point** will use a plus symbol (+) as the default point shape. Once the graph is created, you can change the point shape and alter its size.

X Mode. For line plots, you must choose an X Mode to define the style of the line plot. Select **Sort**, **Overlay**, or **As Read**. For **Overlay** or **As Read**, you should specify a maximum dataset size used to sample the first *n* records. Otherwise, the default 2,000 records will be used.

Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified **Min** and **Max** values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Automatic Y range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified **Min** and **Max** values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Automatic Z range. Only when a 3-D graph is specified on the Plot tab. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified **Min** and **Max** values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Jitter. Also known as **agitation**, jitter is useful for point plots of a dataset in which many values are repeated. In order to see a clearer distribution of values, you can use jitter to distribute the points randomly around the actual value.

Note to users of earlier versions of IBM SPSS Modeler : The jitter value used in a plot uses a different metric in this release of IBM SPSS Modeler. In earlier versions, the value was an actual number, but it is now a proportion of the frame size. This means that agitation values in old streams are likely to be too large. For this release, any nonzero agitation values will be converted to the value 0.2.

Maximum number of records to plot. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 records. Performance is enhanced for large datasets when you select the **Bin** or **Sample** options. Alternatively, you can choose to plot all data points by selecting **Use all data**, but you should note that this may dramatically decrease the performance of the software.

Note: When X Mode is set to **Overlay** or **As Read**, these options are disabled and only the first *n* records are used.

- **Bin.** Select to enable binning when the dataset contains more than the specified number of records. Binning divides the graph into fine grids before actually plotting and counts the number of points that would appear in each of the grid cells. In the final graph, one point is plotted per cell at the bin centroid (average of all point locations in the bin). The size of the plotted symbols indicates the number of points in that region (unless you have used size as an overlay). Using the centroid and size to represent the number of points makes the binned plot a superior way to represent large datasets, because it prevents overplotting in dense regions (undifferentiated masses of color) and reduces symbol artifacts (artificial patterns of density). Symbol artifacts occur when certain symbols (particularly the plus symbol [+]) collide in a way that produces dense areas not present in the raw data.
- **Sample.** Select to randomly sample the data to the number of records entered in the text field. The default is 2,000.

Plot Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Z label. Available only for 3-D graphs, either accept the automatically generated *z*-axis label or select **Custom** to specify a custom label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using a Plot Graph

Plots and multiplots are essentially plots of *X* against *Y*. For example, if you are exploring potential fraud in agricultural grant applications, you might want to plot the income claimed on the application versus the income estimated by a neural net. Using an overlay, such as crop type, will illustrate whether there is a relationship between claims (value or number) and type of crop.

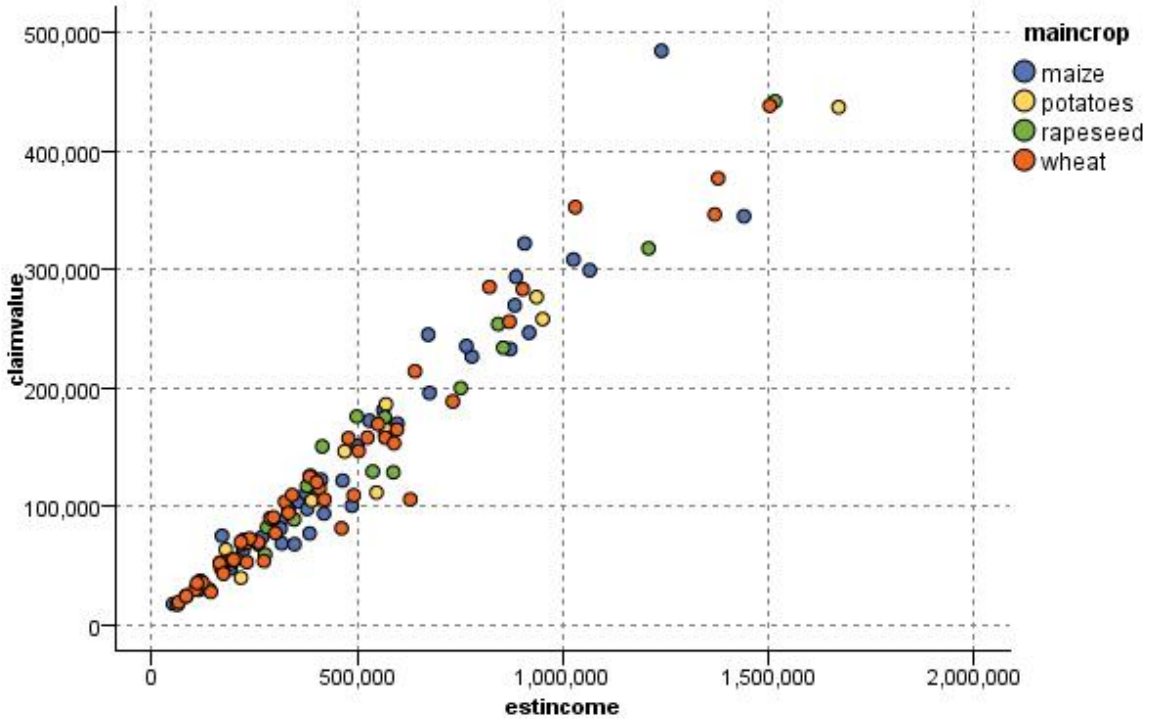


Figure 23. Plot of the relationship between estimated income and claim value with main crop type as an overlay

Since plots, multiplots, and evaluation charts are two-dimensional displays of *Y* against *X*, it is easy to interact with them by defining regions, marking elements, or even drawing bands. You can also generate nodes for the data represented by those regions, bands, or elements. See the topic "Exploring Graphs" on page 245 for more information.

Multiplot Node

A multiplot is a special type of plot that displays multiple *Y* fields over a single *X* field. The *Y* fields are plotted as colored lines and each is equivalent to a Plot node with Style set to **Line** and X Mode set to **Sort**. Multiplots are useful when you have time sequence data and want to explore the fluctuation of several variables over time.

Multiplot Plot Tab

X field. From the list, select the field to display on the horizontal *x* axis.

Y fields. Select one or more fields from the list to display over the range of *X* field values. Use the Field Chooser button to select multiple fields. Click the delete button to remove fields from the list.

Overlay. There are several ways to illustrate categories for data values. For example, you might use an animation overlay to display multiple plots for each value in the data. This is useful for sets containing upwards of 10 categories. When used for sets with more than 15 categories, you may notice a decrease in performance. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Normalize. Select to scale all *Y* values to the range 0–1 for display on the graph. Normalizing helps you explore the relationship between lines that might otherwise be obscured due to differences in the range of values for each series and is recommended when plotting multiple lines on the same graph, or when comparing plots in side-by-side panels. (Normalizing is not necessary when all data values fall within a similar range.)

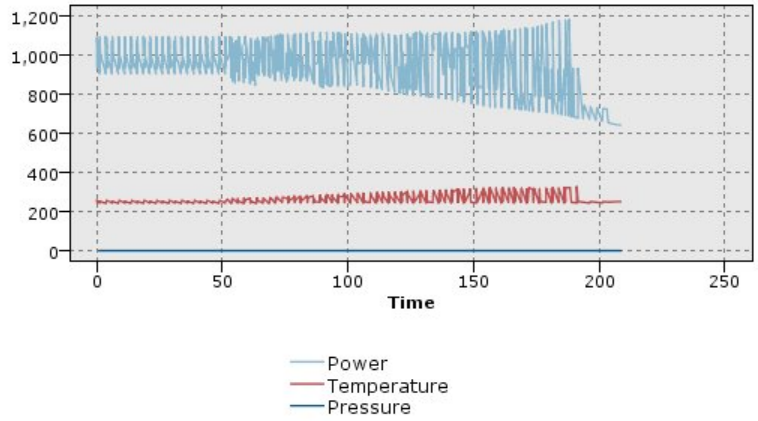


Figure 24. Standard multiplot showing power-plant fluctuation over time (note that without normalizing, the plot for Pressure is impossible to see)

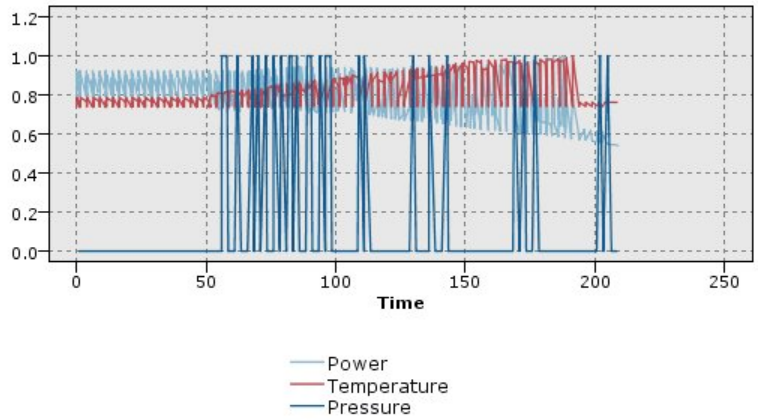


Figure 25. Normalized multiplot showing a plot for Pressure

Overlay function. Select to specify a known function to compare to actual values. For example, to compare actual versus predicted values, you can plot the function y = x as an overlay. Specify a function for **y =** in the text box. The default function is y = x, but you can specify any sort of function, such as a quadratic function or an arbitrary expression, in terms of x.

Note: Overlay functions are not available for a panel or animation graph.

When number of records greater than. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 points. Performance is enhanced for large datasets when you select the **Bin** or **Sample** options. Alternatively, you can choose to plot all data points by selecting **Use all data**, but you should note that this may dramatically decrease the performance of the software.

Note: When X Mode is set to **Overlay** or **As Read**, these options are disabled and only the first *n* records are used.

- **Bin.** Select to enable binning when the dataset contains more than the specified number of records. Binning divides the graph into fine grids before actually plotting and counts the number of connections that would appear in each of the grid cells. In the final graph, one connection is used per cell at the bin centroid (average of all connection points in the bin).
- **Sample.** Select to randomly sample the data to the specified number of records.

Multiplot Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using a Multiplot Graph

Plots and multiplots are essentially plots of *X* against *Y*. For example, if you are exploring potential fraud in agricultural grant applications, you might want to plot the income claimed on the application versus the income estimated by a neural net. Using an overlay, such as crop type, will illustrate whether there is a relationship between claims (value or number) and type of crop.

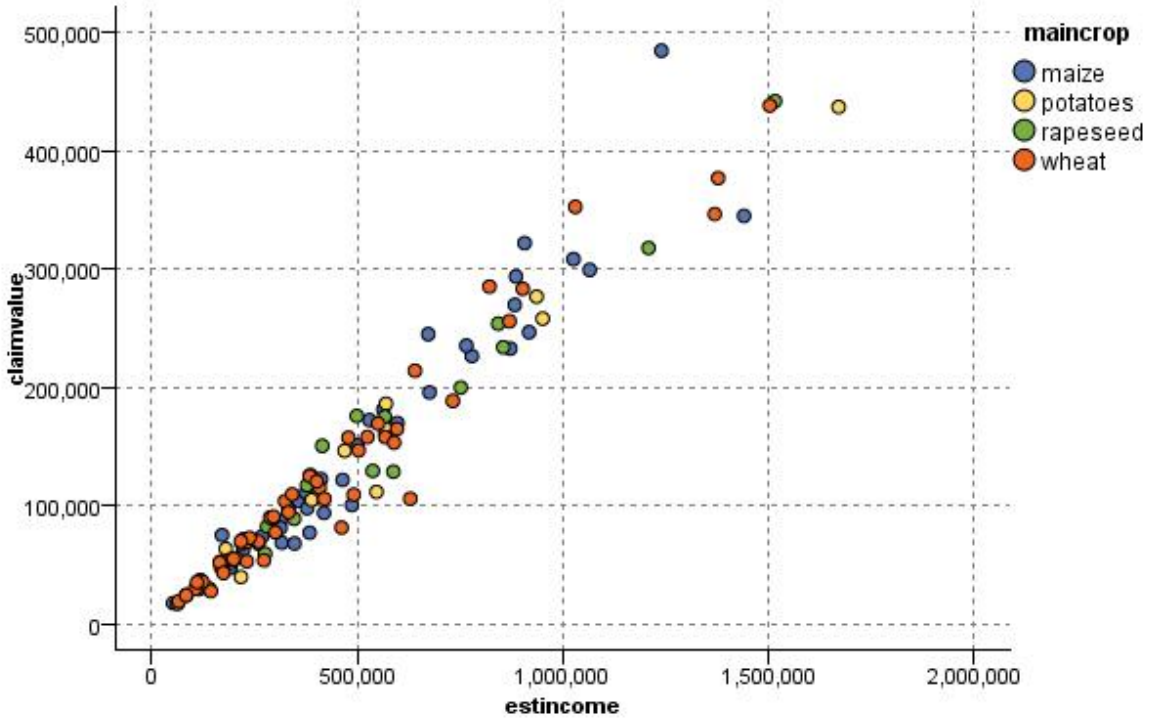


Figure 26. Plot of the relationship between estimated income and claim value with main crop type as an overlay

Since plots, multiplots, and evaluation charts are two-dimensional displays of *Y* against *X*, it is easy to interact with them by defining regions, marking elements, or even drawing bands. You can also generate nodes for the data represented by those regions, bands, or elements. See the topic "Exploring Graphs" on page 245 for more information.

Time Plot Node

Time Plot nodes enable you to view one or more time series plotted over time. The series you plot must contain numeric values and are assumed to occur over a range of time in which the periods are uniform.

In SPSS Modeler versions 17.1 and earlier, you usually use a Time Intervals node before a Time Plot node to create a *TimeLabel* field, which is used by default to label the *x* axis in the graphs.

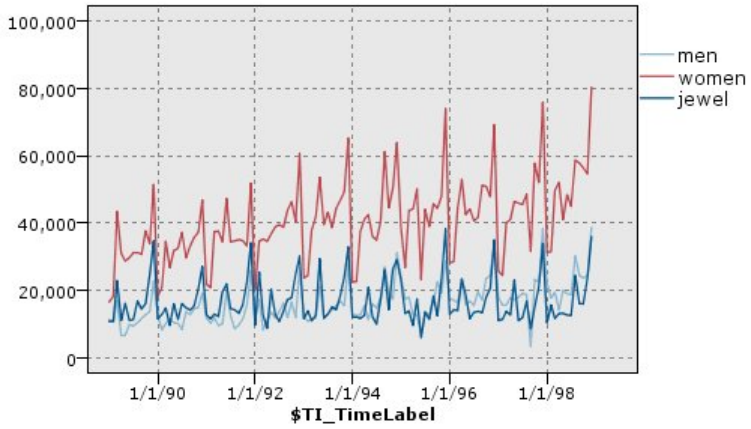


Figure 27. Plotting sales of men's and women's clothing and jewelry over time

Creating Interventions and Events

You can create Event and Intervention fields from the time plot by generating a derive (flag or nominal) node from the context menus. For example, you could create an event field in the case of a rail strike, where the drive state is True if the event happened and False otherwise. For an Intervention field, for a price rise for example, you could use a derive count to identify the date of the rise, with 0 for the old price and 1 for the new price. See the topic "Derive node" on page 136 for more information.

Time Plot Tab

Plot. Provides a choice of how to plot time series data.

- **Selected series.** Plots values for selected time series. If you select this option when plotting confidence intervals, deselect the **Normalize** check box.
- **Selected Time Series models**. Used in conjunction with a Time Series model, this option plots all the related fields (actual and predicted values, as well as confidence intervals) for one or more selected time series. This option disables some other options in the dialog box. This is the preferred option if plotting confidence intervals.

Series. Select one or more fields with time series data you want to plot. The data must be numeric.

X axis label. Choose either the default label or a single field to use as the label for the *x* axis in plots. If you choose Default, the system uses the TimeLabel field created from a Time Intervals node upstream (for streams created in SPSS Modeler versions 17.1 and earlier), or sequential integers if there is no upstream Time Intervals node.

Display series in separate panels. Specifies whether each series is displayed in a separate panel. Alternatively, if you do not choose to panel, all time series are plotted on the same graph, and smoothers will not be available. When plotting all time series on the same graph, each series will be represented by a different color.

Normalize. Select to scale all *Y* values to the range 0–1 for display on the graph. Normalizing helps you explore the relationship between lines that might otherwise be obscured due to differences in the range of values for each series and is recommended when plotting multiple lines on the same graph, or when

comparing plots in side-by-side panels. (Normalizing is not necessary when all data values fall within a similar range.)

Display. Select one or more elements to display in your plot. You can choose from lines, points, and (LOESS) smoothers. Smoothers are available only if you display the series in separate panels. By default, the line element is selected. Make sure you select at least one plot element before you run the graph node; otherwise, the system will return an error stating that you have selected nothing to plot.

Limit records. Select this option if you want to limit the number of records plotted. Specify the number of records, read from the beginning of your data file, that will be plotted in the **Maximum number of records to plot** option. By default this number is set to 2,000. If you want to plot the last *n* records in your data file, you can use a Sort node prior to this node to arrange the records in descending order by time.

Time Plot Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Layout. For time plots only, you can specify whether time values are plotted along a horizontal or vertical axis.

Using a Time Plot Graph

Once you have created a time plot graph, there are several options for adjusting the graph display and generating nodes for further analysis. See the topic "Exploring Graphs" on page 245 for more information.

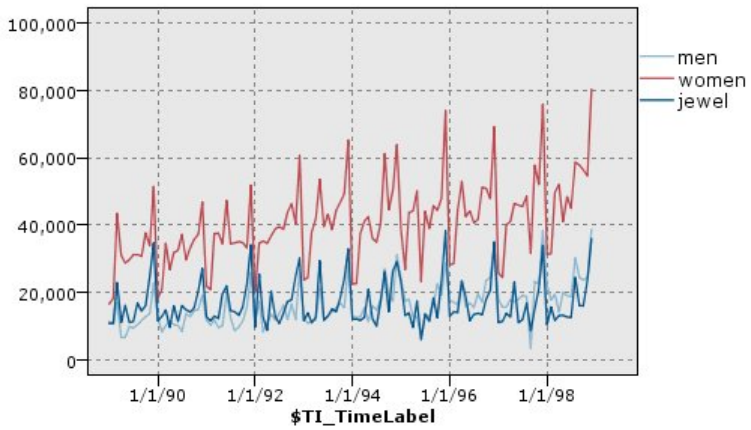


Figure 28. Plotting sales of men's and women's clothing and jewelry over time

Once you have created a time plot, defined bands, and examined the results, you can use options on the Generate menu and the context menu to create Select or Derive nodes. See the topic "Generating Nodes from Graphs" on page 252 for more information.

Distribution Node

A distribution graph or table shows the occurrence of symbolic (non-numeric) values, such as mortgage type or gender, in a dataset. A typical use of the Distribution node is to show imbalances in the data that can be rectified by using a Balance node before creating a model. You can automatically generate a Balance node using the Generate menu in the distribution graph or table window.

You can also use the Graphboard node to produce bar of counts graphs. However, you have more options to choose from in this node. See the topic "Available Built-in Graphboard Visualization Types" on page 176 for more information.

Note: To show the occurrence of numeric values, you should use a Histogram node.

Distribution Plot Tab

Plot. Select the type of distribution. Select **Selected fields** to show the distribution of the selected field. Select **All flags (true values)** to show the distribution of true values for flag fields in the dataset.

Field. Select a nominal or flag field for which to show the distribution of values. Only fields that have not been explicitly set as numeric appear on the list.

Overlay. Select a nominal or flag field to use as a color overlay, illustrating the distribution of its values within each value of the specified field. For example, you can use marketing campaign response (*pep*) as an overlay for number of children (*children*) to illustrate responsiveness by family size. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Normalize by color. Select to scale bars so that all bars take up the full width of the graph. The overlay values equal a proportion of each bar, making comparisons across categories easier.

Sort. Select the method used to display values in the distribution graph. Select **Alphabetic** to use alphabetical order or **By count** to list values in decreasing order of occurrence.

Proportional scale. Select to scale the distribution of values so that the value with the largest count fills the full width of the plot. All other bars are scaled against this value. Deselecting this option scales bars according to the total counts of each value.

Distribution Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using a Distribution Node

Distribution nodes are used to show the distribution of symbolic values in a dataset. They are frequently used before manipulation nodes to explore the data and correct any imbalances. For example, if instances of respondents without children occur much more frequently than other types of respondents, you might want to reduce these instances so that a more useful rule can be generated in later data mining operations. A Distribution node will help you to examine and make decisions about such imbalances.

The Distribution node is unusual in that it produces both a graph and a table to analyze your data.

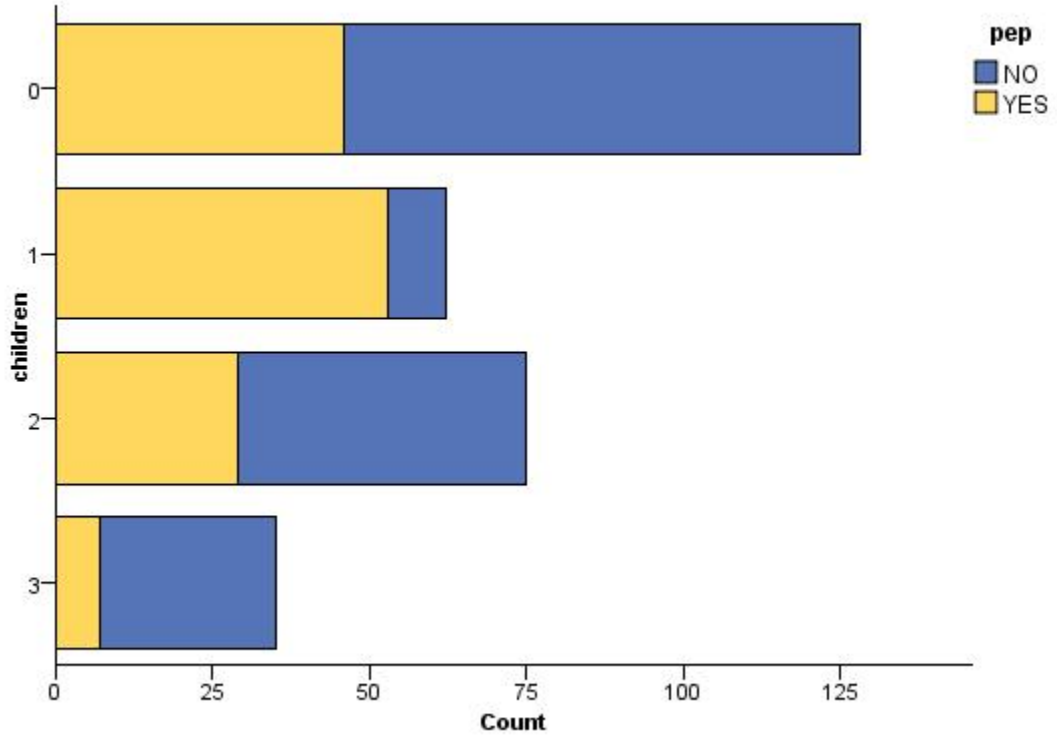


Figure 29. Distribution graph showing the number of people with or without children who responded to a marketing campaign

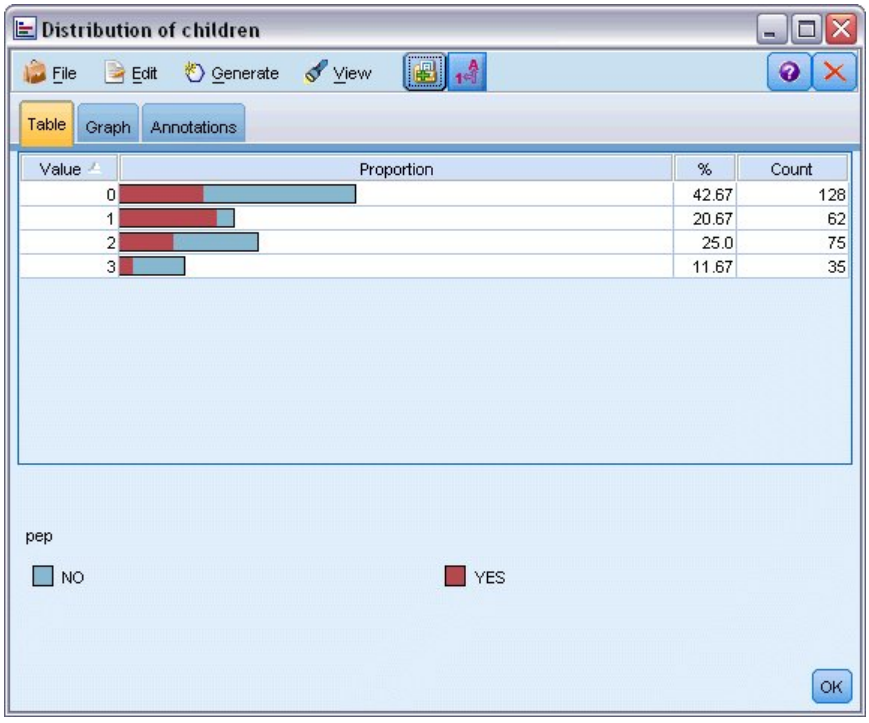


Figure 30. Distribution table showing the proportion of people with or without children who responded to a marketing campaign

Once you have created a distribution table and graph and examined the results, you can use options from the menus to group values, copy values, and generate a number of nodes for data preparation. In

addition, you can copy or export the graph and table information for use in other applications, such as MS Word or MS PowerPoint. See the topic "Printing, saving, copying, and exporting graphs" on page 266 for more information.

To Select and Copy Values from a Distribution Table

1. Click and hold the mouse button while dragging it over the rows to select a set of values. You can also use the Edit menu to **Select All** values.
2. From the Edit menu, choose **Copy Table** or **Copy Table (inc. field names)**.
3. Paste to the clipboard or into the desired application.

   Note: The bars do not get copied directly. Instead, the table values are copied. This means that overlaid values will not be displayed in the copied table.

To Group Values from a Distribution Table

1. Select values for grouping using the Ctrl+click method.
2. From the Edit menu, choose **Group**.

Note: When you group and ungroup values, the graph on the Graph tab is automatically redrawn to show the changes.

You can also:

- Ungroup values by selecting the group name in the distribution list and choosing **Ungroup** from the Edit menu.
- Edit groups by selecting the group name in the distribution list and choosing **Edit group** from the Edit menu. This opens a dialog box where values can be shifted to and from the group.

Generate Menu Options

You can use options on the Generate menu to select a subset of data, derive a flag field, regroup values, reclassify values, or balance the data from either a graph or table. These operations generate a data preparation node and place it on the stream canvas. To use the generated node, connect it to an existing stream. See the topic "Generating Nodes from Graphs" on page 252 for more information.

Histogram Node

Histogram nodes show the occurrence of values for numeric fields. They are often used to explore the data before manipulations and model building. Similar to the Distribution node, Histogram nodes are frequently used to reveal imbalances in the data. While you can also use the Graphboard node to produce a histogram, you have more options to choose from in this node. See the topic "Available Built-in Graphboard Visualization Types" on page 176 for more information.

Note: To show the occurrence of values for symbolic fields, you should use a Distribution node.

Histogram Plot Tab

Field. Select a numeric field for which to show the distribution of values. Only fields that have not been explicitly defined as symbolic (categorical) will be listed.

Overlay. Select a symbolic field to show categories of values for the specified field. Selecting an overlay field converts the histogram to a stacked chart with colors used to represent different categories of the overlay field. Using the Histogram node, there are three types of overlays: color, panel, and animation. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Histogram Options Tab

Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified **Min** and **Max** values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Bins. Select either **By number** or **By width**.

- Select **By number** to display a fixed number of bars whose width depends on the range and the number of bins specified. Indicate the number of bins to be used in the graph in the **No. of bins** option. Use the arrows to adjust the number.
- Select **By width** to create a graph with bars of a fixed width. The number of bins depends on the specified width and the range of values. Indicate the width of the bars in the **Bin width** option.

Normalize by color. Select to adjust all bars to the same height, displaying overlaid values as a percentage of the total cases in each bar.

Show normal curve. Select to add a normal curve to the graph showing the mean and variance of the data.

Separate bands for each color. Select to display each overlaid value as a separate band on the graph.

Histogram Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Using Histograms

Histograms show the distribution of values in a numeric field whose values range along the *x* axis. Histograms operate similarly to collections graphs. Collections show the distribution of values for one numeric field *relative to the values of another*, rather than the occurrence of values for a single field.

Once you have created a graph, you can examine the results and define bands to split values along the *x* axis or define regions. You can also mark elements within the graph. See the topic "Exploring Graphs" on page 245 for more information.

You can use options on the Generate menu to create Balance, Select, or Derive nodes using the data in the graph or more specifically within bands, regions, or marked elements. This type of graph is frequently used before manipulation nodes to explore the data and correct any imbalances by generating a Balance node from the graph to use in the stream. You can also generate a Derive Flag node to add a field showing which band each record falls into or a Select node to select all records within a particular set or range of values. Such operations help you to focus on a particular subset of data for further exploration. See the topic "Generating Nodes from Graphs" on page 252 for more information.

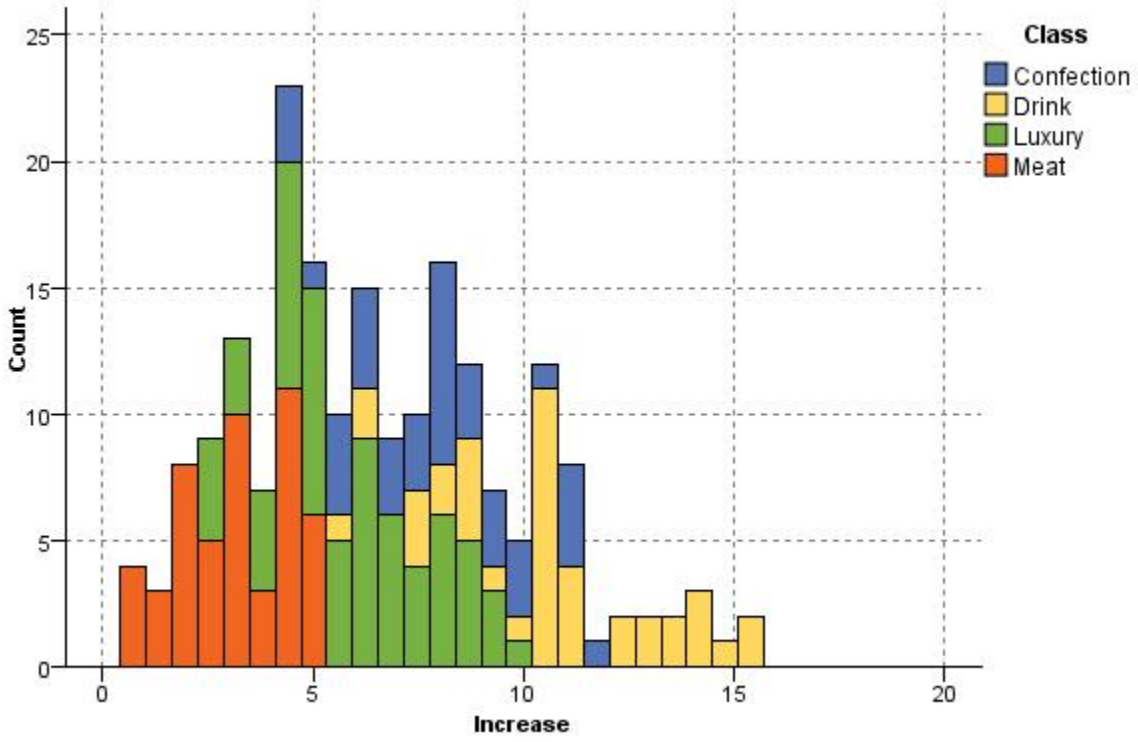


Figure 31. Histogram showing the distribution of increased purchases by category due to promotion

Collection Node

Collections are similar to histograms except that collections show the distribution of values for one numeric field relative to the values of another, rather than the occurrence of values for a single field. A collection is useful for illustrating a variable or field whose values change over time. Using 3-D graphing, you can also include a symbolic axis displaying distributions by category. Two dimensional Collections are shown as stacked bar charts, with overlays where used. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Collection Plot Tab

Collect. Select a field whose values will be collected and displayed over the range of values for the field specified in **Over**. Only fields that have not been defined as symbolic are listed.

Over. Select a field whose values will be used to display the field specified in **Collect**.

By. Enabled when creating a 3-D graph, this option enables you to select a nominal or flag field used to display the collection field by categories.

Operation. Select what each bar in the collection graph represents. Options include **Sum**, **Mean**, **Max**, **Min**, and **Standard Deviation**.

Overlay. Select a symbolic field to show categories of values for the selected field. Selecting an overlay field converts the collection and creates multiple bars of varying colors for each category. This node has three types of overlays: color, panel, and animation. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Collection Options Tab

Automatic X range. Select to use the entire range of values in the data along this axis. Deselect to use an explicit subset of values based on your specified **Min** and **Max** values. Either enter values or use the arrows. Automatic ranges are selected by default to enable rapid graph building.

Bins. Select either **By number** or **By width**.

- Select **By number** to display a fixed number of bars whose width depends on the range and the number of bins specified. Indicate the number of bins to be used in the graph in the **No. of bins** option. Use the arrows to adjust the number.
- Select **By width** to create a graph with bars of a fixed width. The number of bins depends on the specified width and the range of values. Indicate the width of the bars in the **Bin width** option.

Collection Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

Over label. Either accept the automatically generated label, or select **Custom** to specify a label.

Collect label. Either accept the automatically generated label, or select **Custom** to specify a label.

By label. Either accept the automatically generated label, or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

The following example shows where the appearance options are placed on a 3-D version of the graph.

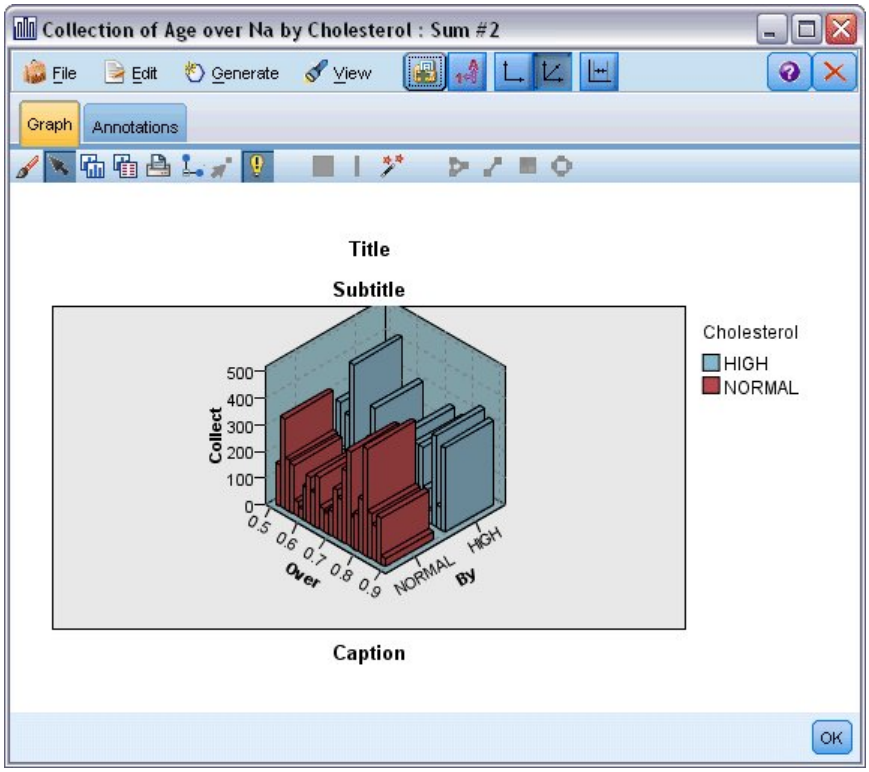


Figure 32. Position of graph appearance options on 3-D Collection graph

Using a Collection Graph

Collections show the distribution of values for one numeric field *relative to the values of another*, rather than the occurrence of values for a single field. Histograms operate similarly to collections graphs. Histograms show the distribution of values in a numeric field whose values range along the *x* axis.

Once you have created a graph, you can examine the results and define bands to split values along the *x* axis or define regions. You can also mark elements within the graph. See the topic "Exploring Graphs" on page 245 for more information.

You can use options on the Generate menu to create Balance, Select, or Derive nodes using the data in the graph or more specifically within bands, regions, or marked elements. This type of graph is frequently used before manipulation nodes to explore the data and correct any imbalances by generating a Balance node from the graph to use in the stream. You can also generate a Derive Flag node to add a field showing which band each record falls into or a Select node to select all records within a particular set or range of values. Such operations help you to focus on a particular subset of data for further exploration. See the topic "Generating Nodes from Graphs" on page 252 for more information.

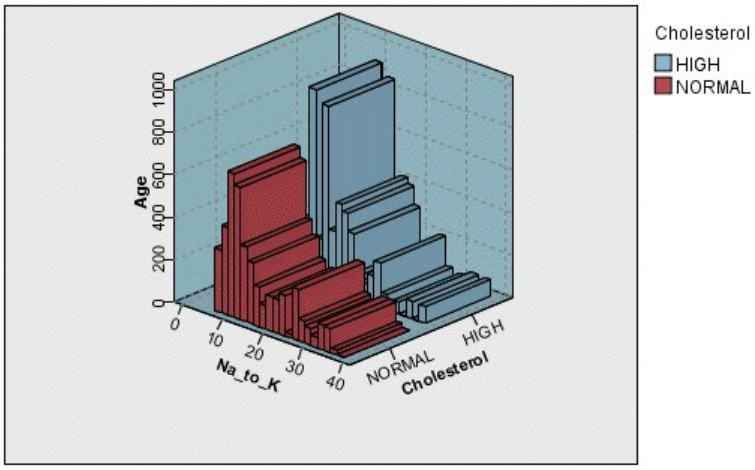


Figure 33. 3-D collection graph showing sum of Na_to_K over Age for both high and normal cholesterol levels

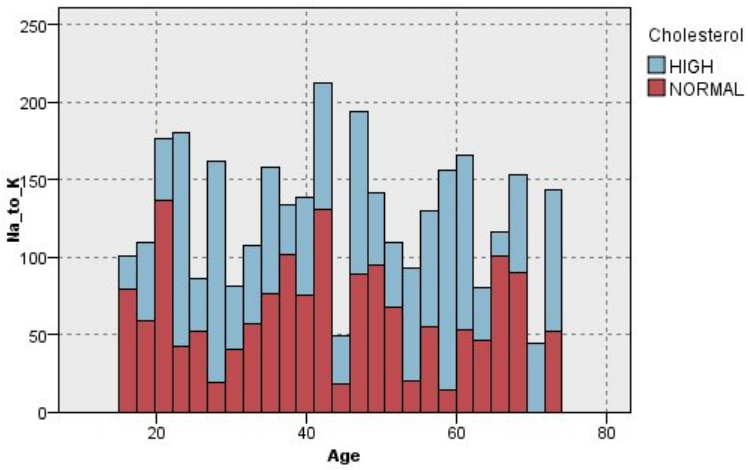


Figure 34. Collection graph without z axis displayed but with Cholesterol as color overlay

Web Node

Web nodes show the strength of relationships between values of two or more symbolic fields. The graph displays connections using varying types of lines to indicate connection strength. You can use a Web

node, for example, to explore the relationship between the purchase of various items at an e-commerce site or a traditional retail outlet.

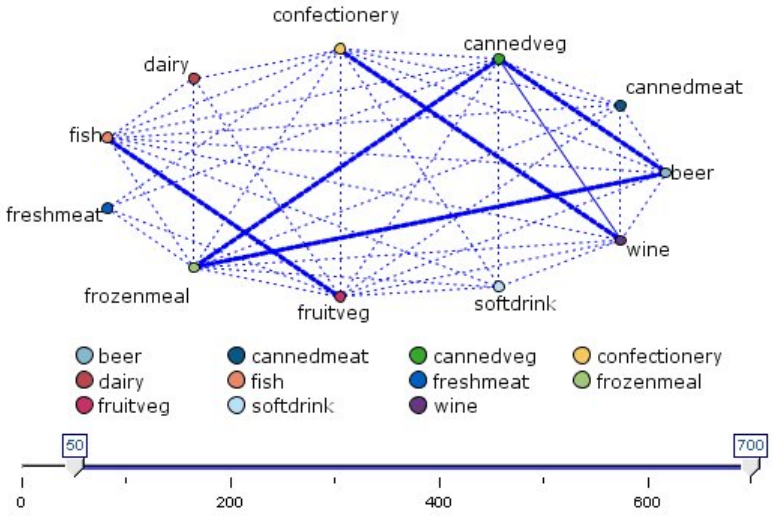


Figure 35. Web graph showing relationships between the purchase of grocery items

Directed Webs

Directed Web nodes are similar to Web nodes in that they show the strength of relationships between symbolic fields. However, directed web graphs show connections only from one or more From fields to a single To field. The connections are unidirectional in the sense that they are one-way connections.

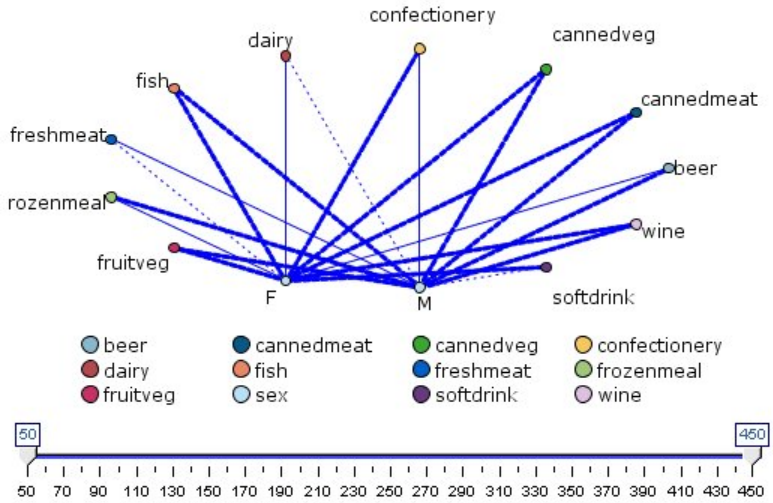


Figure 36. Directed web graph showing the relationship between the purchase of grocery items and gender

Like Web nodes, the graph displays connections using varying types of lines to indicate connection strength. You can use a Directed Web node, for example, to explore the relationship between gender and a proclivity for certain purchase items.

Web Plot Tab

Web. Select to create a web graph illustrating the strength of relationships between all specified fields.

Directed web. Select to create a directional web graph illustrating the strength of relationships between multiple fields and the values of one field, such as gender or religion. When this option is selected, a To Field is activated and the Fields control below is renamed From Fields for additional clarity.

To Field (directed web only). Select a flag or nominal field used for a directed web. Only fields that have not been explicitly set as numeric are listed.

Fields/From Fields. Select fields to create a web graph. Only fields that have not been explicitly set as numeric are listed. Use the Field Chooser button to select multiple fields or select fields by type.

Note: For a directed web, this control is used to select From fields.

Show true flags only. Select to display only true flags for a flag field. This option simplifies the web display and is often used for data where the occurrence of positive values is of special importance.

Line values are. Select a threshold type from the drop-down list.

- **Absolute** sets thresholds based on the number of records having each pair of values.
- **Overall percentages** shows the absolute number of cases represented by the link as a proportion of all of the occurrences of each pair of values represented in the web graph.
- **Percentages of smaller field/value** and **Percentages of larger field/value** indicate which field/value to use for evaluating percentages. For example, suppose 100 records have the value *drugY* for the field *Drug* and only 10 have the value *LOW* for the field *BP*. If seven records have both values *drugY* and *LOW*, this percentage is either 70% or 7%, depending on which field you are referencing, smaller (*BP*) or larger (*Drug*).

Note: For directed web graphs, the third and fourth options above are not available. Instead, you can select **Percentage of "To" field/value** and **Percentage of "From" field/value**.

Strong links are heavier. Selected by default, this is the standard way of viewing links between fields.

Weak links are heavier. Select to reverse the meaning of links displayed in bold lines. This option is frequently used for fraud detection or examination of outliers.

Web Options Tab

The Options tab for Web nodes contains a number of additional options to customize the output graph.

Number of Links. The following options are used to control the number of links displayed in the output graph. Some of these options, such as **Weak links above** and **Strong links above**, are also available in the output graph window. You can also use a slider control in the final graph to adjust the number of links displayed.

- **Maximum number of links to display.** Specify a number indicating the maximum number of links to show on the output graph. Use the arrows to adjust the value.
- **Show only links above.** Specify a number indicating the minimum value for which to show a connection in the web. Use the arrows to adjust the value.
- **Show all links.** Specify to display all links regardless of minimum or maximum values. Selecting this option may increase processing time if there are a large number of fields.

Discard if very few records. Select to ignore connections that are supported by too few records. Set the threshold for this option by entering a number in **Min. records/line**.

Discard if very many records. Select to ignore strongly supported connections. Enter a number in **Max. records/line**.

Weak links below. Specify a number indicating the threshold for weak connections (dotted lines) and regular connections (normal lines). All connections below this value are considered weak.

Strong links above. Specify a threshold for strong connections (heavy lines) and regular connections (normal lines). All connections above this value are considered strong.

Link Size. Specify options for controlling the size of links:

- **Link size varies continuously.** Select to display a range of link sizes reflecting the variation in connection strengths based on actual data values.
- **Link size shows strong/normal/weak categories.** Select to display three strengths of connections--strong, normal, and weak. The cutoff points for these categories can be specified above as well as in the final graph.

Web Display. Select a type of web display:

- **Circle layout.** Select to use the standard web display.
- **Network layout.** Select to use an algorithm to group together the strongest links. This is intended to highlight strong links using spatial differentiation as well as weighted lines.
- **Directed Layout.** Select to create a directed web display that uses the **To Field** selection from the Plot tab as the focus for the direction.
- **Grid Layout.** Select to create a web display that is laid out in a regularly spaced grid pattern.

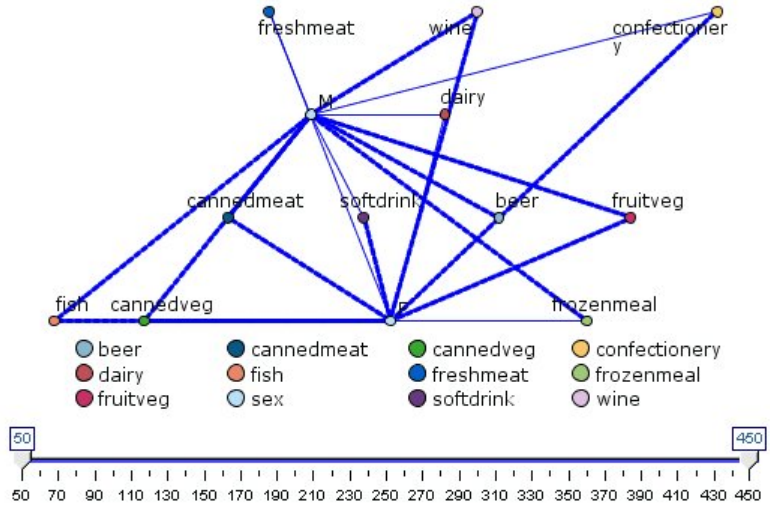


Figure 37. Web graph showing strong connections from frozenmeal and cannedveg to other grocery items

Note: When filtering the displayed links (using either the slider in the web graph or the **Show only links above** control on the Web node's Options tab), you may end up in a situation where all the links that remain to be displayed are of a single value (in other words, they're either all weak links, all medium links, or all strong links, as defined by the **Weak links below** and **Strong links above** controls on the Web node's Options tab). If this happens, all the links are all displayed with the medium width line in the web graph output.

Web Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

Show legend. You can specify whether the legend is displayed. For plots with a large number of fields, hiding the legend may improve the appearance of the plot.

Use labels as nodes. You can include the label text within each node rather than displaying adjacent labels. For plots with a small number of fields, this may result in a more readable chart.

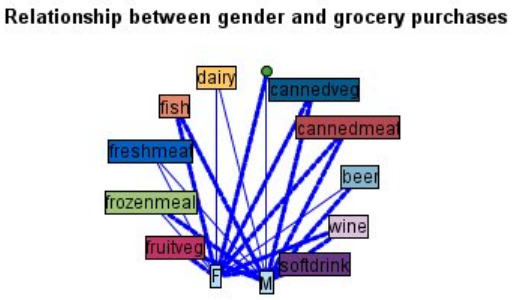


Figure 38. Web graph showing labels as nodes

Using a Web Graph

Web nodes are used to show the strength of relationships between values of two or more symbolic fields. Connections are displayed in a graph with varying types of lines to indicate connections of increasing strength. You can use a Web node, for example, to explore the relationship between cholesterol levels, blood pressure, and the drug that was effective in treating the patient's illness.

- Strong connections are shown with a heavy line. This indicates that the two values are strongly related and should be further explored.
- Medium connections are shown with a line of normal weight.
- Weak connections are shown with a dotted line.
- If no line is shown between two values, this means either that the two values never occur in the same record or that this combination occurs in a number of records below the threshold specified in the Web node dialog box.

Once you have created a Web node, there are several options for adjusting the graph display and generating nodes for further analysis.

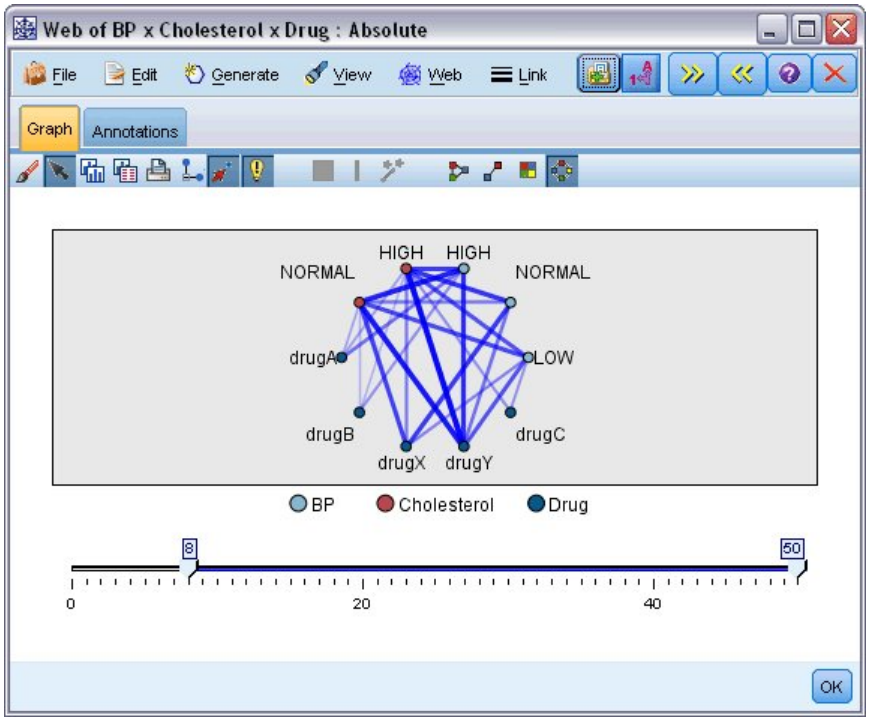


Figure 39. Web graph indicating a number of strong relationships, such as normal blood pressure with DrugX and high cholesterol with DrugY

For both Web nodes and Directed Web nodes, you can:

- Change the layout of the web display.
- Hide points to simplify the display.
- Change the thresholds controlling line styles.
- Highlight lines between values to indicate a "selected" relationship.
- Generate a Select node for one or more "selected" records or a Derive Flag node associated with one or more relationships in the web.

To adjust points

- **Move** points by clicking the mouse on a point and dragging it to the new location. The web will be redrawn to reflect the new location.
- **Hide** points by right-clicking on a point in the web and choosing **Hide** or **Hide and Replan** from the context menu. **Hide** simply hides the selected point and any lines associated with it. **Hide and Replan** redraws the web, adjusting for any changes you have made. Any manual moves are undone.
- **Show** all hidden points by choosing **Reveal All** or **Reveal All and Replan** from the Web menu in the graph window. Selecting **Reveal All and Replan** redraws the web, adjusting to include all previously hidden points and their connections.

To select, or "highlight," lines

Selected lines are highlighted in red.

1. To select a single line, left-click the line.
2. To select multiple lines, do one of the following:

- Using the cursor, draw a circle around the points whose lines you want to select.
- Hold down the Ctrl key and left-click the individual lines you want to select.

You can deselect all selected lines by clicking the graph background, or by choosing **Clear Selection** from the Web menu in the graph window.

To view the web using a different layout

From the Web menu, choose **Circle Layout**, **Network Layout**, **Directed Layout**, or **Grid Layout** to change the layout of the graph.

Toturn the links slider on or off

From the View menu, choose **Links Slider**.

To select or flag records for a single relationship

1. Right-click on the line representing the relationship of interest.
2. From the context menu, choose **Generate Select Node For Link** or **Generate Derive Node For Link**.

A Select node or Derive node is automatically added to the stream canvas with the appropriate options and conditions specified:

- The Select node selects all records in the given relationship.
- The Derive node generates a flag indicating whether the selected relationship holds true for records in the entire dataset. The flag field is named by joining the two values in the relationship with an underscore, such as `LOW_drugC` or `drugC_LOW`.

To select of flag records for a group of relationships

1. Select the line(s) in the web display representing relationships of interest.

2. From the Generate menu in the graph window, choose **Select Node ("And")**, **Select Node ("Or")**, **Derive Node ("And")**, or **Derive Node ("Or")**.

- The "Or" nodes give the disjunction of conditions. This means that the node will apply to records for which any of the selected relationships hold.
- The "And" nodes give the conjunction of conditions. This means that the node will apply only to records for which all selected relationships hold. An error occurs if any of the selected relationships are mutually exclusive.

After you have completed your selection, a Select node or Derive node is automatically added to the stream canvas with the appropriate options and conditions specified.

Note: When filtering the displayed links (using either the slider in the web graph or the **Show only links above** control on the Web node's Options tab), you may end up in a situation where all the links that remain to be displayed are of a single value (in other words, they're either all weak links, all medium links, or all strong links, as defined by the **Weak links below** and **Strong links above** controls on the Web node's Options tab). If this happens, all the links are all displayed with the medium width line in the web graph output.

Adjusting Web Thresholds

After you have created a web graph, you can adjust the thresholds controlling line styles using the toolbar slider to change the minimum visible line. You can also view additional threshold options by clicking the yellow double-arrow button on the toolbar to expand the web graph window. Then click the **Controls** tab to view additional options.

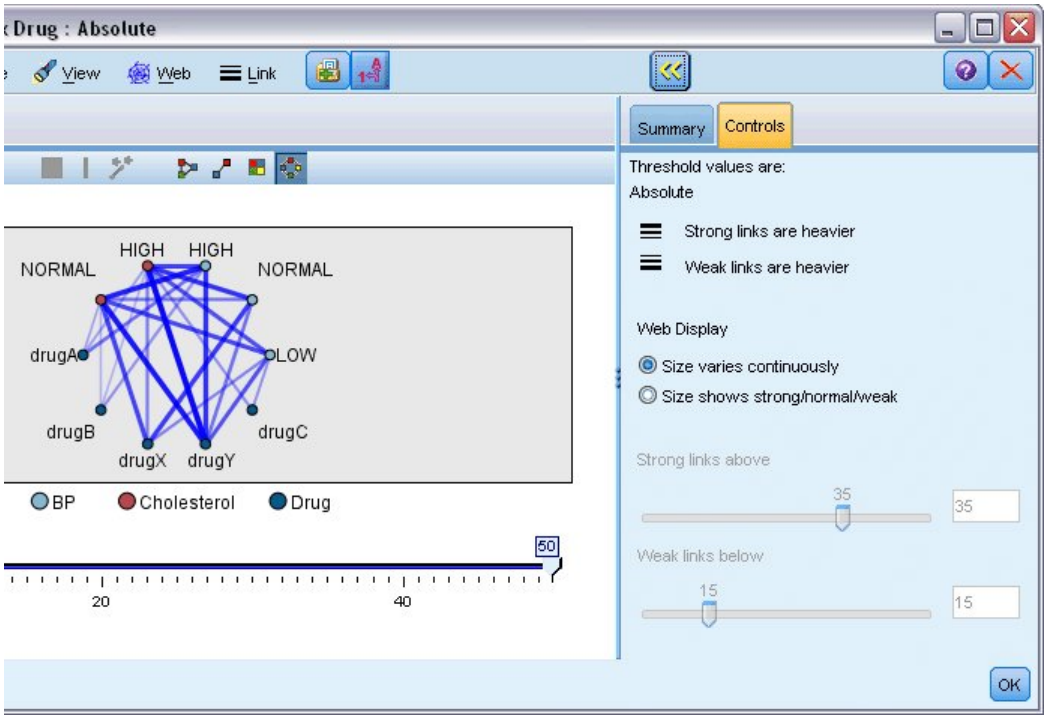


Figure 40. Expanded window featuring display and threshold options

Threshold values are. Shows the type of threshold selected during creation in the Web node dialog box.

Strong links are heavier. Selected by default, this is the standard way of viewing links between fields.

Weak links are heavier. Select to reverse the meaning of links displayed in bold lines. This option is frequently used for fraud detection or examination of outliers.

Web Display. Specify options for controlling the size of links in the output graph:

- **Size varies continuously.** Select to display a range of link sizes reflecting the variation in connection strengths based on actual data values.

- **Size shows strong/normal/weak.** Select to display three strengths of connections--strong, normal, and weak. The cutoff points for these categories can be specified above as well as in the final graph.

Strong links above. Specify a threshold for strong connections (heavy lines) and regular connections (normal lines). All connections above this value are considered strong. Use the slider to adjust the value or enter a number in the field.

Weak links below. Specify a number indicating the threshold for weak connections (dotted lines) and regular connections (normal lines). All connections below this value are considered weak. Use the slider to adjust the value or enter a number in the field.

After you have adjusted the thresholds for a web, you can replan, or redraw, the web display with the new threshold values through the web menu located on the web graph toolbar. Once you have found settings that reveal the most meaningful patterns, you can update the original settings in the Web node (also called the Parent Web node) by choosing **Update Parent Node** from the Web menu in the graph window.

Creating a web summary

You can create a web summary document that lists strong, medium, and weak links by clicking the yellow double-arrow button on the toolbar to expand the web graph window. Then click the **Summary** tab to view tables for each type of link. Tables can be expanded and collapsed using the toggle buttons for each.

To print the summary, choose the following from the menu in the web graph window:

File > **Print Summary**

Evaluation node

The Evaluation node offers an easy way to evaluate and compare predictive models to choose the best model for your application. Evaluation charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (**quantiles**), and then plotting the value of the business criterion for each quantile, from highest to lowest. Multiple models are shown as separate lines in the plot.

Outcomes are handled by defining a specific value or range of values as a **hit**. Hits usually indicate success of some sort (such as a sale to a customer) or an event of interest (such as a specific medical diagnosis). You can define hit criteria on the Options tab of the dialog box, or you can use the default hit criteria as follows:

- **Flag** output fields are straightforward; hits correspond to *true* values.
- For **Nominal** output fields, the first value in the set defines a hit.
- For **Continuous** output fields, hits equal values greater than the midpoint of the field's range.

There are six types of evaluation charts, each of which emphasizes a different evaluation criterion.

Gains Charts

Gains are defined as the proportion of total hits that occurs in each quantile. Gains are computed as (number of hits in quantile / total number of hits) × 100%.

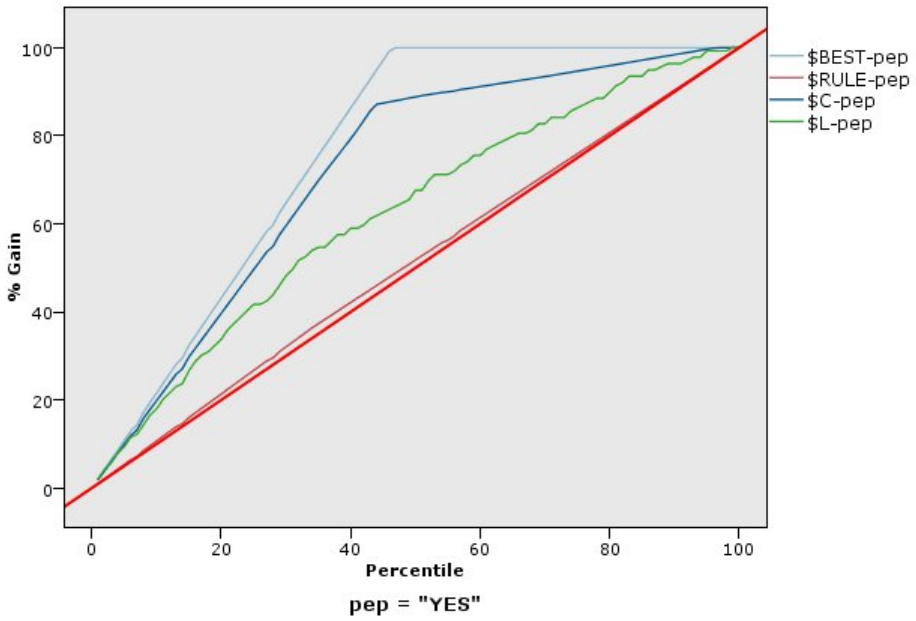


Figure 41. Gains chart (cumulative) with baseline, best line, and business rule displayed

Lift Charts

Lift compares the percentage of records in each quantile that are hits with the overall percentage of hits in the training data. It is computed as (hits in quantile / records in quantile) / (total hits / total records).

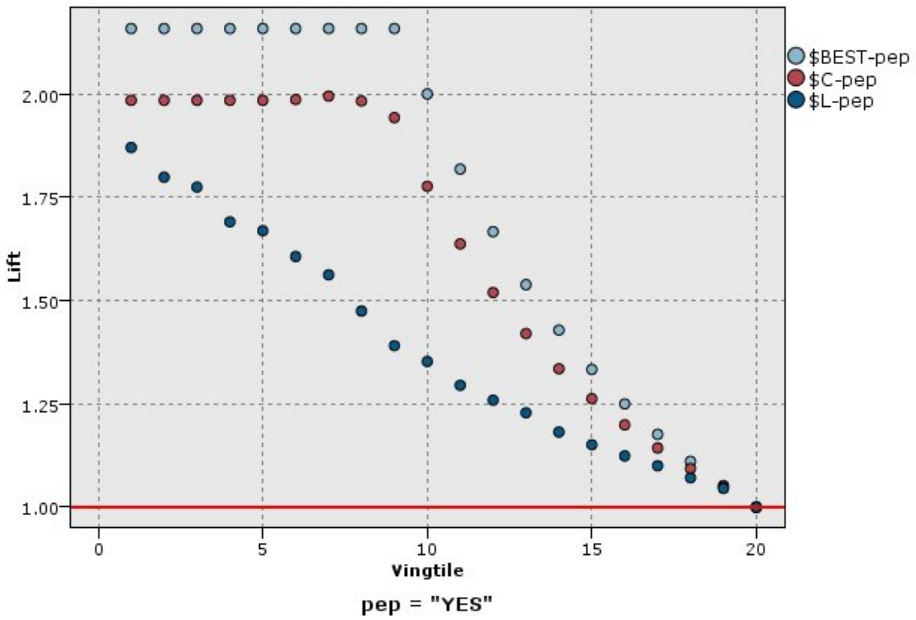


Figure 42. Lift chart (cumulative) using points and best line

Response Charts

Response is simply the percentage of records in the quantile that are hits. Response is computed as (hits in quantile / records in quantile) × 100%.

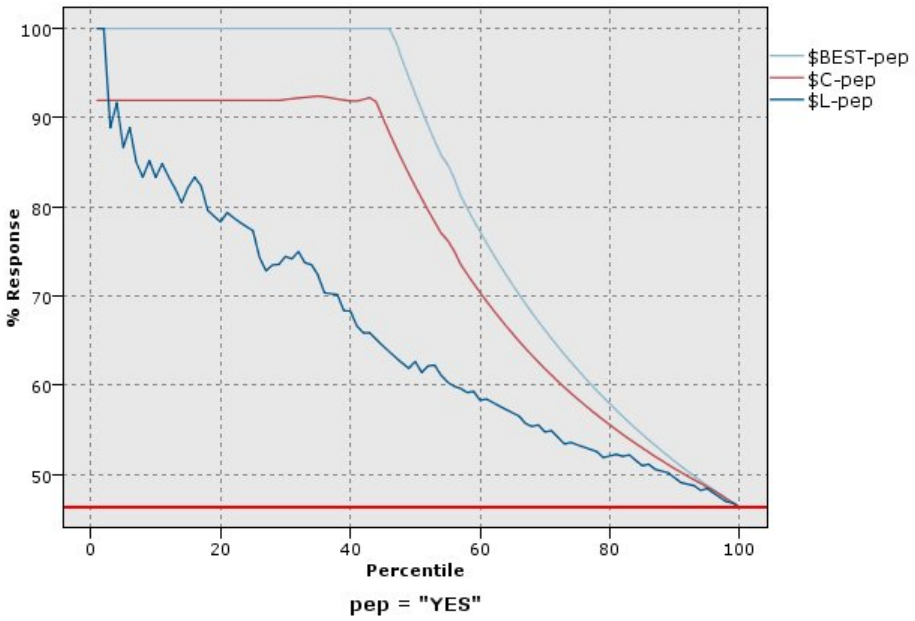


Figure 43. Response chart (cumulative) with best line

Profit Charts

Profit equals the **revenue** for each record minus the **cost** for the record. Profits for a quantile are simply the sum of profits for all records in the quantile. Revenues are assumed to apply only to hits, but costs apply to all records. Profits and costs can be fixed or can be defined by fields in the data. Profits are computed as (sum of revenue for records in quantile – sum of costs for records in quantile).

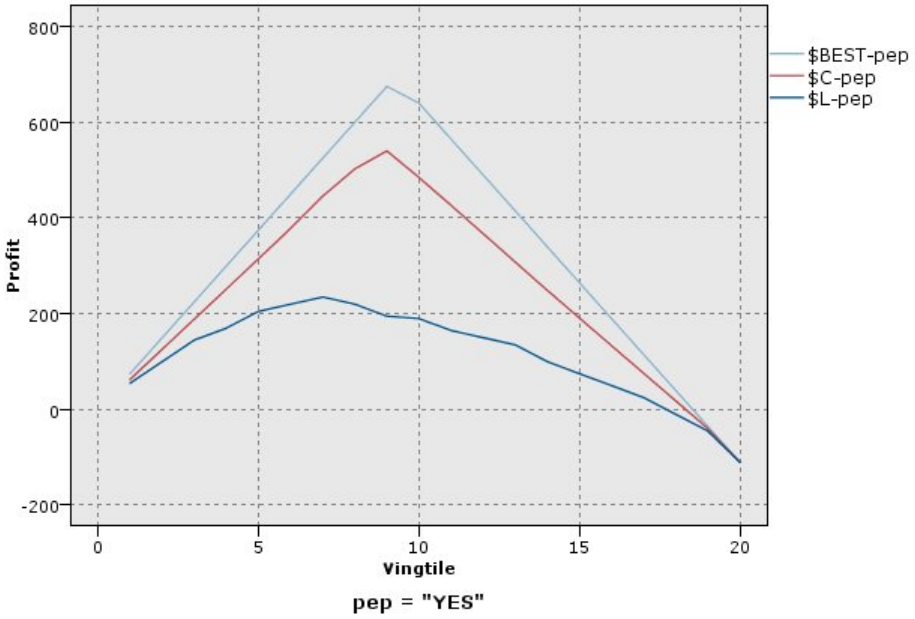


Figure 44. Profit chart (cumulative) with best line

ROI Charts

ROI (return on investment) is similar to profit in that it involves defining revenues and costs. ROI compares profits to costs for the quantile. ROI is computed as (profits for quantile / costs for quantile) × 100%.

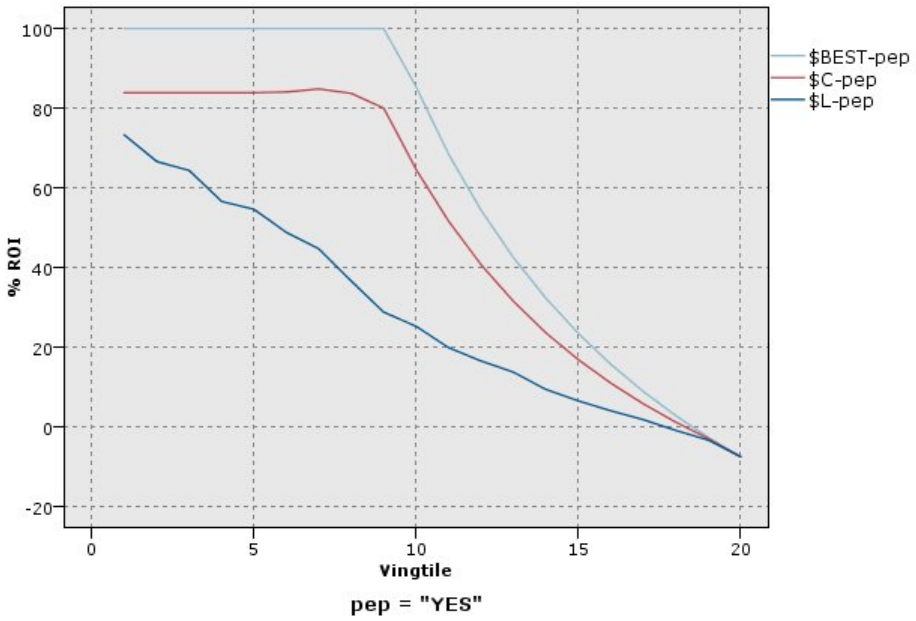


Figure 45. ROI chart (cumulative) with best line

ROC Charts

ROC (receiver operator characteristic) can only be used with binary classifiers. ROC can be used to visualize, organize and select classifiers based on their performance. A ROC chart plots the true positive rate (or sensitivity) against the false positive rate of the classifier. A ROC chart depicts the relative trade-offs between benefits (true positives) and costs (false positives). A true positive is an instance that is a hit and is classified as a hit. Therefore the true positive rate is calculated as the number of true positives / number of instances that are actually hits. A false positive is an instance that is a miss and is classified as a hit. Therefore the false positive rate is calculated as the number of false positives / number of instances that are actually misses.

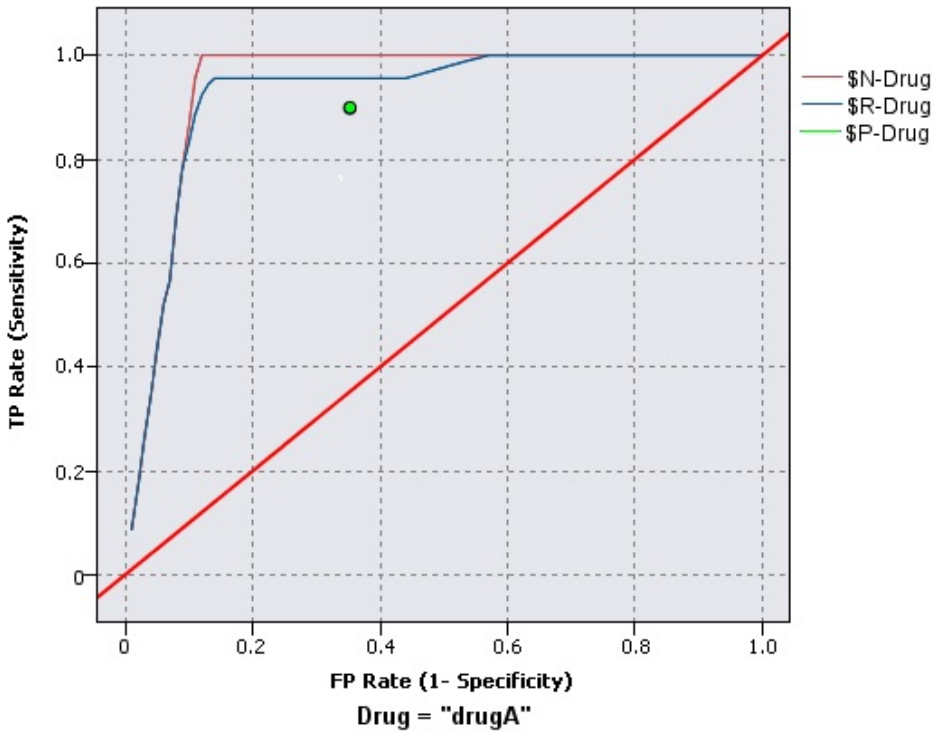


Figure 46. ROC chart with best line

Evaluation charts can also be cumulative, so that each point equals the value for the corresponding quantile plus all higher quantiles. Cumulative charts usually convey the overall performance of models better, whereas noncumulative charts often excel at indicating particular problem areas for models.

Note: The Evaluation node doesn't support the use of commas in field names. If you have field names containing commas, you must either remove the commas or surround the field name in quotes.

Evaluation Plot Tab

Chart type. Select one of the following types: **Gains**, **Response**, **Lift**, **Profit**, **ROI** (return on investment), or **ROC** (receiver operator characteristic).

Cumulative plot. Select to create a cumulative chart. Values in cumulative charts are plotted for each quantile plus all higher quantiles. (**Cumulative plot** is not available for ROC charts.)

Include baseline. Select to include a baseline in the plot, indicating a perfectly random distribution of hits where confidence becomes irrelevant. (**Include baseline** is not available for Profit and ROI charts.)

Include best line. Select to include a best line in the plot, indicating perfect confidence (where hits = 100% of cases). (**Include best line**) is not available for ROC charts.)

Use profit criteria for all chart types. Select to use the profit criteria (cost, revenue, and weight) when calculating the evaluation measures, instead of the normal count of hits. For models with certain numeric targets, such as a model that predicts the revenue obtained from a customer in response to an offer, the value of the target field gives a better measure of the performance of the model than the count of hits. Selecting this option enables the **Costs**, **Revenue**, and **Weight** fields for Gains, Response, and Lift charts. To use the profit criteria for these three chart types, it is recommended that you set **Revenue** to be the target field, **Cost** to be 0.0 so that profit is equal to revenue, and that you specify a user defined hit condition of "true" so that all records are counted as hits. (**Use profit criteria for all chart types** is not available for ROC charts.)

Find predicted/predictor fields using. Select either **Model output field metadata** to search for the predicted fields in the graph using their metadata, or select **Field name format** to search for them by name.

Plot score fields. Select this check box to enable the score fields chooser. Then select one or more range, or continuous, score fields; that is, fields which are not strictly predictive models but which might be useful to rank records in terms of propensity to be a hit. The Evaluation node can compare any combination of one or more score fields with one or more predictive models. A typical example might be to compare several RFM fields with your best predictive model.

Target. Select the target field using the field chooser. Choose any instantiated flag or nominal field with two or more values.

Note: This target field is only applicable to score fields (predictive models define their own targets), and is ignored if a custom hit criterion is set on the Options tab.

Split by partition. If a partition field is used to split records into training, test, and validation samples, select this option to display a separate evaluation chart for each partition. See the topic "Partition Node" on page 155 for more information.

Note: When splitting by partition, records with null values in the partition field are excluded from the evaluation. This will never be an issue if a Partition node is used, since Partition nodes do not generate null values.

Plot. Select the size of quantiles to plot in the chart from the drop-down list. Options include **Quartiles**, **Quintiles**, **Deciles**, **Vingtiles**, **Percentiles**, and **1000-tiles**. (**Plot** is not available for ROC charts.)

Style. Select **Line** or **Point**.

For all chart types except ROC charts, additional controls enable you to specify costs, revenue, and weights.

- **Costs.** Specify the cost associated with each record. You can select **Fixed** or **Variable** costs. For fixed costs, specify the cost value. For variable costs, click the Field Chooser button to select a field as the cost field. (**Costs** is not available for ROC charts.)
- **Revenue.** Specify the revenue associated with each record that represents a hit. You can select **Fixed** or **Variable** costs. For fixed revenue, specify the revenue value. For variable revenue, click the Field Chooser button to select a field as the revenue field. (**Revenue** is not available for ROC charts.)
- **Weight.** If the records in your data represent more than one unit, you can use frequency weights to adjust the results. Specify the weight associated with each record, using **Fixed** or **Variable** weights. For fixed weights, specify the weight value (the number of units per record). For variable weights, click the Field Chooser button to select a field as the weight field. (**Weight** is not available for ROC charts.)

Evaluation Options tab

The Options tab for evaluation charts provides flexibility in defining hits, scoring criteria, and business rules displayed in the chart. You can also set options for exporting the results of the model evaluation.

User defined hit. Select to specify a custom condition used to indicate a hit. This option is useful for defining the outcome of interest rather than deducing it from the type of target field and the order of values.

- **Condition.** When **User defined hit** is selected above, you must specify a CLEM expression for a hit condition. For example, `@TARGET = "YES"` is a valid condition indicating that a value of *Yes* for the target field will be counted as a hit in the evaluation. The specified condition will be used for all target fields. To create a condition, type in the field or use the Expression Builder to generate a condition expression. If the data are instantiated, you can insert values directly from the Expression Builder.

User defined score. Select to specify a condition used for scoring cases before assigning them to quantiles. The default score is calculated from the predicted value and the confidence. Use the Expression field to create a custom scoring expression.

- **Expression.** Specify a CLEM expression used for scoring. For example, if a numeric output in the range 0–1 is ordered so that lower values are better than higher, you might define a hit as `@TARGET < 0.5` and the associated score as `1 – @PREDICTED`. The score expression must result in a numeric value. To create a condition, type in the field or use the Expression Builder to generate a condition expression.

Include business rule. Select to specify a rule condition reflecting criteria of interest. For example, you may want to display a rule for all cases where `mortgage = "Y" and income >= 33000`. Business rules are drawn on the chart and labeled in the key as *Rule*. (**Include business rule** is not supported for ROC charts.)

- **Condition.** Specify a CLEM expression used to define a business rule in the output chart. Simply type in the field or use the Expression Builder to generate a condition expression. If the data are instantiated, you can insert values directly from the Expression Builder.

Export results to file. Select to export the results of the model evaluation to a delimited text file. You can read this file to perform specialized analyses on the calculated values. Set the following options for export:

- **Filename.** Enter the filename for the output file. Use the ellipsis button **(...)** to browse to the desired folder.
- **Delimiter.** Enter a character, such as a comma or space, to use as the field delimiter.

Include field names. Select this option to include field names as the first line of the output file.

New line after each record. Select this option to begin each record on a new line.

Evaluation Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Text. Either accept the automatically generated text label, or select **Custom** to specify a label.

X label. Either accept the automatically generated *x*-axis (horizontal) label or select **Custom** to specify a label.

Y label. Either accept the automatically generated *y*-axis (vertical) label or select **Custom** to specify a label.

Display gridline. Selected by default, this option displays a gridline behind the plot or graph that enables you to more easily determine region and band cutoff points. Gridlines are always displayed in white unless the graph background is white; in this case, they are displayed in gray.

Reading the Results of a Model Evaluation

The interpretation of an evaluation chart depends to a certain extent on the type of chart, but there are some characteristics common to all evaluation charts. For cumulative charts, higher lines indicate better models, especially on the left side of the chart. In many cases, when comparing multiple models the lines will cross, so that one model will be higher in one part of the chart and another will be higher in a different part of the chart. In this case, you need to consider what portion of the sample you want (which defines a point on the *x* axis) when deciding which model to choose.

Most of the noncumulative charts will be very similar. For good models, noncumulative charts should be high toward the left side of the chart and low toward the right side of the chart. (If a noncumulative chart shows a sawtooth pattern, you can smooth it out by reducing the number of quantiles to plot and re-executing the graph.) Dips on the left side of the chart or spikes on the right side can indicate areas where the model is predicting poorly. A flat line across the whole graph indicates a model that essentially provides no information.

Gains charts. Cumulative gains charts always start at 0% and end at 100% as you go from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right (shown in the chart if **Include baseline** is selected).

Lift charts. Cumulative lift charts tend to start above 1.0 and gradually descend until they reach 1.0 as you go from left to right. The right edge of the chart represents the entire dataset, so the ratio of hits in

cumulative quantiles to hits in data is 1.0. For a good model, lift should start well above 1.0 on the left, remain on a high plateau as you move to the right, and then trail off sharply toward 1.0 on the right side of the chart. For a model that provides no information, the line will hover around 1.0 for the entire graph. (If **Include baseline** is selected, a horizontal line at 1.0 is shown in the chart for reference.)

Response charts. Cumulative response charts tend to be very similar to lift charts except for the scaling. Response charts usually start near 100% and gradually descend until they reach the overall response rate (total hits / total records) on the right edge of the chart. For a good model, the line will start near or at 100% on the left, remain on a high plateau as you move to the right, and then trail off sharply toward the overall response rate on the right side of the chart. For a model that provides no information, the line will hover around the overall response rate for the entire graph. (If **Include baseline** is selected, a horizontal line at the overall response rate is shown in the chart for reference.)

Profit charts. Cumulative profit charts show the sum of profits as you increase the size of the selected sample, moving from left to right. Profit charts usually start near 0, increase steadily as you move to the right until they reach a peak or plateau in the middle, and then decrease toward the right edge of the chart. For a good model, profits will show a well-defined peak somewhere in the middle of the chart. For a model that provides no information, the line will be relatively straight and may be increasing, decreasing, or level depending on the cost/revenue structure that applies.

ROI charts. Cumulative ROI (return on investment) charts tend to be similar to response charts and lift charts except for the scaling. ROI charts usually start above 0% and gradually descend until they reach the overall ROI for the entire dataset (which can be negative). For a good model, the line should start well above 0%, remain on a high plateau as you move to the right, and then trail off rather sharply toward the overall ROI on the right side of the chart. For a model that provides no information, the line should hover around the overall ROI value.

ROC charts. ROC curves generally have the shape of a cumulative gains chart. The curve starts at the (0,0) coordinate and ends at the (1,1) coordinate as you go from left to right. A chart that rises steeply toward the (0,1) coordinate then levels off indicates a good classifier. A model that classifies instances at random as hits or misses will follow the diagonal from lower left to upper right (shown in the chart if **Include baseline** is selected). If no confidence field is provided for a model, the model is plotted as a single point. The classifier with the optimum threshold of classification is located closest to the (0,1) coordinate, or upper left corner, of the chart. This location represents a high number of instances that are correctly classified as hits, and a low number of instances that are incorrectly classified as hits. Points above the diagonal line represent good classification results. Points below the diagonal line represent poor classification results that are worse than if the instances were classified at random.

Using an Evaluation Chart

Using the mouse to explore an evaluation chart is similar to using a histogram or collection graph. The *x* axis represents model scores across the specified quantiles, such as vingtiles or deciles.

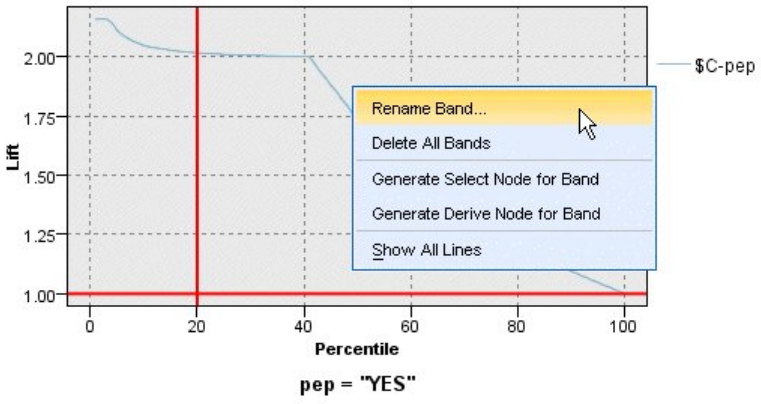


Figure 47. Working with an evaluation chart

You can partition the *x* axis into bands just as you would for a histogram by using the splitter icon to display options for automatically splitting the axis into equal bands. See the topic "Exploring Graphs" on page 245 for more information. You can manually edit the boundaries of bands by selecting **Graph Bands** from the Edit menu.

Once you have created an evaluation chart, defined bands, and examined the results, you can use options on the Generate menu and the context menu to automatically create nodes based upon selections in the graph. See the topic "Generating Nodes from Graphs" on page 252 for more information.

When generating nodes from an evaluation chart, you will be prompted to select a single model from all available models in the chart.

Select a model and click **OK** to generate the new node onto the stream canvas.

Map Visualization Node

The Map Visualization node can accept multiple input connections and display geospatial data on a map as a series of layers. Each layer is a single geospatial field; for example, the base layer might be a map of a country, then above that you might have one layer for roads, one layer for rivers, and one layer for towns.

Although most geospatial datasets ordinarily contain a single geospatial field, when there are multiple geospatial fields in a single input you can choose which fields to display. Two fields from the same input connection cannot be displayed at the same time; however, you can copy and paste the incoming connection and display a different field from each.

Map Visualization Plot Tab

Layers

This table displays information about the inputs to the map node. The order of the layers dictates the order in which the layers are displayed in both the map preview and the visual output when the node is executed. The top row in the table is the 'top' layer and the bottom row is the 'base' layer; in other words, each layer is shown on the map in front of the layer directly below it in the table.

Note: Where a layer in the table contains a three-dimensional geospatial field only the x and y axes are plotted. The z axis is ignored.

Name
Names are automatically created for each layer and are made up using the following format: `tag[source node:connected node]`. By default, the tag is shown as a number, with 1 representing the first input that is connected, 2 for the second input, and so on. If required, press the **Edit Layer** button to change the tag in the **Change Map Layer Options** dialog box. For example, you might change the tag to be "roads" or "cities" to reflect the data input.

Type
Shows the measurement type icon of the geospatial field that is selected as the layer. If the input data contains multiple fields with a geospatial measurement type the default selection uses the following sort order:

1. Point
2. Linestring
3. Polygon
4. MultiPoint
5. MultiLinestring
6. MultiPolygon

Note: If there are two fields with the same measurement type, the first field (alphabetically by name) is selected by default.

Symbol

Note: This column is only completed for Point and MultiPoint fields.

Shows the symbol that is used for Point or MultiPoint fields. If required, press the **Edit Layer** button to change the symbol in the **Change Map Layer Options** dialog box.

Color

Shows the color that is selected to represent layer on the map. If required, press the **Edit Layer** button to change the color in the **Change Map Layer Options** dialog box. The color is applied to different items depending on the measurement type.

- For Points or MultiPoints, the color is applied to the symbol for the layer.
- For Linestrings and Polygons, the color is applied to the entire shape. Polygons always have a black outline; the color that is shown in the column is the color that is used to fill the shape.

Preview

This pane shows a preview of the current selection of inputs in the **Layers** table. The preview takes into account the order of layers, symbol, color, and any other display settings that are associated with the layers and, when possible, updates the display whenever the settings change. If you change details elsewhere in your stream, such as the geospatial fields to use as layers, or if you amend details such as associated aggregation functions, you might need to click the **Refresh Data** button to update the preview.

Use the **Preview** to set your display settings before you run your stream. To protect against time delays that could be caused by using a large dataset, the preview samples each layer and creates a display from the first 100 records.

Change map layers

You can use the **Change Map Layer Options** dialog box to amend various details of any layer that is shown on the **Plot** tab of the Map Visualization node.

Input details

Tag

By default, the tag is a number; you can replace this number with a more meaningful tag to help identify the layer on the map. For example, the tag might be the name of the data input, such as "Cities".

Layer field

If you have more than one geospatial field in your input data, use this option to select the field that you want to display as a layer on the map.

By default, the layers you can choose from are in the following sort order.

- Point
- Linestring
- Polygon
- MultiPoint
- MultiLinestring
- MultiPolygon

Display settings

Use hex binning

Note: This option affects only point and multipoint fields.

Hex (hexagonal) binning combines proximate points (based on their x and y coordinates) into a single point to display on the map. The single point is shown as a hexagon, but is effectively rendered as a polygon.

Because the hexagon is rendered as a polygon, any point fields with hex binning turned on are traeted as polygons. This means that if you choose to **Order by type** on the map node dialog box, any point layers that have hex binning applied are rendered above polygon layers but below linestring and point layers.

If you use hex binning for a multipoint field, the field is first converted into a point field by binning the multipoint values to calculate the central point. Central points are used to calculate the hex bins.

Aggregation

Note: This column is only available when you select the **Use hex binning** check box and also select an **Overlay**.

If you select an **Overlay** field for a points layer that uses hex binning, all of the values in that field must be aggregated for all points within the hexagon. Specify an aggregation function for any overlay fields you want to apply to the map. The available aggregation functions depend on the measurement type.

- Aggregation functions for a Continuous measurement type, with Real or Integer storage:
  - Sum
  - Mean
  - Min
  - Max
  - Median
  - First Quartile
  - Third Quartile
- Aggregation functions for a Continuous measurement type, with Time, Date, or Timestamp storage:
  - Mean
  - Min
  - Max
- Aggregation functions for Nominal or Categorical measurement types:
  - Mode
  - Min
  - Max
- Aggregation functions for a Flag measurement type:
  - True, if any are true
  - False, if any are false

Color

Use this option to choose either a standard color to apply to all features of the geospatial field, or an overlay field, which colors the features based on the values from another field in the data.

If you select **Standard**, you can choose a color from the palette of colors that are shown in the **Chart Category Color Order** pane in the Display tab of the **User Options** dialog box.

If you select **Overlay**, you can choose any field from the data source that contains the geospatial field that was selected as the **Layer field**.

- For nominal or categorical overlay fields, the color palette that you can choose from is the same as that shown for the **Standard** color options.

- For continuous and ordinal overlay fields a second drop-down list is displayed, from which you select a color. When you select a color, the overlay is applied by varying the saturation of that color according to the values in the continuous or ordinal field. The highest value uses the color chosen from the drop-down list and lower values are shown by correspondingly lower saturations.

Symbol

Note: Only enabled for point and multipoint measurement types.

Use this option to choose whether to have a **Standard** symbol, which is applied to all records of your geospatial field, or an **Overlay** symbol, which changes the symbol icon for the points based on the values of another field in your data.

If you select **Standard**, you can choose one of the default symbols from a drop-down list to represent the point data on the map.

If you select **Overlay**, you can choose any nominal, ordinal, or categorical field from the data source that contains the geospatial field that was selected as the **Layer field**. For each value in the overlay field, a different symbol is displayed on the map.

For example, your data might contain a point field that represents the locations of shops and the overlay might be a shop type field. In this example, all food shops might be identified on the map by a cross symbol and all electronics shops by a square symbol.

Size

Note: Only enabled for point, multipoint, linestring and multilinestring measurement types.

Use this option to choose whether to have a **Standard** size, which is applied to all records of your geospatial field, or an **Overlay** size, which changes the size of the symbol icon or line thickness based on the values of another field in your data.

If you select **Standard**, you can choose a pixel width value. Options available are 1, 2, 3, 4, 5, 10, 20, or 30.

If you select **Overlay**, you can choose any field from the data source that contains the geospatial field that was selected as the **Layer field**. The thickness of the line or point varies depending on the value of the chosen field.

Transparency

Use this option to choose whether to have a **Standard** transparency, which is applied to all records of your geospatial field, or an **Overlay** transparency, which changes the transparency of the symbol, line, or polygon based on the values of another field in your data.

If you select **Standard**, you can choose from a selection of transparency levels that start from 0% (opaque) and increase in 10% increments to 100% (transparent).

If you select **Overlay**, you can choose any field from the data source that contains the geospatial field that was selected as the **Layer field**. A different level of transparency is displayed on the map for each value in the overlay field. The transparency is applied to the color chosen from the color drop-down list for the point, line, or polygon.

Data Label

Note: This option is not available if you select the **Use hex binning** check box.

Use this option to select a field to use as data labels on the map. For example, if applied to a polygon layer, the data label might be the name field, containing the name of each polygon. If you select the name field, those names are displayed on the map.

Map Visualization Appearance Tab

You can specify appearance options before graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Caption. Enter the text to use for the graph's caption.

t-SNE node

t-Distributed Stochastic Neighbor Embedding (t-SNE)© is a tool for visualizing high-dimensional data. It converts affinities of data points to probabilities. The affinities in the original space are represented by Gaussian joint probabilities and the affinities in the embedded space are represented by Student's t-distributions. This allows t-SNE to be particularly sensitive to local structure and has a few other advantages over existing techniques: [1]

- Revealing the structure at many scales on a single map
- Revealing data that lie in multiple, different, manifolds, or clusters
- Reducing the tendency to crowd points together at the center

The t-SNE node in SPSS Modeler is implemented in Python and requires the `scikit-learn`© Python library. For details about t-SNE and the scikit-learn library, see:

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

The Python tab on the Nodes Palette contains this node and other Python nodes. The t-SNE node is also available on the Graphs tab.

[1] References:

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE node Expert options

Choose **Simple** mode or **Expert** mode depending on which options you want to set for the t-SNE node.

Visualization type. Select **2D** or **3D** to specify whether to draw the graph as two-dimensional or three-dimensional.

Method. Select **Barnes Hut** or **Exact**. By default, the gradient calculation algorithm uses Barnes-Hut approximation which runs must faster than the Exact method. Barnes-Hut approximation allows the t-SNE technique to be applied to large, real-world datasets. The Exact algorithm will do a better job of avoiding nearest-neighbor errors.

Init. Select **Random** or **PCA** for the initialization of embedding.

Target Field. Select the target field to show as a colormap on the output graph. The graph will use one color if no target field is specified here.

Optimization

Perlexity. The perplexity is related to the number of nearest neighbors that are used in other manifold learning algorithms. Larger datasets usually required a larger perplexity. Consider selecting a value between **5** and **50**. Default is **30**, and the range is **2 - 9999999**.

Early exaggeration. This setting controls how tight natural clusters in the original space will be in the embedded space, and how much space will be between them. Default is **12**, and the range is **2 - 9999999**.

Learning rate. If the learning rate is too high, the data might look like a "ball" with any point approximately equidistant from its nearest neighbors. If the learning rate is too low, most points may look

compressed in a dense cloud with few outliers. If the cost function gets stuck in a bad local minimum, increasing the learning rate may help. Default is **200**, and the range is **0 - 9999999**.

Max iterations. The maximum number of iterations for the optimization. Default is **1000**, and the range is **250 - 9999999**.

Angular size. The angular size of a distant node as measured from a point. Enter a value between **0** and **1**. Default is **0.5**.

Random seed

Set random seed. Select this option and click **Generate** to generate the seed used by the random number generator.

Optimization stop condition

Max iterations without progress. The maximum number of iterations without progress to perform before stopping the optimization, used after 250 initial iterations with early exaggeration. Note that progress is only checked every 50 iterations, so this value is rounded to the next multiple of 50. Default is **300**, and the range is **0 - 9999999**.

Min gradient norm. If the gradient norm is below this minimum threshold, the optimization will stop. Default is **1.0E-7**.

Metric. The metric to use when calculating distance between instances in a feature array. If the metric is a string, it must be one of the options allowed by `scipy.spatial.distance.pdist` for its metric parameter, or a metric listed in `pairwise.PAIRWISE_DISTANCE_FUNCTIONS`. Select one of the available metric types. Default is **euclidean**.

When number of records greater than. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 points. Performance is enhanced for large datasets when you select the **Bin** or **Sample** options. Alternatively, you can choose to plot all data points by selecting **Use all data**, but you should note that this may dramatically decrease the performance of the software.

- **Bin.** Select to enable binning when the dataset contains more than the specified number of records. Binning divides the graph into fine grids before actually plotting and counts the number of connections that would appear in each of the grid cells. In the final graph, one connection is used per cell at the bin centroid (average of all connection points in the bin).
- **Sample.** Select to randomly sample the data to the specified number of records.

The following table shows the relationship between the settings on the Expert tab of the SPSS Modeler t-SNE node dialog and the Python t-SNE library parameters.

Table 37. Node properties mapped to Python library parameters

SPSS Modeler setting	Script name (property name)	Python t-SNE parameter
Mode	`mode_type`	
Visualization type	`n_components`	`n_components`
Method	`method`	`method`
Initialization of embedding	`init`	`init`
Target	`target_field`	`target_field`
Perplexity	`perplexity`	`perplexity`
Early exaggeration	`early_exaggeration`	`early_exaggeration`
Learning rate	`learning_rate`	`learning_rate`
Max iterations	`n_iter`	`n_iter`

Table 37. Node properties mapped to Python library parameters (continued)		
SPSS Modeler setting	**Script name (property name)**	**Python t-SNE parameter**
Angular size	`angle`	`angle`
Set random seed	`enable_random_seed`	
Random seed	`random_seed`	`random_state`
Max iterations without progress	`n_iter_without_progress`	`n_iter_without_progress`
Min gradient norm	`min_grad_norm`	`min_grad_norm`
Perform t-SNE with multiple perplexities	`isGridSearch`	

t-SNE node Output options

Specify options for the t-SNE node output on the **Output** tab.

Output name. Specify the name of the output that is produced when the node runs. If you select **Auto**, the name of the output is automatically set.

Output to screen. Select this option to generate and display the output in a new window. The output is also added to the Output manager.

Output to file. Select this option to save the output to a file. Doing so enables the **File name** and **File type** fields. The t-SNE node requires access to this output file if you want to create plots using other fields for comparison purposes – or to use its output as predictors in classification or regression models. The t-SNE model creates a result file of x, y (and z) coordinate fields that is most easily accessed using a Fixed File source node.

Accessing and plotting t-SNE data

If you use the **Ouptut to file** option to save t-SNE output to files, you can then create plots using other fields for comparison purposes – or use the output as predictors in classification or regression models. The t-SNE model creates a result file of x, y (and z) coordinate fields that is most easily accessed using a Fixed File source node. This section provides example information.

1. In the t-SNE node dialog, open the **Output** tab.
2. Select **Output to file** and type a file name. Use the default HTML file type. When you run the model, this will generate three output files in your output location:
   - A text file (`result_xxxxxx.txt`)
   - An HTML file (the file name you specified)
   - A PNG file (`tsne_chart_yyyyyy.png`)

   The text file will contain the data you need, but for technical reasons it may be in standard or scientific format. If it's in scientific format (1.11111111e+01), then you need to create a new stream that recognizes the format:

Accessing t-SNE plot data when the text file is in scientific numeric format

1. Create a new stream (**File** > **New Stream**).
2. Go to **Tools** > **Stream Properties** > **Options**, select **Number formats**, and select **Scientific (#.###E +##)** for the number display format.
3. Add a Fixed File source node to your canvas, and use the following settings on the File tab:
   - Skip header lines: 1
   - Record length: 54

- tSNE_x Start: 3, Length: 16
- tSNE_y Start: 20, Length: 16
- tSNE_z Start: 36, Length: 16

4. On the Type tab, the numbers should be recognized as Real. Click Read Values and you should see field values similar to:

Table 38. Example field values

Field	Measurement	Values
tSNE_x	Continuous	[-7.07176703,7.14338837]
tSNE_y	Continuous	[-9.2188112,8.89647667]
tSNE_x	Continuous	[-9.95892882,9.95742482]

5. Add a Select node to the stream so you can delete the following bottom two rows of text in the file which are read as nulls:

```
*****************
Perform t-SNE (total time 9.5s)
```

On the Settings tab of the Select node, select **Discard** for the mode and use the condition `@NULL(tSNE_x)` to delete the rows.

6. Add a Type node and a Flat File export node to the stream to create a Var. File source node that will copy and paste back into your original stream.

Accessing t-SNE plot data when the text file is in standard numeric format

1. Create a new stream (**File** > **New Stream**).
2. Add a Fixed File source node to your canvas. The following three nodes are all that's required to access the t-SNE data.

Figure 48. Stream for accessing t-SNE plot data in standard numeric format

3. Use the following settings on the File tab of the Fixed File source node:
   - Skip header lines: 1
   - Record length: 29
   - tSNE_x Start: 3, Length: 12
   - tSNE_y Start: 16, Length: 12
4. On the Filter tab, you can rename `field1` and `field2` to `tsneX` and `tsneY`.
5. Add a Merge node to connect it to your stream by using the **Order** merge method.
6. You can now use a Plot node to plot `tsneX` versus `tsneY` and color it with your field under investigation.

t-SNE model nuggets

t-SNE model nuggets contain all of the information captured by the t-SNE model. The following tabs are available.

Graph

The **Graph** tab displays chart output for the t-SNE node. A pyplot scatter chart shows the low dimensions result. If you didn't select the **Perform t-SNE with multiple perplexities** option on the Expert tab of the t-SNE node, only one graph is included rather than six graphs with different perplexities.

Text output

The **Text output** tab displays the results of the t-SNE algorithm. If you chose the **2D** visualization type on the Expert tab of the t-SNE node, the result here is the point value in two dimensions. If you chose **3D**, the result is the point value in three dimensions.

E-Plot (Beta) node

E-Plot (Beta) nodes show the relationship between numeric fields. The E-Plot (Beta) node is similar to the Plot node, but its options differ and it uses new graphing capabilities. Use the node to play around with new graphing features in SPSS Modeler.

The E-Plot (Beta) node provides scatterplots, line charts, and bar charts to illustrate the relationship between numeric fields. The new graphing interface in this node is intuitive and modern, very customizable, and the data charts are interactive. For more information, see "Using an e-plot graph" on page 243.

E-Plot (Beta) node Plot tab

Plots show values of a *Y* field against values of an *X* field. Often, these fields correspond to a dependent variable and an independent variable, respectively.

X field. From the list, select the field to display on the horizontal *x* axis.

Y field. From the list, select the field to display on the vertical *y* axis.

Overlay. There are different ways to illustrate categories for data values. For example, you might use a *maincrop* field as a color overlay to indicate the *estincome* and *claimvalue* values for the main crop grown by claim applicants. Select fields for color mapping, size mapping, and shape mapping in the output. Also select any other interested fields to include in the interactive output. See the topic "Aesthetics, Overlays, Panels, and Animation" on page 168 for more information.

Once you have set options for an e-plot, you can run the plot directly from the dialog box by clicking **Run**. You may, however, want to use the Options tab for additional specifications.

E-Plot (Beta) node Options tab

Maximum number of records to plot. Specify a method for plotting large datasets. You can specify a maximum dataset size or use the default 2,000 records. Performance is enhanced for large datasets when you select the **Sample** option. The Sample option randomly samples the data to the number of records entered in the text field. Alternatively, you can choose to plot all data points by selecting **Use all data**, but you should note that this may dramatically decrease the performance of the software.

E-Plot (Beta) Appearance tab

You can specify a title and subtitle before graph creation, if desired. These options can also be specified or changed after graph creation.

Title. Enter the text to use for the graph's title.

Subtitle. Enter the text to use for the graph's subtitle.

Using an e-plot graph

The E-Plot (Beta) node provides scatterplots, line charts, and bar charts to illustrate the relationship between numeric fields. The new graphing interface introduced in this beta node includes many new and improved capabilities.

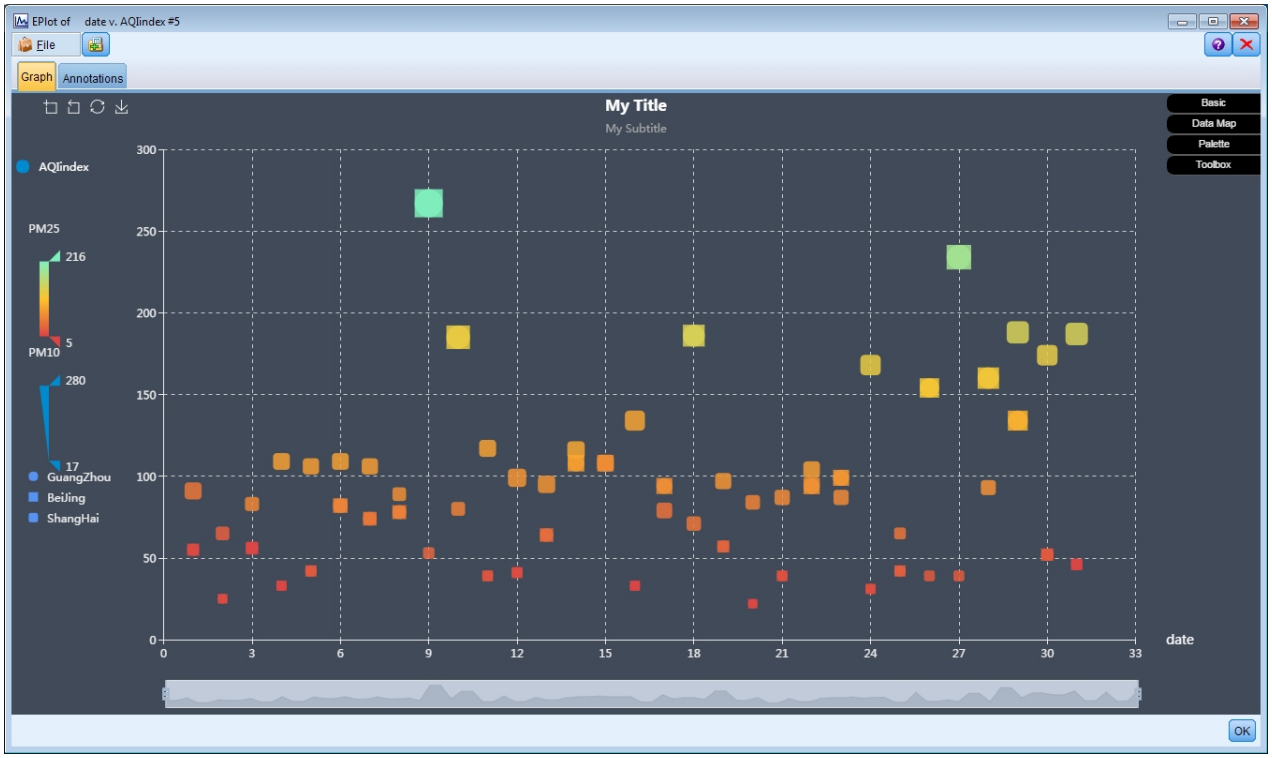


Figure 49. E-Plot (Beta) scatterplot graph

On the Graph tab, the upper left corner provides a toolbar for zooming in on a specific section of the chart, zooming back out, returning to the initial full view, or saving the chart for external use:

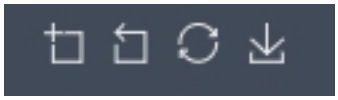


Figure 50. Toolbar

At the bottom of the window, you can use the slider to zoom in on a specific section of the chart. Move the small rectangular controls on the right and the left to zoom. To use this slider, you must first turn it on in the Toolbox options area.

Figure 51. Zoom slider

The left side of the window provides controls for changing the range of values shown. To use these controls, you must first specify options in the Data Map options area. In the example below, a field named `PM25` is selected for the color map, a field named `PM10` is selected for the size map, and a field called `City` is selected for the shape map. You can hover over the vertical color bars to highlight the corresponding areas of the graph, or slide the upper and lower triangles.

Figure 52. Range controls

On the right side of the window, a set of expandable options is available that you can use to interact with the data and change the chart's appearance in real time:

Figure 53. Expandable options

Basic options

Figure 54. Basic options

Select the dark or light theme, specify a title and a subtitle, select a chart type (scatter, line, or bar), and select the series shown on the Y axis. If you select **Line** chart, only the fields on the Y axis will be displayed and only the fields on the Y axis will be available in the Data Map options for color map and size map. If you select **Bar** chart, only color map options will be available in the Data Map options. For the series, all **Interested** fields you selected on the Plot tab of the E-Plot node will be available here.

Data Map options

Figure 55. Data map options

Select a continuous field or a categorical field for the **Color Map**. If a continuous field is selected, colors from green to red will be shown. The lower the value is, the closer to red its color will be – and the higher the value is, the closer to green its color will be. If a categorical field is selected, the field color will be displayed according to the defined color palette.
Size Map only supports continuous fields. The lower the value is on the chart, the smaller its plot size will be.
Shape Map only supports categorical fields. The shape displayed on the map is defined by a categorical field that splits the visualizaton into elements of different shapes – one for each category.

Palette options

Figure 56. Palette options

Use the Palette if you want to customize the colors used for the title and the series. Select the title or the series from the drop-down, click **Edit Predefined Colors**, and then click **More** to select a color. Or you can use the RGB or Hex fields to specify an exact color.

Toolbox options

Figure 57. Toolbox options

Use the Toolbox options to turn on or off the zoom slider, set gridline properties, and turn on or off mouse tracking. Mouse tracking shows your exact coordinate position when hovering over the chart.

Exploring Graphs

While Edit mode allows you to edit the graph's layout and look, Explore mode allows you to analytically explore the data and values represented by the graph. The main goal of exploration is to analyze the data and then identify values using bands, regions, and marking to generate Select, Derive, or Balance nodes. To select this mode, choose **View > Explore Mode** from the menus (or click the toolbar icon).

While some graphs can use all of the exploration tools, others accept only one. Explore mode includes:

- Defining and editing bands, which are used to split the values along a scale *x* axis. See the topic "Using Bands" on page 246 for more information.
- Defining and editing regions, which are used to identify a group of values within the rectangular area. See the topic "Using Regions" on page 249 for more information.
- Marking and unmarking elements to hand select the values that could be used to generate a Select or Derive node. See the topic "Using Marked Elements" on page 251 for more information.
- Generating nodes using the values identified by bands, regions, marked elements, and web links to use in your stream. See the topic "Generating Nodes from Graphs" on page 252 for more information.

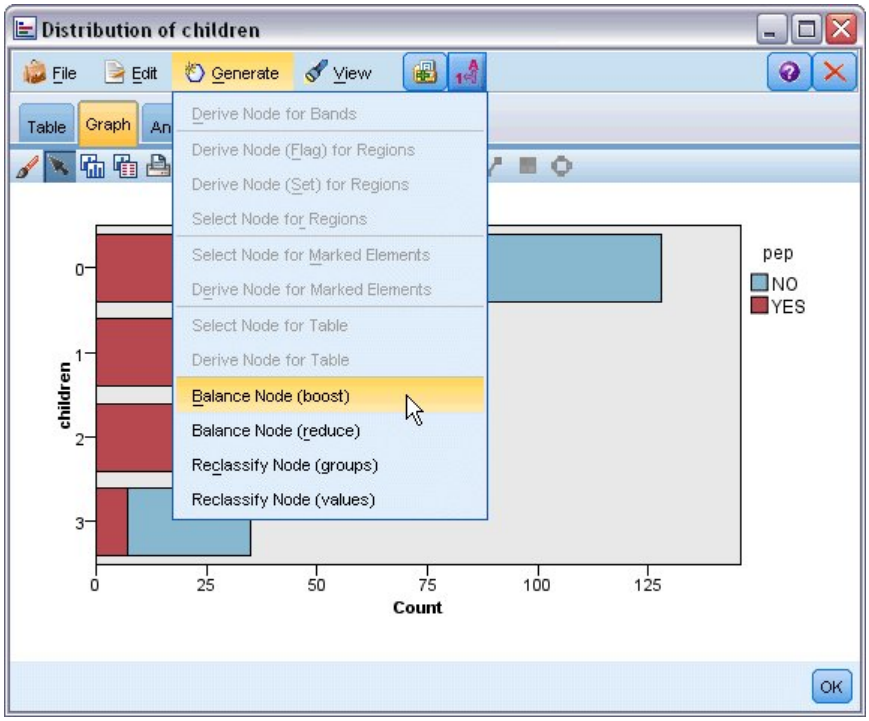


Figure 58. Graph with the generate menu showing

Using Bands

In any graph with a scale field on the *x* axis, you can draw vertical band lines to split the range of values on the *x* axis. If a graph has multiple panels, a band line drawn in one panel is also represented in the other panels as well.

Not all graphs accept bands. Some of those graphs which can have bands include: histograms, bar charts and distributions, plots (line, scatter, time, etc.), collections, and evaluation charts. In graphs with paneling, bands appears in all panels. And in some cases in a SPLOM, you will see a horizontal band line since the axis on which the field/variable band was drawn has been flipped.

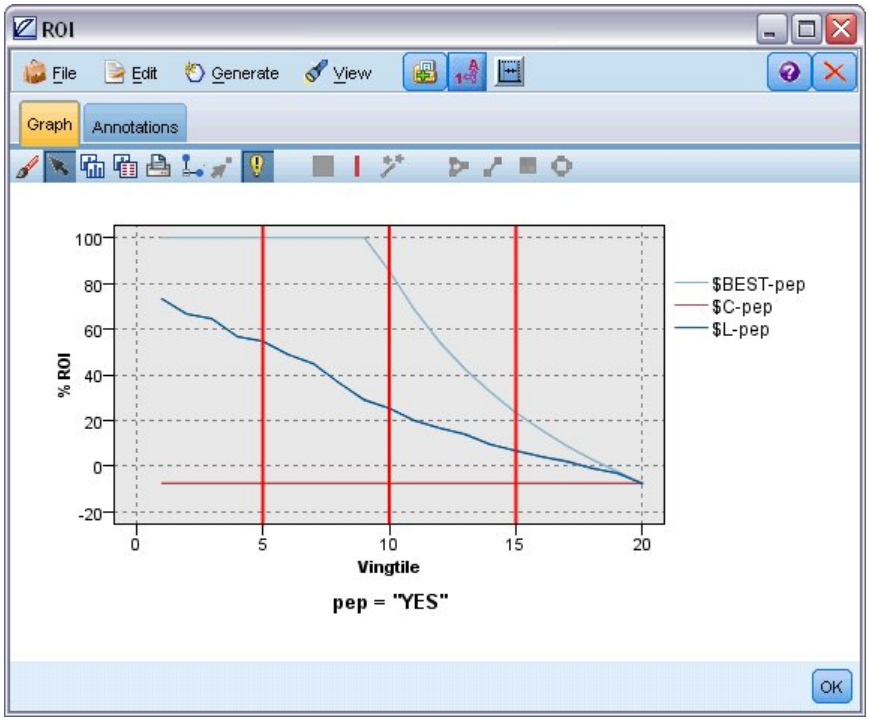


Figure 59. Graph with three bands

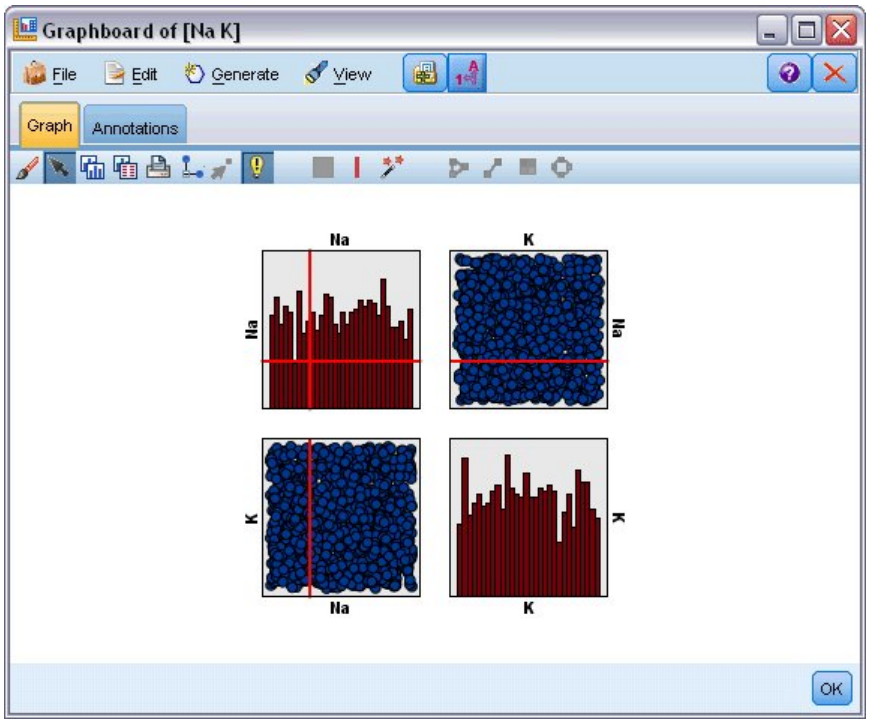


Figure 60. SPLOM with bands

Defining Bands

In a graph without bands, adding a band line splits the graph into two bands. The band line value represents the starting point, also referred to as the lower bound, of the second band when reading the graph from left to right. Likewise, in a graph with two bands, adding a band line splits one of those bands into two, which results in three bands. By default, bands are named *bandN*, where *N* equals the number of bands from left to right on the *x* axis.

Once you have a defined a band, you can drag-and-drop the band to reposition it on the *x* axis. You can see more shortcuts by right-clicking within a band for tasks such as renaming, deleting, or generating nodes for that specific band.

To define bands:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Draw Band button.

Figure 61. Draw Bands toolbar button

3. In a graph that accepts bands, click the *x*-axis value point at which you want to define a band line.

Note: Alternatively, click the **Split Graph into Bands** toolbar icon and enter the number of equal bands you want and click **Split**.

Figure 62. Splitter icon used to expand the toolbar with options for splitting into bands

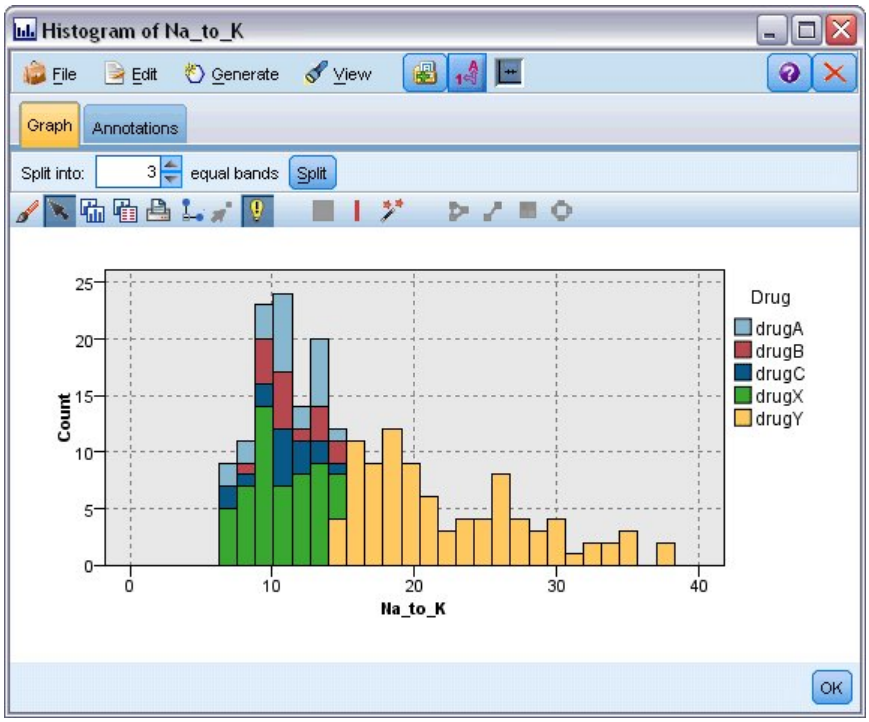


Figure 63. Creating equal bands toolbar with bands enabled

Editing, Renaming, and Deleting Bands

You can edit the properties of existing bands in the Edit Graph Bands dialog box or through context menus in the graph itself.

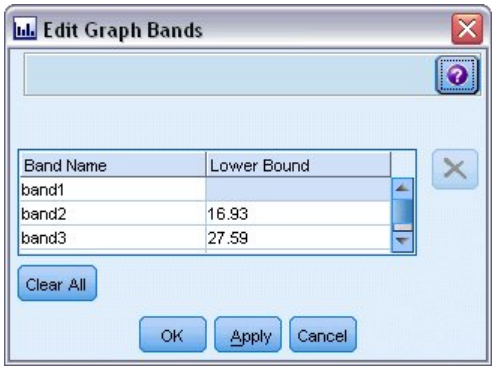


Figure 64. Edit Graph Bands dialog box

To edit bands:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Draw Band button.
3. From the menus, choose **Edit > Graph Bands**. The Edit Graph Bands dialog box opens.
4. If you have multiple fields in your graph (such as with SPLOM graphs), you can select the field you want in the drop-down list.
5. Add a new band by typing a name and lower bound. Press the Enter key to begin a new row.
6. Edit a band's boundary by adjusting the **Lower Bound** value.
7. Rename a band by entering a new band name.
8. Delete a band by selecting the line in the table and clicking the delete button.
9. Click **OK** to apply your changes and close the dialog box.

Note: Alternatively, you can delete and rename bands directly in the graph by right-clicking the band's line and choosing the option you want from the context menus.

Using Regions

In any graph with two scale (or range) axes, you can draw regions to group values within a rectangular area you draw, called a region. A **region** is an area of the graph described by its minimum and maximum *X* and *Y* values. If a graph has multiple panels, a region drawn in one panel is also represented in the other panels as well.

Not all graphs accept regions. Some of those graphs that accept regions include: plots (line, scatter, bubble, time, etc.), SPLOM, and collections. These regions are drawn in X,Y space and cannot, therefore, be defined in 1-D, 3-D, or animated plots. In graphs with paneling, regions appear in all panels. With a scatterplot matrix (SPLOM), a corresponding region will appear in the corresponding upper plots but not on the diagonal plots since they show only one scale field.

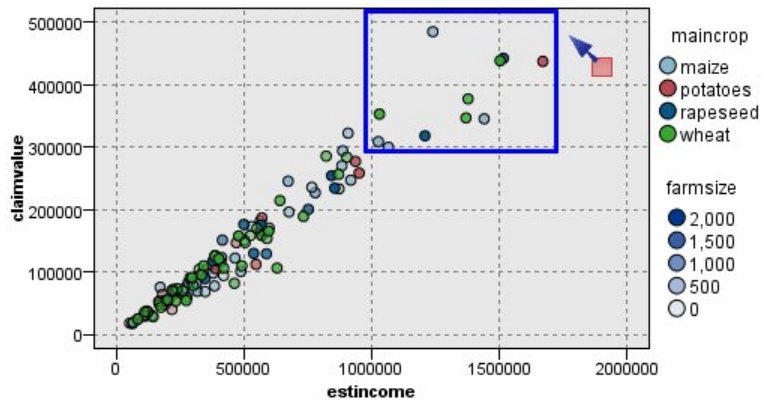


Figure 65. Defining a region of high claim values

Defining Regions

Wherever you define a region, you are creating a grouping of values. By default, each new region is called *Region<N>*, where *N* corresponds to the number of regions already created.

Once you have a defined a region, you can right-click the region line to get some basic shortcuts. However, you can see many other shortcuts by right-clicking inside the region (not on the line) for tasks such as renaming, deleting, or generating Select and Derive nodes for that specific region.

You can select subsets of records on the basis of their inclusion in a particular region or in one of several regions. You can also incorporate region information for a record by producing a Derive node to flag records based on their inclusion in a region. See the topic "Generating Nodes from Graphs" on page 252 for more information.

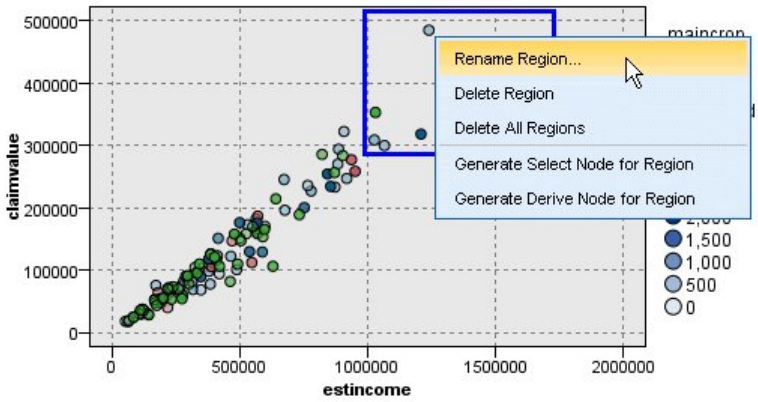


Figure 66. Exploring the region of high claim values

To define regions:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Draw Region button.

Figure 67. Draw Region toolbar button

3. In a graph that accepts regions, click and drag your mouse to draw the rectangular region.

Editing, Renaming and Deleting Regions

You can edit the properties of existing regions in the Edit Graph Regions dialog box or through context menus in the graph itself.

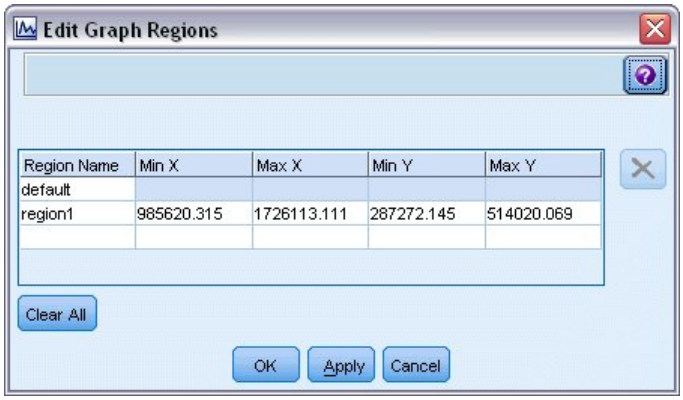


Figure 68. Specifying properties for the defined regions

To edit regions:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Draw Region button.
3. From the menus, choose **Edit > Graph Regions**. The Edit Graph Regions dialog box opens.
4. If you have multiple fields in your graph (for example, SPLOM graphs), you must define the field for the region in the *Field A* and *Field B* columns.
5. Add a new region on a new line by typing a name, selecting field names (if applicable) and defining the maximum and minimum boundaries for each field. Press the Enter key to begin a new row.
6. Edit existing region boundaries by adjusting the **Min** and **Max** values for *A* and *B*.
7. Rename a region by changing the region's name in the table.
8. Delete a region by selecting the line in the table and clicking the delete button.
9. Click **OK** to apply your changes and close the dialog box.

Note: Alternatively, you can delete and rename regions directly in the graph by right-clicking the region's line and choosing the option you want from the context menus.

Using Marked Elements

You can mark elements, such as bars, slices, and points, in any graph. Lines, areas, and surfaces cannot be marked in graphs other than time plot, multiplot, and evaluation graphs since lines refers to fields in those cases. Whenever you mark an element, you are essentially highlighting all of the data represented by that element. In any graph where the same case is represented in more than one place (such as SPLOM), marking is synonymous with brushing. You can mark elements in graphs and even within bands and regions. Whenever you mark an element and then go back into Edit mode, the marking still remains visible.

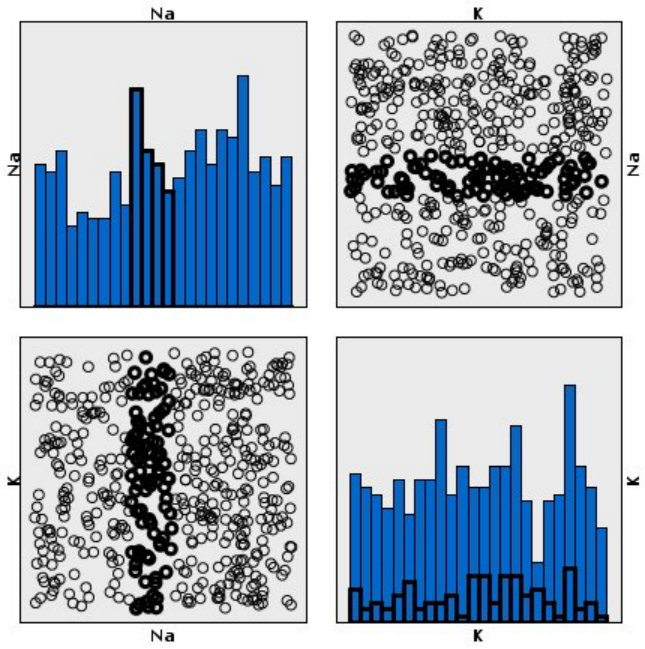


Figure 69. Marking elements in a SPLOM

You can mark and unmark elements by clicking on elements in the graph. When you first click an element to mark it, the element appears with a thick border color to indicate that it has been marked. If you click the element again, the border disappears and the element is no longer marked. To mark multiple elements, you can either hold down the Ctrl key while clicking elements, or you can drag the mouse around each of the elements you want marked using the "magic wand". Remember that if you click another area or element without holding down the Ctrl key, all previously marked elements are cleared.

You can generate Select and Derive nodes from the marked elements in your graph. See the topic “Generating Nodes from Graphs” on page 252 for more information.

To mark elements:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Mark Elements button.
3. Either click on the element you require, or click and drag your mouse to draw a line around the region containing multiple elements.

Generating Nodes from Graphs

One of the most powerful features offered by IBM SPSS Modeler graphs is the ability to generate nodes from a graph or a selection within the graph. For example, from a time plot graph, you can generate Derive and Select nodes based on a selection or region of data, effectively "subsetting" the data. For example, you might use this powerful feature to identify and exclude outliers.

Whenever you can draw a band, you can also generate a Derive node. In graphs with two scale axes, you can generate Derive or Select nodes from the regions drawn in your graph. In graphs with marked elements, you can generate Derive nodes, Select nodes, and in some cases Filter nodes from these elements. Balance node generation is enabled for any graph showing a distribution of counts.

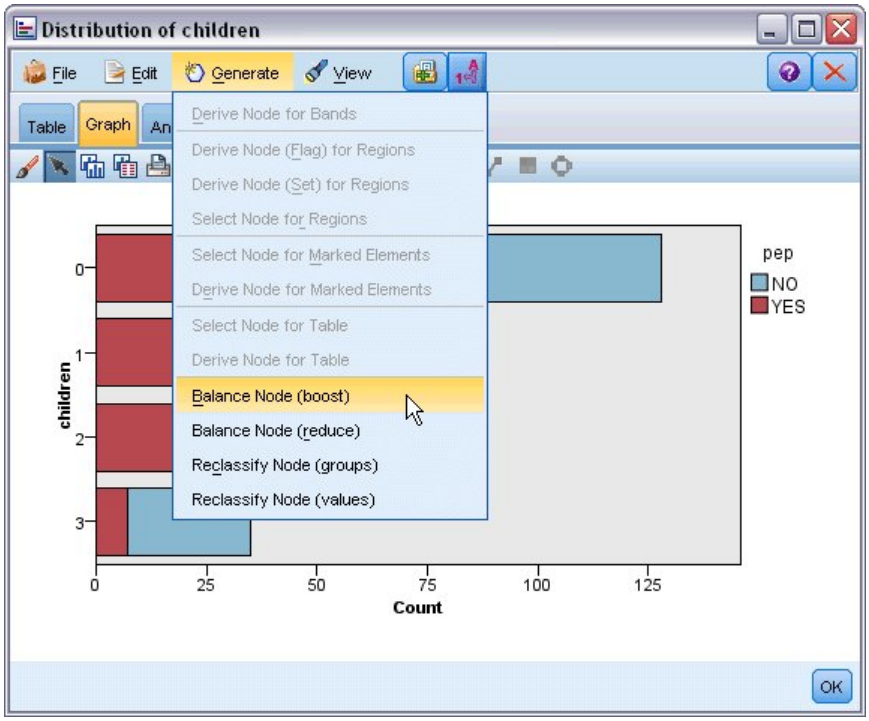


Figure 70. Graph with the generate menu showing

Whenever you generate a node, it is placed on the stream canvas directly so that you can connect it to an existing stream. The following nodes can be generated from graphs: Select, Derive, Balance, Filter, and Reclassify.

Select Nodes

Select nodes can be generated to test for inclusion of the records within a region and exclusion of all records falling outside the region or the reverse for downstream processing.

- **For bands**. You can generate a Select node that includes or excludes the records within that band. **Select node for Bands only** is only available through contextual menus since you need to select which band to use in the Select node.
- **For regions**. You can generate a Select node that includes or excludes the records within a region.
- **For marked elements**. You can generate Select nodes to capture the records corresponding to the marked elements or web graph links.

Derive Nodes

Derive nodes can be generated from regions, bands, and marked elements. All graphs can produce Derive nodes. In the case of evaluation charts, a dialog box for selecting the model appears. In the case of web graphs, **Derive Node ("And")** and **Derive Node ("Or")** are possible.

- **For bands**. You can generate a Derive node that produces a category for each interval marked on the axis, using the band names listed in the Edit Bands dialog box as category names.
- **For regions**. You can generate a Derive node (**Derive as flag**) that creates a flag field called *in_region* with the flags set to *T* for records inside any region and *F* for records outside all regions. You can also generate a Derive node (**Derive as set**) that produces a set with a value for each region with a new field called *region* for each record, which takes as its value the name of the region into which the records fall. Records falling outside all regions receive the name of the default region. Value names become the region names listed in the Edit regions dialog box.
- **For marked elements**. You can generate a Derive node that calculates a flag that is *True* for all marked elements and *False* for all other records.

Balance Nodes

Balance nodes can be generated to correct imbalances in the data, such as reducing the frequency of common values (use **Balance Node (reduce)** menu option) or boosting the occurrence of infrequent values (use **Balance Node (boost)** menu option). Balance node generation is enabled for any graph showing a distribution of counts, such as Histogram, Dot, Collection, Bar of Counts, Pie of Counts, and Multiplot.

Filter Nodes

Filter nodes can be generated to rename or filter fields based on the lines or nodes marked in the graph. In the case of evaluation charts, the best fit line does not generate a filter node.

Reclassify Nodes

Reclassify nodes can be generated to recode values. This option is used for distribution graphs. You can generate a Reclassify node for **groups** to recode specific values of a displayed field depending upon their inclusion in a group (select groups using Ctrl+click on the **Tables** tab). You can also generate a reclassify node for **values** to recode data into an existing set of numerous values, such as reclassifying data into a standard set of values in order to merge financial data from various companies for analysis.

Note: If the values are predefined, you can read them into IBM SPSS Modeler as a flat file and use a distribution to display all values. Then generate a Reclassify (values) node for this field directly from the chart. Doing so will put all the target values in the Reclassify node's *New values* column (drop-down list).

When setting options for the Reclassify node, the table enables a clear mapping from old set values to new values you specify:

- **Original value.** This column lists existing values for the select field(s).
- **New value.** Use this column to type new category values or select one from the drop-down list. When you automatically generate a Reclassify node using values from a Distribution chart, these values are included in the drop-down list. This allows you to quickly map existing values to a known set of values. For example, healthcare organizations sometimes group diagnoses differently based upon network or locale. After a merger or acquisition, all parties will be required to reclassify new or even existing data in a consistent fashion. Rather than manually typing each target value from a lengthy list, you can read the master list of values in to IBM SPSS Modeler, run a Distribution chart for the *Diagnosis* field, and generate a Reclassify (values) node for this field directly from the chart. This process will make all of the target Diagnosis values available from the New Values drop-down list.

For more information about the Reclassify Node, see "Setting Options for the Reclassify Node" on page 144.

Generating Nodes from Graphs

You can use the Generate menu in the graph output window to generate nodes. The generated node will be placed on the stream canvas. To use the node, connect it to an existing stream.

To generate a node from a graph:

1. Verify that you are in Explore mode. From the menus, choose **View > Explore Mode**.
2. In the Explore mode toolbar, click the Region button.
3. Define bands, regions, or any marked elements needed to generate your node.
4. From the Generate menu, choose the kind of node you want to produce. Only those which are possible are enabled.

Note: Alternatively, you can also generate nodes directly from the graph by right-clicking and choosing the generate option you want from the context menus.

Editing Visualizations

While Explore mode allows you to analytically explore the data and values represented by the visualization, Edit mode allows you to change the visualization's layout and look. For example, you can change the fonts and colors to match your organization's style guide. To select this mode, choose **View > Edit Mode** from the menus (or click the toolbar icon).

In Edit mode, there are several toolbars that affect different aspects of the visualization's layout. If you find that there are any you don't use, you can hide them to increase the amount of space in the dialog box in which the graph is displayed. To select or deselect toolbars, click the relevant toolbar name on the View menu.

Note: To add further detail to your visualizations, you can apply title, footnote, and axis labels. See the topic "Adding Titles and Footnotes" on page 264 for more information.

You have several options for editing a visualization in **Edit mode**. You can:

- Edit text and format it.
- Change the fill color, transparency, and pattern of frames and graphic elements.
- Change the color and dashing of borders and lines.
- Rotate and change the shape and aspect ratio of point elements.
- Change the size of graphic elements (such as bars and points).
- Adjust the space around items by using margins and padding.
- Specify formatting for numbers.
- Change the axis and scale settings.
- Sort, exclude, and collapse categories on a categorical axis.
- Set the orientation of panels.
- Apply transformations to a coordinate system.
- Change statistics, graphic element types, and collision modifiers.
- Change the position of the legend.
- Apply visualization stylesheets.

The following topics describe how to perform these various tasks. It is also recommended that you read the general rules for editing graphs.

How to Switch to Edit Mode

From the menus choose:

View > **Edit Mode**

General Rules for Editing Visualizations

Edit Mode

All edits are done in Edit mode. To enable Edit mode, from the menus choose:

View > **Edit Mode**

Selection

The options available for editing depend on selection. Different toolbar and properties palette options are enabled depending on what is selected. Only the enabled items apply to the current selection. For example, if an axis is selected, the Scale, Major Ticks, and Minor Ticks tabs are available in the properties palette.

Here are some tips for selecting items in the visualization:

- Click an item to select it.
- Select a graphic element (such as points in a scatterplot or bars in a bar chart) with a single click. After initial selection, click again to narrow the selection to groups of graphic elements or a single graphic element.
- Press Esc to deselect everything.

Palettes

When an item is selected in the visualization, the various palettes are updated to reflect the selection. The palettes contain controls for making edits to the selection. Palettes may be toolbars or a panel with multiple controls and tabs. Palettes can be hidden, so ensure the necessary palette is displayed for making edits. Check the View menu for palettes that are currently displayed.

You can reposition the palettes by clicking and dragging the empty space in a toolbar palette or the left side of other palettes. Visual feedback lets you know where you can dock the palette. For non-toolbar palettes, you can also click the close button to hide the palette and the undock button to display the palette in a separate window. Click the help button to display help for the specific palette.

Automatic Settings

Some settings provide an **-auto-** option. This indicates that automatic values are applied. Which automatic settings are used depends on the specific visualization and data values. You can enter a value to override the automatic setting. If you want to restore the automatic setting, delete the current value and press Enter. The setting will display **-auto-** again.

Removing/Hiding Items

You can remove/hide various items in the visualization. For example, you can hide the legend or axis label. To delete an item, select it and press Delete. If the item does not allow deletion, nothing will happen. If you accidentally delete an item, press Ctrl+Z to undo the deletion.

State

Some toolbars reflect the state of the current selection, others don't. The properties palette always reflects state. If a toolbar does *not* reflect state, this is mentioned in the topic that describes the toolbar.

Editing and Formatting Text

You can edit text in place and change the formatting of an entire text block. Note that you can't edit text that is linked directly to data values. For example, you can't edit a tick label because the content of the label is derived from the underlying data. However, you can format any text in the visualization.

How to Edit Text in Place

1. Double-click the text block. This action selects all the text. All toolbars are disabled at this time, because you cannot change any other part of the visualization while editing text.
2. Type to replace the existing text. You can also click the text again to display a cursor. Position the cursor where you want and enter the additional text.

How to Format Text

1. Select the frame containing the text. Do not double-click the text.
2. Format text using the font toolbar. If the toolbar is not enabled, make sure only the *frame* containing the text is selected. If the text itself is selected, the toolbar will be disabled.

You can change the font:

- Color
- Family (for example, Arial or Verdana)
- Size (the unit is pt unless you indicate a different unit, such as pc)
- Weight
- Alignment relative to the text frame

Formatting applies to all the text in a frame. You can't change the formatting of individual letters or words in any particular block of text.

Changing Colors, Patterns, Dashings, and Transparency

Many different items in a visualization have a fill and border. The most obvious example is a bar in a bar chart. The color of the bars is the fill color. They may also have a solid, black border around them.

There are other less obvious items in the visualization that have fill colors. If the fill color is transparent, you may not know there is a fill. For example, consider the text in an axis label. It appears as if this text is "floating" text, but it actually appears in a frame that has a transparent fill color. You can see the frame by selecting the axis label.

Any frame in the visualization can have a fill and border style, including the frame around the whole visualization. Also, any fill has an associated opacity/transparency level that can be adjusted.

How to Change the Colors, Patterns, Dashing, and Transparency

1. Select the item you want to format. For example, select the bars in a bar chart or a frame containing text. If the visualization is split by a categorical variable or field, you can also select the group that corresponds to an individual category. This allows you to change the default aesthetic assigned to that group. For example, you can change the color of one of the stacking groups in a stacked bar chart.
2. To change the fill color, the border color, or the fill pattern, use the color toolbar.

   Note: This toolbar does not reflect the state of the current selection.

   To change a color or fill, you can click the button to select the displayed option or click the drop-down arrow to choose another option. For colors, notice there is one color that looks like white with a red, diagonal line through it. This is the transparent color. You could use this, for example, to hide the borders on bars in a histogram.

   - The first button controls the fill color. If the color is associated with a continuous or ordinal field, this button changes the fill color for the color associated with the highest value in the data. You can use the Color tab on the properties palette to change the color associated with the lowest value and missing data. The color of the elements will change incrementally from the Low color to the High color as the values of the underlying data increase.
   - The second button controls the border color.
   - The third button controls the fill pattern. The fill pattern uses the border color. Therefore, the fill pattern is visible only if there is a visible border color.
   - The fourth control is a slider and text box that control the opacity of the fill color and pattern. A lower percentage means less opacity and more transparency. 100% is fully opaque (no transparency).
3. To change the dashing of a border or line, use the line toolbar.

   Note: This toolbar does not reflect the state of the current selection.

   As with the other toolbar, you can click the button to select the displayed option or click the drop-down arrow to choose another option.

Rotating and Changing the Shape and Aspect Ratio of Point Elements

You can rotate point elements, assign a different predefined shape, or change the aspect ratio (the ratio of width to height).

How to Modify Point Elements

1. Select the point elements. You cannot rotate or change the shape and aspect ratio of individual point elements.
2. Use the symbol toolbar to modify the points.
   - The first button allows you to change the shape of the points. Click the drop-down arrow and select a predefined shape.
   - The second button allows you to rotate the points to a specific compass position. Click the drop-down arrow and then drag the needle to the position you want.
   - The third button allows you to change the aspect ratio. Click the drop-down arrow and then click and drag the rectangle that appears. The shape of the rectangle represents the aspect ratio.

Changing the size of graphic elements

You can change the size of the graphic elements in the visualization. These include bars, lines, and points among others. If the graphic element is sized by a variable or field, the specified size is the *minimum* size.

How to change the size of the graphic elements

1. Select the graphic elements you want to resize.
2. Use the slider to change the size.

Specifying Margins and Padding

If there is too much or too little spacing around or inside a frame in the visualization, you can change its margin and padding settings. The **margin** is the amount of space between the frame and other items around it. The **padding** is the amount of space between the border of the frame and the *contents* of the frame.

How to Specify Margins and Padding

1. Select the frame for which you want to specify margins and padding. This can be a text frame, the frame around the legend, or even the data frame displaying the graphic elements (such as bars and points).
2. Use the Margins tab on the properties palette to specify the settings. All sizes are in pixels unless you indicate a different unit (such as cm or in).

Formatting Numbers

You can specify the format for numbers in tick labels on a continuous axis or data value labels displaying a number. For example, you can specify that numbers displayed in the tick labels are shown in thousands.

How to Specify Number Formats

1. Select the continuous axis tick labels or the data value labels if they contain numbers.
2. Click the **Format** tab on the properties palette.
3. Select the number formatting options you want:

   Prefix. A character to display at the beginning of the number. For example, enter a dollar sign ($) if the numbers are salaries in U.S. dollars.

   Suffix. A character to display at the end of the number. For example, enter a percentage sign (%) if the numbers are percentages.

   Min. Integer Digits. Minimum number of digits to display in the integer part of a decimal representation. If the actual value does not contain the minimum number of digits, the integer part of the value will be padded with zeros.

   Max. Integer Digits. Maximum number of digits to display in the integer part of a decimal representation. If the actual value exceeds the maximum number of digits, the integer part of the value will be replaced with asterisks.

   Min. Decimal Digits. Minimum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value does not contain the minimum number of digits, the decimal part of the value will be padded with zeros.

   Max. Decimal Digits. Maximum number of digits to display in the decimal part of a decimal or scientific representation. If the actual value exceeds the maximum number of digits, the decimal is rounded to the appropriate number of digits.

   Scientific. Whether to display numbers in scientific notation. Scientific notation is useful for very large or very small numbers. **-auto-** lets the application determine when scientific notation is appropriate.

   Scaling. A scale factor, which is a number by which the original value is divided. Use a scale factor when the numbers are large, but you don't want the label to extend too much to accommodate the

number. If you change the number format of the tick labels, be sure to edit the axis title to indicate how the number should be interpreted. For example, assume your scale axis displays salaries and the labels are 30,000, 50,000, and 70,000. You might enter a scale factor of `1000` to display 30, 50, and 70. You should then edit the scale axis title to include the text in thousands.

Parentheses for -ve. Whether parentheses should be displayed around negative values.

Grouping. Whether to display a character between groups of digits. Your computer's current locale determines which character is used for digit grouping.

Changing the Axis and Scale Settings

There are several options for modifying axes and scales.

How to change axis and scale settings

1. Select any part of the axis (for example, the axis label or tick labels).
2. Use the Scale, Major Ticks, and Minor Ticks tabs on the properties palette to change the axis and scale settings.

Scale tab

Note: The Scale tab does not appear for graphs where the data is pre-aggregated (for example, histograms).

Type. Specifies whether the scale is linear or transformed. Scale transformations help you understand the data or make assumptions necessary for statistical inference. On scatterplots, you might use a transformed scale if the relationship between the independent and dependent variables or fields is nonlinear. Scale transformations can also be used to make a skewed histogram more symmetric so that it resembles a normal distribution. Note that you are transforming only the scale on which the data are displayed; you are not transforming the actual data.

- **linear.** Specifies a linear, untransformed scale.
- **log.** Specifies a base-10 log transformed scale. To accommodate zero and negative values, this transformation uses a modified version of the log function. This "safe log" function is defined as `sign(x) * log(1 + abs(x))`. So `safeLog(-99)` equals:

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```

- **power.** Specifies a power transformed scale, using an exponent of 0.5. To accommodate negative values, this transformation uses a modified version of the power function. This "safe power" function is defined as `sign(x) * pow(abs(x), 0.5)`. So `safePower(-100)` equals:

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

Min/Max/Nice Low/Nice High. Specifies the range for the scale. Selecting **Nice Low** and **Nice High** allows the application to select an appropriate scale based on the data. The minimum and maximum are "nice" because they are typically whole values greater or less than the maximum and minimum data values. For example, if the data range from 4 to 92, a nice low and high for scale may be 0 and 100 rather than the actual data minimum and maximum. Be careful that you don't set a range that is too small and hides important items. Also note that you cannot set an explicit minimum and maximum if the **Include zero** option is selected.

Low Margin/High Margin. Create margins at the low and/or high end of the axis. The margin appears perpendicular to the selected axis. The unit is pixels unless you indicate a different unit (such as cm or in). For example, if you set the **High Margin** to 5 for the vertical axis, a horizontal margin of 5 px runs along the top of the data frame.

Reverse. Specifies whether the scale is reversed.

Include zero. Indicates that the scale should include 0. This option is commonly used for bar charts to ensure the bars begin at 0, rather than a value near the height of the smallest bar. If this option is selected, **Min** and **Max** are disabled because you cannot set a custom minimum and maximum for the scale range.

Major Ticks/Minor Ticks Tabs

Ticks or **tick marks** are the lines that appear on an axis. These indicate values at specific intervals or categories. **Major ticks** are the tick marks with labels. These are also longer than other tick marks. **Minor ticks** are tick marks that appear between the major tick marks. Some options are specific to the tick type, but most options are available for major and minor ticks.

Show ticks. Specifies whether major or minor ticks are displayed on a graph.

Show gridlines. Specifies whether gridlines are displayed at the major or minor ticks. **Gridlines** are lines that cross a whole graph from axis to axis.

Position. Specifies the position of the tick marks relative to the axis.

Length. Specifies the length of the tick marks. The unit is pixels unless you indicate a different unit (such as cm or in).

Base. *Applies only to major ticks.* Specifies the value at which the first major tick appears.

Delta. *Applies only to major ticks.* Specifies the difference between major ticks. That is, major ticks will appear at every *n*th value, where *n* is the delta value.

Divisions. *Applies only to minor ticks.* Specifies the number of minor tick divisions between major ticks. The number of minor ticks is one less than the number of divisions. For example, assume that there are major ticks at 0 and 100. If you enter 2 as the number of minor tick divisions, there will be *one* minor tick at 50, dividing the 0–100 range and creating *two* divisions.

Editing Categories

You can edit the categories on a categorical axis in several ways:

- Change the sort order for displaying the categories.
- Exclude specific categories.
- Add a category that does not appear in the data set.
- Collapse/combine small categories into one category.

How to Change the Sort Order of Categories

1. Select a categorical axis. The Categories palette displays the categories on the axis.

   Note: If the palette is not visible, make sure that you have it enabled. From the View menu in IBM SPSS Modeler, choose **Categories**.
2. In the Categories palette, select a sorting option from the drop-down list:

   Custom. Sort categories based on the order in which they appear in the palette. Use the arrow buttons to move categories to the top of the list, up, down, and to the bottom of the list.

   Data. Sort categories based on the order in which they occur in the dataset.

   Name. Sort categories alphabetically, using the names as displayed in the palette. This may be either the value or label, depending on whether the toolbar button to display values and labels is selected.

   Value. Sort categories by the underlying data value, using the values displayed in parentheses in the palette. Only data sources with metadata (such as IBM SPSS Statistics data files) support this option.

   Statistic. Sort categories based on the calculated statistic for each category. Examples of statistics include counts, percentages, and means. This option is available only if a statistic is used in the graph.

How to Add a Category

By default, only categories that appear in the data set are available. You can add a category to the visualization if needed.

1. Select a categorical axis. The Categories palette displays the categories on the axis.

   Note: If the palette is not visible, make sure that you have it enabled. From the View menu in IBM SPSS Modeler, choose **Categories**.

2. In the Categories palette, click the add category button:

Figure 71. Add category button

3. In the Add a new category dialog box, enter a name for the category.
4. Click **OK**.

How to Exclude Specific Categories

1. Select a categorical axis. The Categories palette displays the categories on the axis.

   Note: If the palette is not visible, make sure that you have it enabled. From the View menu in IBM SPSS Modeler, choose **Categories**.
2. In the Categories palette, select a category name in the Include list, and then click the X button. To move the category back, select its name in the Excluded list, and then click the arrow to the right of the list.

How to Collapse/Combine Small Categories

You can combine categories that are so small you don't need to display them separately. For example, if you have a pie chart with many categories, consider collapsing categories with a percentage less than 10. Collapsing is available only for statistics that are additive. For example, you can't add means together because means are not additive. Therefore, combining/collapsing categories using a mean is not available.

1. Select a categorical axis. The Categories palette displays the categories on the axis.

   Note: If the palette is not visible, make sure that you have it enabled. From the View menu in IBM SPSS Modeler, choose **Categories**.
2. In the Categories palette, select **Collapse** and specify a percentage. Any categories whose percentage of the total is less than the specified number are combined into one category. The percentage is based on the statistic shown in the chart. Collapsing is available only for count-based and summation (sum) statistics.

Changing the Orientation Panels

If you are using panels in your visualization, you can change their orientation.

How to Change the Orientation of the Panels

1. Select any part of the visualization.
2. Click the **Panels** tab on the properties palette.
3. Select an option from **Layout**:

Table. Lays out panels like a table, in that there is a row or column assigned to every individual value.

Transposed. Lays out panels like a table, but also swaps the original rows and columns. This option is not the same as transposing the graph itself. Note that the *x* axis and the *y* axis are unchanged when you select this option.

List. Lays out panels like a list, in that each cell represents a combination of values. Columns and rows are no longer assigned to individual values. This option allows the panels to wrap if needed.

Transforming the Coordinate System

Many visualizations are displayed in a flat, rectangular coordinate system. You can transform the coordinate system as needed. For example, you can apply a polar transformation to the coordinate system, add oblique drop shadow effects, and transpose the axes. You can also undo any of these transformations if they are already applied to the current visualization. For example, a pie chart is drawn in a polar coordinate system. You can undo the polar transformation and display the pie chart as a single stacked bar in a rectangular coordinate system.

How to Transform the Coordinate System

1. Select the coordinate system that you want to transform. You select the coordinate system by selecting the frame around the individual graph.
2. Click the **Coordinates** tab on the properties palette.
3. Select the transformations that you want to apply to the coordinate system. You can also deselect a transformation to undo it.

   Transposed. Changing the orientation of the axes is called **transposing**. It is similar to swapping the vertical and horizontal axes in a 2-D visualization.

   Polar. A polar transformation draws the graphic elements at a specific angle and distance from the center of the graph. A pie chart is a 1-D visualization with a polar transformation that draws the individual bars at specific angles. A radar chart is a 2-D visualization with a polar transformation that draws graphic elements at specific angles and distances from the center of the graph. A 3-D visualization would also include an additional depth dimension.

   Oblique. An oblique transformation adds a 3-D effect to the graphic elements. This transformation adds depth to the graphic elements, but the depth is purely decorative. It is not influenced by particular data values.

   Same Ratio. Applying the same ratio specifies that the same distance on each scale represents the same difference in data values. For example, 2cm on both scales represent a difference of 1000.

   Pre-transform inset %. If axes are clipped after the transformation, you may want to add insets to the graph before applying the transformation. The insets shrink the dimensions by a certain percentage before any transformations are applied to the coordinate system. You have control over the lower *x*, upper *x*, lower *y*, and upper *y* dimensions, in that order.

   Post-transform inset %. If you want to change the aspect ratio of the graph, you can add insets to the graph after applying the transformation. The insets shrink the dimensions by a certain percentage after any transformations are applied to the coordinate system. These insets can also be applied even if no transformation is applied to the graph. You have control over the lower *x*, upper *x*, lower *y*, and upper *y* dimensions, in that order.

Changing Statistics and Graphic Elements

You can convert a graphic element to another type, change the statistic used to draw the graphic element, or specify the collision modifier that determines what happens when graphic elements overlap.

How to Convert a Graphic Element

1. Select the graphic element that you want to convert.
2. Click the **Element** tab on the properties palette.
3. Select a new graphic element type from the Type list.

Table 39. Graphic element types

Graphic Element Type	Description
Point	A marker identifying a specific data point. A point element is used in scatterplots and other related visualizations.
Interval	A rectangular shape drawn at a specific data value and filling the space between an origin and another data value. An interval element is used in bar charts and histograms.
Line	A line that connects data values.
Path	A line that connects data values in the order they appear in the dataset.

Table 39. Graphic element types (continued)	
Graphic Element Type	**Description**
Area	A line that connects data elements with the area between the line and an origin filled in.
Polygon	A multi-sided shape enclosing a data region. A polygon element could be used in a binned scatterplot or a map.
Schema	An element consisting of a box with whiskers and markers indicating outliers. A schema element is used for boxplots.

How to Change the Statistic

1. Select the graphic element whose statistic you want to change.
2. Click the **Element** tab on the properties palette.

How to Specify the Collision Modifier

The collision modifier determines what happens when graphic elements overlap.

1. Select the graphic element for which you want to specify the collision modifier.
2. Click the **Element** tab on the properties palette.
3. From the Modifier drop-down list, select a collision modifier. **-auto-** lets the application determine which collision modifier is appropriate for the graphic element type and statistic.

   Overlay. Draw graphic elements on top of each other when they have the same value.

   Stack. Stack graphic elements that would normally be superimposed when they have the same data values.

   Dodge. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged symmetrically. That is, the graphic elements are moved to opposite sides of a central position. Dodging is very similar to clustering.

   Pile. Move graphic elements next to other graphic elements that appear at the same value, rather than superimposing them. The graphic elements are arranged asymmetrically. That is, the graphic elements are piled on top of one another, with the graphic element on the bottom positioned at a specific value on the scale.

   Jitter (normal). Randomly reposition graphic elements at the same data value using a normal distribution.

   Jitter (uniform). Randomly reposition graphic elements at the same data value using a uniform distribution.

Changing the Position of the Legend

If a graph includes a legend, the legend is typically displayed to the right of a graph. You can change this position if needed.

How to Change the Legend Position

1. Select the legend.
2. Click the **Legend** tab on the properties palette.
3. Select a position.

Copying a Visualization and Visualization Data

The General palette includes buttons for copying the visualization and its data.

Figure 72. Copy visualization button

Copying the visualization. This action copies the visualization to the clipboard as an image. Multiple image formats are available. When you paste the image into another application, you can choose a "paste special" option to select one of the available image formats for pasting.

Figure 73. Copy visualization data button

Copying the visualization data. This action copies the underlying data that is used to draw the visualization. The data is copied to the clipboard as plain text or HTML-formatted text. When you paste the data into another application, you can choose a "paste special" option to choose one of these formats for pasting.

Graphboard Editor Keyboard Shortcuts

Table 40. Keyboard shortcuts

Shortcut Key	Function
Ctrl+Space	Toggle between Explore and Edit mode
Delete	Delete a visualization item
Ctrl+Z	Undo
Ctrl+Y	Redo
F2	Display outline for selecting items in the graph

Adding Titles and Footnotes

For all graph types you can add unique title, footnote, or axis labels to help identify what is shown in the graph.

Adding Titles to Graphs

1. From the menus, choose **Edit > Add Graph Title**. A text box containing **<TITLE>** is displayed above the graph.
2. Verify that you are in Edit mode. From the menus, choose **View > Edit Mode**.
3. Double-click the **<TITLE>** text.
4. Type the required title and press Return.

Adding Footnotes to Graphs

1. From the menus, choose **Edit > Add Graph Footnote**. A text box containing **<FOOTNOTE>** is displayed below the graph.
2. Verify that you are in Edit mode. From the menus, choose **View > Edit Mode**.
3. Double-click the **<FOOTNOTE>** text.
4. Type the required title and press Return.

Using Graph Stylesheets

Basic graph display information, such as colors, fonts, symbols, and line thickness, are controlled by a stylesheet. There is a default stylesheet supplied with IBM SPSS Modeler; however, you can make changes to it if you need. For example, you may have a corporate color scheme for presentations that you want used in your graphs. See the topic "Editing Visualizations" on page 254 for more information.

In the graph nodes, you can use the Edit mode to make style changes to the look of a graph. You can then use the **Edit > Styles** menu to save the changes as a stylesheet to apply to all graphs that you subsequently generate from the current graph node or as a new default stylesheet for all graphs that you produce using IBM SPSS Modeler.

There are five stylesheet options available from the **Styles** option on the Edit menu:

- **Switch Stylesheet.** This displays a list of different, stored, stylesheets that you can select to change the look of your graphs. See the topic "Applying stylesheets" on page 265 for more information.
- **Store Styles in Node.** This stores modifications to the selected graph's styles so that they are applied to any future graphs created from the same graph node in the current stream.
- **Store Styles as Default.** This stores modifications to the selected graph's styles so that they are applied to all future graphs created from any graph node in any stream. After selecting this option, you can use **Apply Default Styles** to change any other existing graphs to use the same styles.
- **Apply Default Styles.** This changes the selected graph's styles to those that are currently saved as the default styles.
- **Apply Original Styles.** This changes a graph's styles back to the ones supplied as the original default.

Applying stylesheets

You can apply a visualization stylesheet that specifies stylistic properties of the visualization. For example, the stylesheet can define fonts, dashings, and colors, among other options. To a certain extent, stylesheets provide a shortcut for edits that you would have to perform manually. Note, however, that a stylesheet is limited to *style* changes. Other changes such as the position of the legend or the scale range are not stored in the stylesheet.

How to Apply a Stylesheet

1. From the menus choose:

   Edit > **Styles** > **Switch Stylesheet**
2. Use the Switch Stylesheet dialog box to select a stylesheet.
3. Click **Apply** to apply the stylesheet to the visualization without closing the dialog. Click **OK** to apply the stylesheet and close the dialog box.

Switch/Select Stylesheet Dialog Box

The table at the top of the dialog box lists all of the visualization stylesheets that are currently available. Some stylesheets are pre-installed, while others may have been created in the IBM SPSS Visualization Designer (a separate product).

The bottom of the dialog box shows example visualizations with sample data. Select one of the stylesheets to apply its styles to the example visualizations. These examples can help you determine how the stylesheet will affect your actual visualization.

The dialog box also offers the following options.

Existing styles. By default, a stylesheet can overwrite all the styles in the visualization. You can change this behavior.

- **Overwrite all styles**. When applying the stylesheet, overwrite all styles in the visualization, including those styles modified in the visualization during the current editing session.

- **Preserve modified styles**. When applying the stylesheet, overwrite only those styles that were *not* modified in the visualization during the current editing session. Styles that were modified during the current editing session are preserved.

Manage. Manage visualization templates, stylesheets, and maps on your computer. You can import, export, rename, and delete visualization templates, stylesheets, and maps on your local machine. See the topic "Managing Templates, Stylesheets, and Map Files" on page 195 for more information.

Location. Change the location in which visualization templates, stylesheets, and maps are stored. The current location is listed to the right of the button. See the topic "Setting the Location of Templates, Stylesheets, and Maps" on page 195 for more information.

Printing, saving, copying, and exporting graphs

Each graph has a number of options that enable you to save or print the graph or export it to another format. Most of these options are available from the File menu. In addition, from the Edit menu, you can choose to copy the graph, the data within it, or the Microsoft Office Drawing Object for use in another application.

Printing

To print the graph, use the **Print** menu item or button. Before you print, you can use **Page Setup** and **Print Preview** to set print options and preview the output.

Saving graphs

To save the graph to an IBM SPSS Modeler output file (`*.cou`), choose **File > Save** or **File > Save As** from the menus.

or

To save the graph in the repository, choose **File > Store Output** from the menus.

Copying graphs

To copy the graph for use in another application, such as MS Word or MS PowerPoint, choose **Edit > Copy Graph** from the menus.

Copying data

To copy the data for use in another application, such as MS Excel or MS Word, choose **Edit > Copy Data** from the menus. By default, the data will be formatted as HTML. Use **Paste Special** in the other application to view other formatting options when pasting.

Copying Microsoft Office Graphic Object

You can copy a graph as a Microsoft Office Graphic Object and use it in Microsoft Office applications such as Excel or PowerPoint. To copy a graph, choose **Edit > Copy Microsoft Office Graphic Object** from the menus. The content will be copied to your clipboard, and will be in binary format by default. Use **Paste Special** in the Microsoft Office application to specify other formatting options when pasting.

Note that some content may not support this feature, in which case the **Copy Microsoft Office Graphic Object** menu option will be disabled. Also note that the appearance of the graph may be different after pasting into an Office application, but the graph data is the same.

There are six types of graph output that can be copied and pasted into Excel: Simple Bar, Stacked Bar, Simple Boxplot, Clustered Boxplot, Simple Scatter, and Grouped Scatter. If using Panel and Animation options for any of these graph types, the **Copy Microsoft Office Graphic Object** option will be disabled in SPSS Modeler. And for other settings such as Optional Aesthetics or Overlay, the option is partially supported. See the following table for details:

Table 41. Copy Microsoft Graphic Object support

Graph Output Template	Modeler Graph Node	Modeler Graph Type	Basic Setting	Optional Aesthetics	Overlay	Microsoft Graphic Object Copy Support	Comments
Simple Bar	Graphboard	Bar	Yes	No	N/A	Yes	
		Bar of Counts	Yes	No	N/A	Yes	
	Distribution	Bar	Yes	N/A	No	Yes	
Stacked Bar	Graphboard	Bar	Yes	Yes	N/A	Yes with limitation	Only Yes for the categorical variable in Optional Aesthetics.
		Bar of Counts	Yes	Yes	N/A	Yes with limitation	Only Yes for the categorical variable in Optional Aesthetics.
	Distribution	Bar	Yes	N/A	Yes	Yes	
Boxplot	Graphboard	Boxplot	Yes	No	N/A	Yes with limitation	Only Yes on Windows.
		Boxplot	Yes	Yes	N/A	No	
Clustered Boxplot	Graphboard	Clustered Boxplot	Yes	No	N/A	Yes with limitation	Only Yes on Windows.
		Clustered Boxplot	Yes	Yes	N/A	No	

Table 41. Copy Microsoft Graphic Object support (continued)

Graph Output Template	Modeler Graph Node	Modeler Graph Type	Basic Setting	Optional Aesthetics	Overlay	Microsoft Graphic Object Copy Support	Comments
Simple Scatter	Graphboard	Bubble Plot	Yes	No	N/A	Yes with limitation	Only Yes for the continuous variables in both X and Y fields, and the categorical variable in Sizes.
		Scatterplot	Yes	No	N/A	Yes with limitation	Only Yes for the continuous variables in both X and Y fields.
	Plot	Point	Yes	N/A	No	Yes with limitation	Only Yes for the continuous variables in both X and Y fields.
Grouped Scatter	Graphboard	Bubble Plot	Yes	Yes	N/A	No	
		Scatterplot	Yes	Yes	N/A	Yes with limitation	Only Yes for the continuous variables in both X and Y fields, and the categorical variable in Optional Aesthetics.
	Plot	Point	Yes	N/A	Yes	Yes with limitation	Only Yes for the continuous variables in both X and Y fields, and the categorical variable in the Overlay option.

Exporting graphs

The **Export Graph** option enables you to export the graph in one of the following formats: Bitmap (.bmp), JPEG (.jpg), PNG (.png), HTML (.html), PDF (.pdf), or ViZml document (.xml) for use in other applications.

Note: When the PDF option is selected, the graphs are exported as high resolution PDF files that are cropped to the size of the graphic.

To export graphs, choose **File > Export Graph** from the menus and then choose the format.

Exporting tables

The **Export Table** option enables you to export the table in one of the following formats: tab delimited (.tab), comma delimited (.csv), or HTML (.html).

To export tables, choose **File > Export Table** from the menus and then choose the format.

Chapter 6. Output Nodes

Overview of Output Nodes

Output nodes provide the means to obtain information about your data and models. They also provide a mechanism for exporting data in various formats to interface with your other software tools.

The following output nodes are available:

The Table node displays the data in table format, which can also be written to a file. This is useful anytime that you need to inspect your data values or export them in an easily readable form.

The Matrix node creates a table that shows relationships between fields. It is most commonly used to show the relationship between two symbolic fields, but it can also show relationships between flag fields or numeric fields.

The Analysis node evaluates predictive models' ability to generate accurate predictions. Analysis nodes perform various comparisons between predicted values and actual values for one or more model nuggets. They can also compare predictive models to each other.

The Data Audit node provides a comprehensive first look at the data, including summary statistics, histograms and distribution for each field, as well as information on outliers, missing values, and extremes. Results are displayed in an easy-to-read matrix that can be sorted and used to generate full-size graphs and data preparation nodes.

The Transform node allows you to select and visually preview the results of transformations before applying them to selected fields.

The Statistics node provides basic summary information about numeric fields. It calculates summary statistics for individual fields and correlations between fields.

The Means node compares the means between independent groups or between pairs of related fields to test whether a significant difference exists. For example, you could compare mean revenues before and after running a promotion or compare revenues from customers who did not receive the promotion with those who did.

The Report node creates formatted reports containing fixed text as well as data and other expressions derived from the data. You specify the format of the report using text templates to define the fixed text and data output constructions. You can provide custom text formatting by using HTML tags in the template and by setting options on the Output tab. You can include data values and other conditional output by using CLEM expressions in the template.

The Set Globals node scans the data and computes summary values that can be used in CLEM expressions. For example, you can use this node to compute statistics for a field called *age* and then use the overall mean of *age* in CLEM expressions by inserting the function `@GLOBAL_MEAN(age)`.

The Simulation Fitting node examines the statistical distribution of the data in each field and generates (or updates) a Simulation Generate node, with the best fitting distribution assigned to each field. The Simulation Generate node can then be used to generate simulated data.

The Simulation Evaluation node evaluates a specified predicted target field, and presents distribution and correlation information about the target field.

Managing output

The Output manager shows the charts, graphs, and tables generated during an IBM SPSS Modeler session. You can always reopen an output by double-clicking it in the manager—you do not have to rerun the corresponding stream or node.

To view the Output manager

Open the View menu and choose **Managers**. Click the **Outputs** tab.

From the Output manager, you can:

- Display existing output objects, such as histograms, evaluation charts, and tables.
- Rename output objects.
- Save output objects to disk or to the IBM SPSS Collaboration and Deployment Services Repository (if available).
- Add output files to the current project.
- Delete unsaved output objects from the current session.
- Open saved output objects or retrieve them from the IBM SPSS Collaboration and Deployment Services Repository (if available).

To access these options, right-click anywhere on the Outputs tab.

Viewing output

On-screen output is displayed in an output browser window. The output browser window has its own set of menus that allow you to print or save the output, or export it to another format. Note that specific options may vary depending on the type of output.

Printing, saving, and exporting data. More information is available as follows:

- To print the output, use the **Print** menu option or button. Before you print, you can use **Page Setup** and **Print Preview** to set print options and preview the output.
- To save the output to an IBM SPSS Modeler output file (`.cou`), choose **Save** or **Save As** from the File menu.
- To save the output in another format, such as text or HTML, choose **Export** from the File menu. See the topic "Exporting Output" on page 274 for more information.

  Note that you can only select these formats if the output contains data that can be sensibly exported in that way. For example, the contents of a decision tree could be exported as text, but the contents of a K-means model would not make sense as text.

- To save the output in a shared repository so that other users can view it using the IBM SPSS Collaboration and Deployment Services Deployment Portal, choose **Publish to Web** from the File menu. Note that this option requires a separate license for IBM SPSS Collaboration and Deployment Services.

Selecting cells and columns. The Edit menu contains various options for selecting, deselecting, and copying cells and columns, as appropriate for the current output type. See "Selecting cells and columns" on page 275 for more information.

Generating new nodes. The Generate menu enables you to generate new nodes based on the contents of the output browser. The options vary depending on the type of output and the items in the output that are currently selected. For details about the node-generation options for a particular type of output, see the documentation for that output.

Publish to Web

The Publish to Web feature enables you to publish certain types of stream output to a central shared IBM SPSS Collaboration and Deployment Services Repository that forms the basis of IBM SPSS Collaboration and Deployment Services. If you use this option, other users who need to view this output can do so by using Internet access and an IBM SPSS Collaboration and Deployment Services account--they do not need to have IBM SPSS Modeler installed.

The following table lists the IBM SPSS Modeler nodes that support the Publish to Web feature. Output from these nodes is stored in the IBM SPSS Collaboration and Deployment Services Repository in output object (*.cou*) format, and can be viewed directly in the IBM SPSS Collaboration and Deployment Services Deployment Portal.

Other types of output can be viewed only if the relevant application (e.g. IBM SPSS Modeler, for stream objects) is installed on the user's machine.

Table 42. Nodes supporting Publish to Web

Node Type	**Node**
Graphs	all
Output	Table
	Matrix
	Data Audit
	Transform
	Means
	Analysis
	Statistics
	Report (HTML)
IBM SPSS Statistics	Statistics Output

Publishing Output to the Web

To publish output to the Web:

1. In an IBM SPSS Modeler stream, execute one of the nodes listed in the table. Doing so creates an output object (for example, a table, matrix or report object) in a new window.
2. From the output object window, choose:

   File > **Publish to Web**

   Note: If you just want to export simple HTML files for use with a standard Web browser, choose **Export** from the File menu and select **HTML**.

3. Connect to the IBM SPSS Collaboration and Deployment Services Repository.

   When you have connected successfully, the Repository: Store dialog is displayed, offering a number of storage options.
4. When you have chosen the storage options you want, click **Store**.

Viewing Published Output Over the Web

You must have an IBM SPSS Collaboration and Deployment Services account set up in order to use this feature. If you have the relevant application installed for the object type you want to view (for example, IBM SPSS Modeler or IBM SPSS Statistics), the output is displayed in the application itself rather than in the browser.

To view published output over the Web:

1. Point your browser to *http://<repos_host>:<repos_port>/peb*

   where *repos_host* and *repos_port* are the hostname and port number for the IBM SPSS Collaboration and Deployment Services host.
2. Enter the login details for your IBM SPSS Collaboration and Deployment Services account.
3. Click on **Content Repository**.
4. Navigate to, or search for, the object you want to view.
5. Click on the object name. For some object types, such as graphs, there may be a delay while the object is rendered in the browser.

Viewing Output in an HTML Browser

From the Advanced tab on the Linear, Logistic, and PCA/Factor model nuggets, you can view the displayed information in a separate browser, such as Internet Explorer. The information is output as HTML, enabling you to save it and reuse it elsewhere, such as on a corporate intranet, or Internet site.

To display the information in a browser, click the launch button, situated below the model icon in the top left of the Advanced tab of the model nugget.

Exporting Output

In the output browser window, you may choose to export the output to another format, such as text or HTML. The export formats vary depending on the type of output, but in general are similar to the file type options available if you select **Save to file** in the node used to generate the output.

Note: You can only select these formats if the output contains data that can be sensibly exported in that way. For example, the contents of a decision tree could be exported as text, but the contents of a K-means model would not make sense as text.

To Export Output

1. In the output browser, open the File menu and choose **Export**. Then choose the file type that you want to create:
   - **Tab Delimited (*.tab).** This option generates a formatted text file containing the data values. This style is often useful for generating a plain-text representation of the information that can be imported into other applications. This option is available for the Table, Matrix, and Means nodes.
   - **Comma Delimited (*.dat).** This option generates a comma-delimited text file containing the data values. This style is often useful as a quick way to generate a data file that can be imported into spreadsheets or other data analysis applications. This option is available for the Table, Matrix, and Means nodes.
   - **Transposed Tab Delimited (*.tab).** This option is identical to the Tab Delimited option, but the data is transposed so that rows represent fields and the columns represent records.
   - **Transposed Comma Delimited (*.dat).** This option is identical to the Comma Delimited option, but the data is transposed so that rows represent fields and the columns represent records.

- **HTML (*.html).** This option writes HTML-formatted output to a file or files.

Selecting cells and columns

A number of nodes, including the Table node, Matrix node, and Means node, generate tabular output. These output tables can be viewed and manipulated in similar ways, including selecting cells, copying all or part of the table to the Clipboard, generating new nodes based on the current selection, and saving and printing the table.

Selecting cells. To select a cell, click it. To select a rectangular range of cells, click one corner of the desired range, drag the mouse to the other corner of the range, and release the mouse button. To select an entire column, click the column heading. To select multiple columns, use Shift-click or Ctrl-click on column headings.

When you make a new selection, the old selection is cleared. By holding down the Ctrl key while selecting, you can add the new selection to any existing selection instead of clearing the old selection. You can use this method to select multiple, noncontiguous regions of the table. The Edit menu also contains the **Select All** and **Clear Selection** options.

Reordering columns. The Table node and Means node output browsers allow you to move columns in the table by clicking a column heading and dragging it to the desired location. You can move only one column at a time.

Table node

The Table node creates a table that lists the values in your data. All fields and all values in the stream are included, making this an easy way to inspect your data values or export them in an easily readable form. Optionally, you can highlight records that meet a certain condition.

Note: Unless you are working with small datasets, it is recommended that you select a subset of the data to pass into the Table node. The Table node cannot display properly when the number of records surpasses a size that can be contained in the display structure (for example, 100 million rows).

Table Node Settings Tab

Highlight records where. You can highlight records in the table by entering a CLEM expression that is true for the records to be highlighted. This option is enabled only when **Output to screen** is selected.

Table Node Format Tab

The Format tab contains options used to specify formatting on a per-field basis. This tab is shared with the Type node. See the topic “Field Format Settings Tab” on page 132 for more information.

Output Node Output Tab

For nodes that generate table-style output, the Output tab enables you to specify the format and location of the results.

Output name. Specifies the name of the output produced when the node is executed. **Auto** chooses a name based on the node that generates the output. Optionally, you can select **Custom** to specify a different name.

Output to screen (the default)**.** Creates an output object to view online. The output object will appear on the Outputs tab of the manager window when the output node is executed.

Output to file. Saves the output to a file when the node is executed. If you choose this option, enter a filename (or navigate to a directory and specify a filename using the File Chooser button) and select a file type. Note that some file types may be unavailable for certain types of output.

Notice:

Output data from output nodes is encoded according to the following rules:

- When executing an output node, the stream encoding value (which is set on the Stream Options tab) will be set to the output.
- After the output is generated, its encoding will not be changed even when the stream encoding is changed.
- When exporting output node output, the output file is exported with the currently defined stream encoding. After the output is created, even if you change the stream encoding, it will not impact the generated output.

Note the following exceptions to these rules:

- All HTML exports are encoded in UTF-8 format.
- Output from the Extension output node is generated by a custom user script. So the encoding is controlled by the script.

The following options are available for saving the output to a file:

- **Data (tab delimited) (*.tab).** This option generates a formatted text file containing the data values. This style is often useful for generating a plain-text representation of the information that can be imported into other applications. This option is available for the Table, Matrix, and Means nodes.
- **Data (comma delimited) (*.dat).** This option generates a comma-delimited text file containing the data values. This style is often useful as a quick way to generate a data file that can be imported into spreadsheets or other data analysis applications. This option is available for the Table, Matrix, and Means nodes.
- **HTML (*.html).** This option writes HTML-formatted output to a file or files. For tabular output (from the Table, Matrix, or Means nodes), a set of HTML files contains a contents panel listing field names and the data in an HTML table. The table may be split over multiple HTML files if the number of rows in the table exceeds the **Lines per page** specification. In this case, the contents panel contains links to all table pages and provides a means of navigating the table. For non-tabular output, a single HTML file is created containing the results of the node.

  Note: If the HTML output contains only formatting for the first page, select **Paginate output** and adjust the **Lines per page** specification to include all output on a single page. Or if the output template for nodes such as the Report node contains custom HTML tags, be sure you have specified **Custom** as the format type.
- **Text File (*.txt).** This option generates a text file containing the output. This style is often useful for generating output that can be imported into other applications, such as word processors or presentation software. This option is not available for some nodes.
- **Output object (*.cou).** Output objects saved in this format can be opened and viewed in IBM SPSS Modeler, added to projects, and published and tracked using the IBM SPSS Collaboration and Deployment Services Repository.

Output view. For the Means node, you can specify whether simple or advanced output is displayed by default. Note you can also toggle between these views when browsing the generated output. See the topic “Means Node Output Browser” on page 294 for more information.

Format. For the Report node, you can choose whether output is automatically formatted or formatted using HTML included in the template. Select **Custom** to allow HTML formatting in the template.

Title. For the Report node, you can specify optional title text that will appear at the top of the report output.

Highlight inserted text. For the Report node, select this option to highlight text generated by CLEM expressions in the Report template. See the topic “Report Node Template Tab” on page 296 for more information. This option is not recommended when using **Custom** formatting.

Lines per page. For the Report node, specify a number of lines to include on each page during **Auto** formatting of the output report.

Transpose data. This option transposes the data before export, so that rows represent fields and the columns represent records.

Note: For large tables, the above options can be somewhat inefficient, especially when working with a remote server. In such cases, using a File output node provides much better performance. See the topic "Flat File Export Node" on page 326 for more information.

Table Browser

The table browser displays tabular data and enables you to perform standard operations including selecting and copying cells, reordering columns, and saving and printing the table. See the topic "Selecting cells and columns" on page 275 for more information. These are the same operations that you can carry out when previewing the data in a node.

Exporting table data. You can export data from the table browser by choosing:

File > **Export**

See the topic "Exporting Output" on page 274 for more information.

Data is exported in the system default encoding format, which is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.

Searching the table. The search button (with the binoculars icon) on the main toolbar activates the search toolbar, enabling you to search the table for specific values. You can search forward or backward in the table, you can specify a case-sensitive search (the **Aa** button), and you can interrupt a search-in-progress with the interrupt search button.

Generating new nodes. The Generate menu contains node generation operations.

- **Select Node ("Records").** Generates a Select node that selects the records for which any cell in the table is selected.
- **Select ("And").** Generates a Select node that selects records containing *all* of the values selected in the table.
- **Select ("Or").** Generates a Select node that selects records containing *any* of the values selected in the table.
- **Derive ("Records").** Generates a Derive node to create a new flag field. The flag field contains *T* for records for which any cell in the table is selected and *F* for the remaining records.
- **Derive ("And").** Generates a Derive node to create a new flag field. The flag field contains *T* for records containing *all* of the values selected in the table and *F* for the remaining records.
- **Derive ("Or").** Generates a Derive node to create a new flag field. The flag field contains *T* for records containing *any* of the values selected in the table and *F* for the remaining records.

Matrix node

The Matrix node enables you to create a table that shows relationships between fields. It is most commonly used to show the relationship between two categorical fields (flag, nominal, or ordinal), but it can also be used to show relationships between continuous (numeric range) fields.

Matrix Node Settings Tab

The Settings tab enables you to specify options for the structure of the matrix.

Fields. Select a field selection type from the following options:

- **Selected.** This option enables you to select a categorical field for the rows and one for the columns of the matrix. The rows and columns of the matrix are defined by the list of values for the selected categorical field. The cells of the matrix contain the summary statistics selected below.
- **All flags (true values).** This option requests a matrix with one row and one column for each flag field in the data. The cells of the matrix contain the counts of double positives for each flag combination. In other words, for a row corresponding to *bought bread* and a column corresponding to *bought cheese*, the cell at the intersection of that row and column contains the number of records for which both *bought bread* and *bought cheese* are true.

- **All numerics.** This option requests a matrix with one row and one column for each numeric field. The cells of the matrix represent the sum of the cross-products for the corresponding pair of fields. In other words, for each cell in the matrix, the values for the row field and the column field are multiplied for each record and then summed across records.

Include missing values. Includes user-missing (blank) and system missing ($null$) values in the row and column output. For example, if the value *N/A* has been defined as user-missing for the selected column field, a separate column labeled *N/A* will be included in the table (assuming this value actually occurs in the data) just like any other category. If this option is deselected, the *N/A* column is excluded regardless of how often it occurs.

Note: The option to include missing values applies only when selected fields are cross-tabulated. Blank values are mapped to $null$ and are excluded from aggregation for the function field when the mode is **Selected** and the content is set to **Function** and for all numeric fields when the mode is set to **All Numerics**.

Cell contents. If you have chosen **Selected** fields above, you can specify the statistic to be used in the cells of the matrix. Select a count-based statistic, or select an overlay field to summarize values of a numeric field based on the values of the row and column fields.

- **Cross-tabulations.** Cell values are counts and/or percentages of how many records have the corresponding combination of values. You can specify which cross-tabulation summaries you want using the options on the Appearance tab. The global chi-square value is also displayed along with the significance. See the topic "Matrix node output browser" on page 279 for more information.
- **Function.** If you select a summary function, cell values are a function of the selected overlay field values for cases having the appropriate row and column values. For example, if the row field is *Region*, the column field is *Product*, and the overlay field is *Revenue*, then the cell in the *Northeast* row and the *Widget* column will contain the sum (or average, minimum, or maximum) of revenue for widgets sold in the northeast region. The default summary function is **Mean**. You can select another function for summarizing the function field. Options include **Mean**, **Sum**, **SDev** (standard deviation), **Max** (maximum), and **Min** (minimum).

Matrix Node Appearance Tab

The Appearance tab enables you to control sorting and highlighting options for the matrix, as well as statistics presented for cross-tabulation matrices.

Rows and columns. Controls the sorting of row and column headings in the matrix. The default is **Unsorted**. Select **Ascending** or **Descending** to sort row and column headings in the specified direction.

Overlay. Enables you to highlight extreme values in the matrix. Values are highlighted based on cell counts (for cross-tabulation matrices) or calculated values (for function matrices).

- **Highlight top.** You can request the highest values in the matrix to be highlighted (in red). Specify the number of values to highlight.
- **Highlight bottom.** You can also request the lowest values in the matrix to be highlighted (in green). Specify the number of values to highlight.

Note: For the two highlighting options, ties can cause more values than requested to be highlighted. For example, if you have a matrix with six zeros among the cells and you request **Highlight bottom 5**, all six zeros will be highlighted.

Cross-tabulation cell contents. For cross-tabulations, you can specify the summary statistics contained in the matrix for cross-tabulation matrices. These options are not available when either the **All Numerics** or **Function** option is selected on the Settings tab.

- **Counts.** Cells include the number of records with the row value that have the corresponding column value. This is the only default cell content.

- **Expected values.** The expected value for number of records in the cell, assuming that there is no relationship between the rows and columns. Expected values are based on the following formula:

```
p(row value) * p(column value) * total number of records
```

- **Residuals.** The difference between observed and expected values.
- **Percentage of row.** The percentage of all records with the row value that have the corresponding column value. Percentages sum to 100 within rows.
- **Percentage of column.** The percentage of all records with the column value that have the corresponding row value. Percentages sum to 100 within columns.
- **Percentage of total.** The percentage of all records having the combination of column value and row value. Percentages sum to 100 over the whole matrix.
- **Include row and column totals.** Adds a row and a column to the matrix for column and row totals.
- **Apply Settings.** (Output Browser only) Enables you to make changes to the appearance of the Matrix node output without having to close and reopen the Output Browser. Make the changes on this tab of the Output Browser, click this button and then select the Matrix tab to see the effect of the changes.

Matrix node output browser

The matrix browser displays cross-tabulated data and enables you to perform operations on the matrix, including selecting cells, copying the matrix to the Clipboard in whole or in part, generating new nodes based on the matrix selection, and saving and printing the matrix. The matrix browser may also be used to display output from certain models, such as Naive Bayes models from Oracle.

The File and Edit menus provide the usual options for printing, saving, exporting output, and for selecting and copying data. See the topic "Viewing output" on page 272 for more information.

Chi-square. For a cross-tabulation of two categorical fields, the global Pearson chi-square is also displayed below the table. This test indicates the probability that the two fields are unrelated, based on the difference between observed counts and the counts you would expect if no relationship exists. For example, if there were no relationship between customer satisfaction and store location, you would expect similar satisfaction rates across all stores. But if customers at certain stores consistently report higher rates than others, you might suspect it wasn't a coincidence. The greater the difference, the smaller the probability that it was the result of chance sampling error alone.

- The chi-square test indicates the probability that the two fields are unrelated, in which case any differences between observed and expected frequencies are the result of chance alone. If this probability is very small—typically less than 5%—then the relationship between the two fields is said to be significant.
- If there is only one column or one row (a one-way chi-square test), the degrees of freedom is the number of cells minus one. For a two-way chi-square, the degrees of freedom is the number of rows minus one times the number of columns minus one.
- Use caution when interpreting the chi-square statistic if any of the expected cell frequencies are less than five.
- The chi-square test is available only for a cross-tabulation of two fields. (When **All flags** or **All numerics** is selected on the Settings tab, this test is not displayed.)

Generate menu. The Generate menu contains node generation operations. These operations are available only for cross-tabulated matrices, and you must have at least one cell selected in the matrix.

- **Select Node.** Generates a Select node that selects the records that match any selected cell in the matrix.
- **Derive Node (Flag).** Generates a Derive node to create a new flag field. The flag field contains *T* for records that match any selected cell in the matrix and *F* for the remaining records.
- **Derive Node (Set).** Generates a Derive node to create a new nominal field. The nominal field contains one category for each contiguous set of selected cells in the matrix.

Analysis Node

The Analysis node enables you to evaluate the ability of a model to generate accurate predictions. Analysis nodes perform various comparisons between predicted values and actual values (your target field) for one or more model nuggets. Analysis nodes can also be used to compare predictive models to other predictive models.

When you execute an Analysis node, a summary of the analysis results is automatically added to the Analysis section on the Summary tab for each model nugget in the executed stream. The detailed analysis results appear on the Outputs tab of the manager window or can be written directly to a file.

Note: Because Analysis nodes compare predicted values to actual values, they are only useful with supervised models (those that require a target field). For unsupervised models such as clustering algorithms, there are no actual results available to use as a basis for comparison.

Analysis Node Analysis Tab

The Analysis tab enables you to specify the details of the analysis.

Coincidence matrices (for symbolic or categorical targets). Shows the pattern of matches between each generated (predicted) field and its target field for categorical targets (either flag, nominal, or ordinal). A table is displayed with rows defined by actual values and columns defined by predicted values, with the number of records having that pattern in each cell. This is useful for identifying systematic errors in prediction. If there is more than one generated field related to the same output field but produced by different models, the cases where these fields agree and disagree are counted and the totals are displayed. For the cases where they agree, another set of correct/wrong statistics is displayed.

Performance evaluation. Shows performance evaluation statistics for models with categorical outputs. This statistic, reported for each category of the output field(s), is a measure of the average information content (in bits) of the model for predicting records belonging to that category. It takes the difficulty of the classification problem into account, so accurate predictions for rare categories will earn a higher performance evaluation index than accurate predictions for common categories. If the model does no better than guessing for a category, the performance evaluation index for that category will be 0.

Evaluation metrics (AUC & Gini, binary classifiers only). For binary classifiers, this options reports the AUC (area under curve) and Gini coefficient evaluation metrics. Both of these evaluation metrics are calculated together for each binary model. The values of the metrics are reported in a table in the analysis output browser.

The AUC evaluation metric is calculated as the area under an ROC (receiver operator characteristic) curve, and is a scalar representation of the expected performance of a classifier. The AUC is always between 0 and 1, with a higher number representing a better classifier. A diagonal ROC curve between the coordinates (0,0) and (1,1) represents a random classifier, and has an AUC of 0.5. Therefore, a realistic classifier will not have and AUC of less than 0.5.

The Gini coefficient evaluation metric is sometimes used as an alternative evaluation metric to the AUC, and the two measures are closely related. The Gini coefficient is calculated as twice the area between the ROC curve and the diagonal, or as Gini = 2AUC - 1. The Gini coefficient is always between 0 and 1, with a higher number representing a better classifier. The Gini coefficient is negative in the unlikely event that the ROC curve is below the diagonal.

Confidence figures (if available). For models that generate a confidence field, this option reports statistics on the confidence values and their relationship to predictions. There are two settings for this option:

- **Threshold for.** Reports the confidence level above which the accuracy will be the specified percentage.
- **Improve accuracy.** Reports the confidence level above which the accuracy is improved by the specified factor. For example, if the overall accuracy is 90% and this option is set to 2.0, the reported value will be the confidence required for 95% accuracy.

Find predicted/predictor fields using. Determines how predicted fields are matched to the original target field.

- **Model output field metadata.** Matches predicted fields to the target based on model field information, allowing a match even if a predicted field has been renamed. Model field information can also be accessed for any predicted field from the Values dialog box using a Type node. See the topic "Using the Values Dialog Box" on page 127 for more information.
- **Field name format.** Matches fields based on the naming convention. For example predicted values generated by a C5.0 model nugget for a target named *response* must be in a field named *$C-response*.

Separate by partition. If a partition field is used to split records into training, test, and validation samples, select this option to display results separately for each partition. See the topic "Partition Node" on page 155 for more information.

Note: When separating by partition, records with null values in the partition field are excluded from the analysis. This will never be an issue if a Partition node is used, since Partition nodes do not generate null values.

User defined analysis. You can specify your own analysis calculation to be used in evaluating your model(s). Use CLEM expressions to specify what should be computed for each record and how to combine the record-level scores into an overall score. Use the functions **@TARGET** and **@PREDICTED** to refer to the target (actual output) value and the predicted value, respectively.

- **If.** Specify a conditional expression if you need to use different calculations depending on some condition.
- **Then.** Specify the calculation if the If condition is true.
- **Else.** Specify the calculation if the If condition is false.
- **Use.** Select a statistic to compute an overall score from the individual scores.

Break down analysis by fields. Shows the categorical fields available for breaking down the analysis. In addition to the overall analysis, a separate analysis will be reported for each category of each breakdown field.

Analysis Output Browser

The analysis output browser displays the results of executing the Analysis node. The usual saving, exporting, and printing options are available from the File menu. See the topic "Viewing output" on page 272 for more information.

When you first browse Analysis output, the results are expanded. To hide results after viewing them, use the expander control to the left of the item to collapse the specific results you want to hide or click the **Collapse All** button to collapse all results. To see results again after collapsing them, use the expander control to the left of the item to show the results or click the **Expand All** button to show all results.

Results for output field. The Analysis output contains a section for each output field for which there is a corresponding prediction field created by a generated model.

Comparing. Within the output field section is a subsection for each prediction field associated with that output field. For categorical output fields, the top level of this section contains a table showing the number and percentage of correct and incorrect predictions and the total number of records in the stream. For numeric output fields, this section shows the following information:

- **Minimum Error.** Shows the minimum error (difference between observed and predicted values).
- **Maximum Error.** Shows the maximum error.
- **Mean Error.** Shows the average (mean) of errors across all records. This indicates whether there is a systematic **bias** (a stronger tendency to overestimate than to underestimate, or vice versa) in the model.
- **Mean Absolute Error.** Shows the average of the absolute values of the errors across all records. Indicates the average magnitude of error, independent of the direction.
- **Standard Deviation.** Shows the standard deviation of the errors.
- **Linear Correlation.** Shows the linear correlation between the predicted and actual values. This statistic varies between –1.0 and 1.0. Values close to +1.0 indicate a strong positive association, so that high predicted values are associated with high actual values and low predicted values are associated with

low actual values. Values close to –1.0 indicate a strong negative association, so that high predicted values are associated with low actual values, and vice versa. Values close to 0.0 indicate a weak association, so that predicted values are more or less independent of actual values. *Note*: A blank entry here indicates that linear correlation cannot be computed in this case, because either the actual or predicted values are constant.

- **Occurrences.** Shows the number of records used in the analysis.

Coincidence Matrix. For categorical output fields, if you requested a coincidence matrix in the analysis options, a subsection appears here containing the matrix. The rows represent actual observed values, and the columns represent predicted values. The cell in the table indicates the number of records for each combination of predicted and actual values.

Performance Evaluation. For categorical output fields, if you requested performance evaluation statistics in the analysis options, the performance evaluation results appear here. Each output category is listed with its performance evaluation statistic.

Confidence Values Report. For categorical output fields, if you requested confidence values in the analysis options, the values appear here. The following statistics are reported for model confidence values:

- **Range.** Shows the range (smallest and largest values) of confidence values for records in the stream data.
- **Mean Correct.** Shows the average confidence for records that are classified correctly.
- **Mean Incorrect.** Shows the average confidence for records that are classified incorrectly.
- **Always Correct Above.** Shows the confidence threshold above which predictions are always correct and shows the percentage of cases meeting this criterion.
- **Always Incorrect Below.** Shows the confidence threshold below which predictions are always incorrect and shows the percentage of cases meeting this criterion.
- **X% Accuracy Above.** Shows the confidence level at which accuracy is *X%*. *X* is approximately the value specified for **Threshold for** in the Analysis options. For some models and datasets, it is not possible to choose a confidence value that gives the exact threshold specified in the options (usually due to clusters of similar cases with the same confidence value near the threshold). The threshold reported is the closest value to the specified accuracy criterion that can be obtained with a single confidence value threshold.
- **X Fold Correct Above.** Shows the confidence value at which accuracy is *X* times better than it is for the overall dataset. *X* is the value specified for **Improve accuracy** in the Analysis options.

Agreement between. If two or more generated models that predict the same output field are included in the stream, you will also see statistics on the **agreement** between predictions generated by the models. This includes the number and percentage of records for which the predictions agree (for categorical output fields) or error summary statistics (for continuous output fields). For categorical fields, it includes an analysis of predictions compared to actual values for the subset of records on which the models agree (generate the same predicted value).

Evaluation Metrics. For binary classifiers, if you requested evaluation metrics in the analysis options, the values of the AUC and Gini coefficient evaluation metrics are shown in a table in this section. The table has one row for each binary classifier model. The evaluation metrics table is shown for each output field, rather than for each model.

Data Audit node

The Data Audit node provides a comprehensive first look at the data you bring into IBM SPSS Modeler, presented in an easy-to-read matrix that can be sorted and used to generate full-size graphs and a variety of data preparation nodes.

- The Audit tab displays a report that provides summary statistics, histograms, and distribution graphs that may be useful in gaining a preliminary understanding of the data. The report also displays the storage icon before the field name.

- The Quality tab in the audit report displays information about outliers, extremes, and missing values, and offers tools for handling these values.

Using the Data Audit node

The Data Audit node can be attached directly to a source node or downstream from an instantiated Type node. You can also generate a number of data preparation nodes based on the results. For example, you can generate a Filter node that excludes fields with too many missing values to be useful in modeling, and generate a SuperNode that imputes missing values for any or all of the fields that remain. This is where the real power of the audit comes in, enabling you not only to assess the current state of your data, but to take action based on the assessment.

Screening or sampling the data. Because an initial audit is particularly effective when dealing with big data, a Sample node may be used to reduce processing time during the initial exploration by selecting only a subset of records. The Data Audit node can also be used in combination with nodes such as Feature Selection and Anomaly Detection in the exploratory stages of analysis.

Data Audit Node Settings Tab

The Settings tab enables you to specify basic parameters for the audit.

Default. You can simply attach the node to your stream and click **Run** to generate an audit report for all fields based on default settings, as follows:

- If there are no Type node settings, all fields are included in the report.
- If there are Type settings (regardless of whether or not they are instantiated), all *Input*, *Target*, and *Both* fields are included in the display. If there is a single *Target* field, use it as the Overlay field. If there is more than one *Target* field specified, no default overlay is specified.

Use custom fields. Select this option to manually select fields. Use the field chooser button on the right to select fields individually or by type.

Overlay field. The overlay field is used in drawing the thumbnail graphs shown in the audit report. In the case of a continuous (numeric range) field, bivariate statistics (covariance and correlation) are also calculated. If a single *Target* field is present based on Type node settings, it is used as the default overlay field as described above. Alternatively, you can select **Use custom fields** in order to specify an overlay.

Display. Enables you to specify whether graphs are available in the output, and to choose the statistics displayed by default.

- **Graphs.** Displays a graph for each selected field; either a distribution (bar) graph, histogram, or scatterplot as appropriate for the data. Graphs are displayed as thumbnails in the initial report, but full-sized graphs and graph nodes can also be generated. See the topic "Data Audit Output Browser" on page 284 for more information.
- **Basic/Advanced statistics.** Specifies the level of statistics displayed in the output by default. While this setting determines the initial display, all statistics are available in the output regardless of this setting. See the topic "Display Statistics" on page 285 for more information.

Median and mode. Calculates the median and mode for all fields in the report. Note that with large datasets, these statistics may increase processing time, since they take longer than others to compute. In the case of the median only, the reported value may be based on a sample of 2000 records (rather than the full dataset) in some cases. This sampling is done on a per-field basis in cases where memory limits would otherwise be exceeded. When sampling is in effect, the results will be labeled as such in the output (*Sample Median* rather than just *Median*). All statistics other than the median are always computed using the full dataset.

Empty or typeless fields. When used with instantiated data, typeless fields are not included in the audit report. To include typeless fields (including empty fields), select **Clear All Values** in any upstream Type nodes. This ensures that data are not instantiated, causing all fields to be included in the report. For example, this may be useful if you want to obtain a complete list of all fields or generate a Filter node that

will exclude those that are empty. See the topic "Filtering Fields with Missing Data" on page 287 for more information.

Data Audit Quality Tab

The Quality tab in the Data Audit node provides options for handling missing values, outliers, and extreme values.

Missing Values

- **Count of records with valid values.** Select this option to show the number of records with valid values for each evaluated field. Note that null (undefined) values, blank values, white spaces and empty strings are always treated as invalid values.
- **Breakdown counts of records with invalid values.** Select this option to show the number of records with each type of invalid value for each field.

Outliers and Extreme Values

Detection method for outliers and extreme values. Two methods are supported:

Standard deviation from the mean. Detects outliers and extremes based on the number of standard deviations from the mean. For example, if you have a field with a mean of 100 and a standard deviation of 10, you could specify 3.0 to indicate that any value below 70 or above 130 should be treated as an outlier.

Interquartile range. Detects outliers and extremes based on the interquartile range, which is the range within which the two central quartiles fall (between the 25th and 75th percentiles). For example, based on the default setting of 1.5, the lower threshold for outliers would be Q1 – 1.5 * IQR and the upper threshold would be Q3 + 1.5*IQR. Note that using this option may slow performance on large datasets.

Data Audit Output Browser

The Data Audit browser is a powerful tool for gaining overview of your data. The Audit tab displays thumbnail graphs, storage icons, and statistics for all fields, while the Quality tab displays information about outliers, extremes, and missing values. Based on the initial graphs and summary statistics, you might decide to recode a numeric field, derive a new field, or reclassify the values of a nominal field. Or you may want to explore further using more sophisticated visualization. You can do this right from the audit report browser using the Generate menu to create any number of nodes that can be used to transform or visualize the data.

- Sort columns by clicking on the column header, or reorder columns using drag and drop. Most standard output operations are also supported. See the topic "Viewing output" on page 272 for more information.
- View values and ranges for fields by double-clicking a field in the Measurement or Unique columns.
- Use the toolbar or Edit menu to show or hide value labels, or to choose the statistics you want to display. See the topic "Display Statistics" on page 285 for more information.
- Verify the storage icons to the left of the field names. Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. See the topic "Setting Field Storage and Formatting" on page 6 for more information.

Viewing and Generating Graphs

If no overlay is selected, the Audit tab displays either bar charts (for nominal or flag fields) or histograms (continuous fields).

For a nominal or flag field overlay, the graphs are colored by the values of the overlay.

For a continuous field overlay, two-dimensional scatterplots are generated rather than one-dimensional bars and histograms. In this case, the *x* axis maps to the overlay field, enabling you to see the same scale on all *x* axes as you read down the table.

- For Flag or Nominal fields, hold the mouse cursor over a bar to display the underlying value or label in a ToolTip.
- For Flag or Nominal fields, use the toolbar to toggle the orientation of thumbnail graphs from horizontal to vertical.
- To generate a full-sized graph from any thumbnail, double-click on the thumbnail, or select a thumbnail and choose **Graph Output** from the Generate menu. *Note*: If a thumbnail graph was based on sampled data, the generated graph will contain all cases if the original data stream is still open.

  You can only generate a graph if the Data Audit node that created the output is connected to the stream.
- To generate a matching graph node, select one or more fields on the Audit tab and choose **Graph Node** from the Generate menu. The resulting node is added to the stream canvas and can be used to re-create the graph each time the stream is run.
- If an overlay set has more than 100 values, a warning is raised and the overlay is not included.

Display Statistics

The Display Statistics dialog box enables you to choose the statistics displayed on the Audit tab. The initial settings are specified in the Data Audit node. See the topic "Data Audit Node Settings Tab" on page 283 for more information.

Minimum. The smallest value of a numeric variable.

Maximum. The largest value of a numeric variable.

Sum. The sum or total of the values, across all cases with nonmissing values.

Range. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Standard Error of Mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

standard deviation. A measure of dispersion around the mean, equal to the square root of the variance. The standard deviation is measured in the same units as the original variable.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

Skewness. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Standard Error of Skewness. The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Kurtosis. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution.

Standard Error of Kurtosis. The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for

kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Unique. Evaluates all effects simultaneously, adjusting each effect for all other effects of any type.

Valid. Valid cases having neither the system-missing value, nor a value defined as user-missing. Note that null (undefined) values, blank values, white spaces and empty strings are always treated as invalid values.

Median. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Mode. The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.

Note that median and mode are suppressed by default in order to improve performance but can be selected on the Settings tab in the Data Audit node. See the topic "Data Audit Node Settings Tab" on page 283 for more information.

Statistics for Overlays

If a continuous (numeric range) overlay field is in use, the following statistics are also available:

Covariance. An unstandardized measure of association between two variables, equal to the cross-product deviation divided by N-1.

Data Audit Browser Quality Tab

The Quality tab in the Data Audit browser displays the results of the data quality analysis and enables you to specify treatments for outliers, extremes, and missing values.

Imputing Missing Values

The audit report lists the percentage of complete records for each field, along with the number of valid, null, and blank values. You can choose to impute missing values for specific fields as appropriate, and then generate a SuperNode to apply these transformations.

1. In the **Impute Missing** column, specify the type of values you want to impute, if any. You can choose to impute blanks, nulls, both, or specify a custom condition or expression that selects the values to impute.

   There are several types of missing values recognized by IBM SPSS Modeler:

   - **Null or system-missing values.** These are nonstring values that have been left blank in the database or source file and have not been specifically defined as "missing" in a source or Type node. System-missing values are displayed as **$null$**. Note that empty strings are not considered nulls in IBM SPSS Modeler, although they may be treated as nulls by certain databases.
   - **Empty strings and white space.** Empty string values and white space (strings with no visible characters) are treated as distinct from null values. Empty strings are treated as equivalent to white space for most purposes. For example, if you select the option to treat white space as blanks in a source or Type node, this setting applies to empty strings as well.
   - **Blank or user-defined missing values.** These are values such as `unknown`, `99`, or `–1` that are explicitly defined in a source node or Type node as missing. Optionally, you can also choose to treat nulls and white space as blanks, which allows them to be flagged for special treatment and to be excluded from most calculations. For example, you can use the `@BLANK` function to treat these values, along with other types of missing values, as blanks.
2. In the **Method** column, specify the method you want to use.

   The following methods are available for imputing missing values:

   Fixed. Substitutes a fixed value (either the field mean, midpoint of the range, or a constant that you specify).

Random. Substitutes a random value based on a normal or uniform distribution.

Expression. Allows you to specify a custom expression. For example, you could replace values with a global variable created by the Set Globals node.

Algorithm. Substitutes a value predicted by a model based on the C&RT algorithm. For each field imputed using this method, there will be a separate C&RT model, along with a Filler node that replaces blanks and nulls with the value predicted by the model. A Filter node is then used to remove the prediction fields generated by the model.

3. To generate a Missing Values SuperNode, from the menus choose:

   Generate > **Missing Values SuperNode**

   The Missing Values SuperNode dialog box is displayed.
4. Select **All fields** or **Selected fields only**, and specify a sample size if desired. (The specified sample is a percentage; by default, 10% of all records are sampled.)
5. Click **OK** to add the generated SuperNode to the stream canvas.
6. Attach the SuperNode to the stream to apply the transformations.

Within the SuperNode, a combination of model nugget, Filler, and Filter nodes is used as appropriate. To understand how it works, you can edit the SuperNode and click **Zoom In**, and you can add, edit, or remove specific nodes within the SuperNode to fine-tune the behavior.

Handling Outliers and Extreme Values

The audit report lists number of outliers and extremes is listed for each field based on the detection options specified in the Data Audit node. See the topic "Data Audit Quality Tab" on page 284 for more information. You can choose to coerce, discard, or nullify these values for specific fields as appropriate, and then generate a SuperNode to apply the transformations.

1. In the **Action** column, specify handling for outliers and extremes for specific fields as desired.

   The following actions are available for handling outliers and extremes:

   - **Coerce.** Replaces outliers and extreme values with the nearest value that would not be considered extreme. For example if an outlier is defined to be anything above or below three standard deviations, then all outliers would be replaced with the highest or lowest value within this range.
   - **Discard.** Discards records with outlying or extreme values for the specified field.
   - **Nullify.** Replaces outliers and extremes with the null or system-missing value.
   - **Coerce outliers / discard extremes.** Discards extreme values only.
   - **Coerce outliers / nullify extremes.** Nullifies extreme values only.
2. To generate the SuperNode, from the menus choose:

   Generate > **Outlier & Extreme SuperNode**

   The Outlier SuperNode dialog box is displayed.
3. Select **All fields** or **Selected fields only**, and then click **OK** to add the generated SuperNode to the stream canvas.
4. Attach the SuperNode to the stream to apply the transformations.

Optionally, you can edit the SuperNode and zoom in to browse or make changes. Within the SuperNode, values are discarded, coerced, or nullified using a series of Select and/or Filler nodes as appropriate.

Filtering Fields with Missing Data

From the Data Audit browser, you can create a new Filter node based on the results of the Quality analysis by using the Generate Filter from Quality dialog box.

Mode. Select the desired operation for specified fields, either **Include** or **Exclude**.

- **Selected fields.** The Filter node will include/exclude the fields selected on the Quality tab. For example you could sort the table on the **% Complete** column, use Shift-click to select the least complete fields, and then generate a Filter node that excludes these fields.
- **Fields with quality percentage higher than.** The Filter node will include/exclude fields where the percentage of complete records is greater than the specified threshold. The default threshold is 50%.

Filtering Empty or Typeless Fields

Note that after data values have been instantiated, typeless or empty fields are excluded from the audit results and from most other output in IBM SPSS Modeler. These fields are ignored for purposes of modeling, but may bloat or clutter the data. If so, you can use the Data Audit browser to generate a Filter node from that removes these fields from the stream.

1. To make sure that all fields are included in the audit, including empty or typeless fields, click **Clear All Values** in the upstream source or Type node, or set Values to *<Pass>* for all fields.
2. In the Data Audit browser, sort on the **% Complete** column, select the fields that have zero valid values (or some other threshold) and use the Generate menu to produce a Filter node which can be added to the stream.

Selecting Records with Missing Data

From the Data Audit browser, you can create a new Select node based on the results of the quality analysis.

1. In the Data Audit browser, choose the Quality tab.
2. From the menu, choose:

   Generate > **Missing Values Select Node**

   The Generate Select Node dialog box is displayed.

Select when record is. Specify whether records should be kept when they are **Valid** or **Invalid**.

Look for invalid values in. Specify where to check for invalid values.

- **All fields.** The Select node will check all fields for invalid values.
- **Fields selected in table.** The Select node will check only the fields currently selected in the Quality output table.
- **Fields with quality percentage higher than.** The Select node will check fields where the percentage of complete records is greater than the specified threshold. The default threshold is 50%.

Consider a record invalid if an invalid value is found in. Specify the condition for identifying a record as invalid.

- **Any of the above fields.** The Select node will consider a record invalid if *any* of the fields specified above contains an invalid value for that record.
- **All of the above fields.** The Select node will consider a record invalid only if *all* of the fields specified above contain invalid values for that record.

Generating Other Nodes for Data Preparation

A variety of nodes used in data preparation can be generated directly from the Data Audit browser, including Reclassify, Binning, and Derive nodes. For example:

- You can derive a new field based on the values of *claimvalue* and *farmincome* by selecting both in the audit report and choosing **Derive** from the Generate menu. The new node is added to the stream canvas.
- Similarly, you may determine, based on audit results, that recoding *farmincome* into percentile-based bins provides a more focused analysis. To generate a Binning node, select the field row in the display and choose **Binning** from the Generate menu.

Once a node is generated and added to the stream canvas, you must attach it to the stream and open the node to specify options for the selected field(s).

Transform Node

Normalizing input fields is an important step before using traditional scoring techniques, such as regression, logistic regression, and discriminant analysis. These techniques carry assumptions about normal distributions of data that may not be true for many raw data files. One approach to dealing with real-world data is to apply transformations that move a raw data element toward a more normal distribution. In addition, normalized fields can easily be compared with each other—for example, income and age are on totally different scales in a raw data file but when normalized, the relative impact of each can be easily interpreted.

The Transform Node provides an output viewer that enables you to perform a rapid visual assessment of the best transformation to use. You can see at a glance whether variables are normally distributed and, if necessary, choose the transformation you want and apply it. You can pick multiple fields and perform one transformation per field.

After selecting the preferred transformations for the fields, you can generate Derive or Filler nodes that perform the transformations and attach these nodes to the stream. The Derive node creates new fields, while the Filler node transforms the existing ones. See the topic "Generating Graphs" on page 291 for more information.

Transform node Fields tab

On the Fields tab, you can specify which fields of the data you want to use for viewing possible transformations and applying them. Only numeric fields can be transformed. Click the field selector button and select one or more numeric fields from the list displayed.

Transform Node Options Tab

The Options tab enables you to specify the type of transformations you want to include. You can choose to include all available transformations or select transformations individually.

In the latter case, you can also enter a number to offset the data for the inverse and log transformations. Doing so is useful in situations where a large proportion of zeros in the data would bias the mean and standard deviation results.

For example, assume that you have a field named *BALANCE* that has some zero values in it, and you want to use the inverse transformation on it. To avoid undesired bias, you would select **Inverse (1/x)** and enter 1 in the **Use a data offset** field. (Note that this offset is not related to that performed by the @OFFSET sequence function in IBM SPSS Modeler.)

All formulas. Indicates that all available transformations should be calculated and shown in the output.

Select formulas. Enables you to select the different transformations to be calculated and shown in the output.

- **Inverse (1/x).** Indicates that the inverse transformation should be displayed in the output.
- **Log (log n).** Indicates that the $\log_n$ transformation should be displayed in the output.
- **Log (log 10).** Indicates that the $\log_{10}$ transformation should be displayed in the output.
- **Exponential.** Indicates that the exponential transformation (e^x) should be displayed in the output.
- **Square Root.** Indicates that the square root transformation should be displayed in the output.

Transform Node Output Tab

The Output tab lets you specify the output format and location of the output. You can also choose to display the results on the screen, or send them to one of the standard file types. See the topic "Output Node Output Tab" on page 275 for more information.

Transform Node Output Viewer

The output viewer enables you to see the results of executing the Transform Node. The viewer is a powerful tool that displays multiple transformations per field in thumbnail views of the transformation, enabling you to compare fields quickly. You can use options on its File menu to save, export, or print the output. See the topic "Viewing output" on page 272 for more information.

For each transformation (other than Selected Transform), a legend is displayed underneath in the format:

```
Mean (Standard deviation)
```

Generating Nodes for the Transformations

The output viewer provides a useful starting point for your data preparation. For example, you might want to normalize the field *AGE* so that you can use a scoring technique (such as logistic regression or discriminant analysis) that assumes a normal distribution. Based upon the initial graphs and summary statistics, you might decide to transform the *AGE* field according to a particular distribution (for example, log). After selecting the preferred distribution, you can then generate a derive node with a standardized transformation to use for scoring.

You can generate the following field operations nodes from the output viewer:

- Derive
- Filler

A Derive node creates new fields with the desired transformations, while the Filler node transforms existing fields. The nodes are placed on the canvas in the form of a SuperNode.

If you select the same transformation for different fields, a Derive or Filler node contains the formulas for that transformation type for all the fields to which that transformation applies. For example, assume that you have selected the fields and transformations, shown in the following table, to generate a Derive node.

Table 43. Example of Derive node generation

Field	**Transformation**
AGE	Current Distribution
INCOME	Log
OPEN_BAL	Inverse
BALANCE	Inverse

The following nodes are contained in the SuperNode:

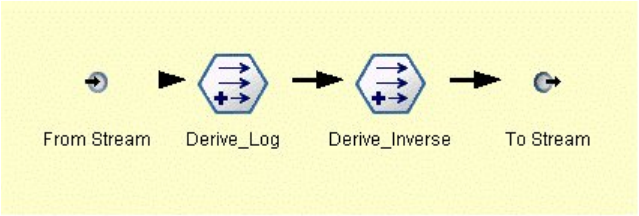


Figure 74. SuperNode on canvas

In this example, the Derive_Log node has the log formula for the *INCOME* field, and the Derive_Inverse node has the inverse formulas for the *OPEN_BAL* and *BALANCE* fields.

To Generate a Node

1. For each field in the output viewer, select the desired transformation.
2. From the Generate menu, choose **Derive Node** or **Filler Node** as desired.

   Doing so displays the Generate Derive Node or Generate Filler Node dialog box, as appropriate.

Choose **Non-standardized transformation** or **Standardized transformation (z-score)** as desired. The second option applies a *z* score to the transformation; *z* scores represent values as a function of distance from the mean of the variable in standard deviations. For example, if you apply the log transformation to the *AGE* field and choose a standardized transformation, the final equation for the generated node will be:

```
(log(AGE)-Mean)/SD
```

Once a node is generated and appears on the stream canvas:

1. Attach it to the stream.
2. For a SuperNode, optionally double-click the node to view its contents.
3. Optionally double-click a Derive or Filler node to modify options for the selected field(s).

Generating Graphs

You can generate full-size histogram output from a thumbnail histogram in the output viewer.

To Generate a Graph

1. Double-click a thumbnail graph in the output viewer.

   or

   Select a thumbnail graph in the output viewer.
2. From the Generate menu, choose **Graph output**.

Doing so displays the histogram with a normal distribution curve overlaid. This enables you to compare how closely each available transformation matches a normal distribution.

Note: You can only generate a graph if the Transform node that created the output is connected to the stream.

Other Operations

From the output viewer, you can also:

- Sort the output grid by the Field column.
- Export the output to an HTML file. See the topic "Exporting Output" on page 274 for more information.

Statistics Node

The Statistics node gives you basic summary information about numeric fields. You can get summary statistics for individual fields and correlations between fields.

Statistics Node Settings Tab

Examine. Select the field or fields for which you want individual summary statistics. You can select multiple fields.

Statistics. Select the statistics to report. Available options include **Count**, **Mean**, **Sum**, **Min**, **Max**, **Range**, **Variance**, **Std Dev**, **Std Error of Mean**, **Median**, and **Mode**.

Correlate. Select the field or fields that you want to correlate. You can select multiple fields. When correlation fields are selected, the correlation between each Examine field and the correlation field(s) will be listed in the output.

Correlation Settings. You can specify options for displaying the strength of correlations in the output.

Correlation Settings

IBM SPSS Modeler can characterize correlations with descriptive labels to help highlight important relationships. The **correlation** measures the strength of relationship between two continuous (numeric range) fields. It takes values between –1.0 and 1.0. Values close to +1.0 indicate a strong positive

association so that high values on one field are associated with high values on the other and low values are associated with low values. Values close to –1.0 indicate a strong negative association so that high values for one field are associated with low values for the other, and vice versa. Values close to 0.0 indicate a weak association, so that values for the two fields are more or less independent.

Using the Correlation Settings dialog box you can control display of correlation labels, change the thresholds that define the categories, and change the labels used for each range. Because the way you characterize correlation values depends greatly on the problem domain, you may want to customize the ranges and labels to fit your specific situation.

Show correlation strength labels in output. This option is selected by default. Deselect this option to omit the descriptive labels from the output.

Correlation Strength. There are two options for defining and labeling the strength of correlations:

- **Define correlation strength by importance (1-p).** Labels correlations based on importance, defined as 1 minus the significance, or 1 minus the probability that the difference in means could be explained by chance alone. The closer this value comes to 1, the greater the chance that the two fields are *not* independent—in other words, that some relationship exists between them. Labeling correlations based on importance is generally recommended over absolute value because it accounts for variability in the data—for example, a coefficient of 0.6 may be highly significant in one dataset and not significant at all in another. By default, importance values between 0.0 and 0.9 are labeled as *Weak*, those between 0.9 and 0.95 are labeled as *Medium*, and those between 0.95 and 1.0 are labeled as *Strong*.
- **Define correlation strength by absolute value.** Labels correlations based on the absolute value of the Pearson's correlation coefficient, which ranges between –1 and 1, as described above. The closer the absolute value of this measure comes to 1, the stronger the correlation. By default, correlations between 0.0 and 0.3333 (in absolute value) are labeled as *Weak*, those between 0.3333 and 0.6666 are labeled as *Medium*, and those between 0.6666 and 1.0 are labeled as *Strong*. Note, however, that the significance of any given value is difficult to generalize from one dataset to another; for this reason, defining correlations based on probability rather than absolute value is recommended in most cases.

Statistics Output Browser

The Statistics node output browser displays the results of the statistical analysis and enables you to perform operations, including selecting fields, and generating new nodes based on the selection, and saving and printing the results. The usual saving, exporting, and printing options are available from the File menu, and the usual editing options are available from the Edit menu. See the topic "Viewing output" on page 272 for more information.

When you first browse Statistics output, the results are expanded. To hide results after viewing them, use the expander control to the left of the item to collapse the specific results you want to hide or click the **Collapse All** button to collapse all results. To see results again after collapsing them, use the expander control to the left of the item to show the results or click the **Expand All** button to show all results.

The output contains a section for each *Examine* field, containing a table of the requested statistics.

- **Count.** The number of records with valid values for the field.
- **Mean.** The average (mean) value for the field across all records.
- **Sum.** The sum of values for the field across all records.
- **Min.** The minimum value for the field.
- **Max.** The maximum value for the field.
- **Range.** The difference between the minimum and maximum values.
- **Variance.** A measure of the variability in the values of a field. It is calculated by taking the difference between each value and the overall mean, squaring it, summing across all of the values, and dividing by the number of records.
- **Standard Deviation.** Another measure of variability in the values of a field, calculated as the square root of the variance.

- **Standard Error of Mean.** A measure of the uncertainty in the estimate of the field's mean if the mean is assumed to apply to new data.
- **Median.** The "middle" value for the field; that is, the value that divides the upper half of the data from the lower half of the data (based on values of the field).
- **Mode.** The most common single value in the data.

Correlations. If you specified correlate fields, the output also contains a section listing the Pearson correlation between the Examine field and each correlate field, and optional descriptive labels for the correlation values. See the topic "Correlation Settings" on page 291 for more information.

Generate menu. The Generate menu contains node generation operations.

- **Filter.** Generates a Filter node to filter out fields that are uncorrelated or weakly correlated with other fields.

Generating a Filter Node from Statistics

The Filter node generated from a Statistics output browser will filter fields based on their correlations with other fields. It works by sorting the correlations in order of absolute value, taking the largest correlations (according to the criterion set in the Generate Filter from Statistics dialog box), and creating a filter that passes all fields that appear in any of those large correlations.

Mode. Decide how to select correlations. **Include** causes fields appearing in the specified correlations to be retained. **Exclude** causes the fields to be filtered.

Include/Exclude fields appearing in. Define the criterion for selecting correlations.

- **Top number of correlations.** Selects the specified number of correlations and includes/excludes fields that appear in any of those correlations.
- **Top percentage of correlations (%).** Selects the specified percentage (*n*%) of correlations and includes/excludes fields that appear in any of those correlations.
- **Correlations greater than.** Selects correlations greater in absolute value than the specified threshold.

Means Node

The Means node compares the means between independent groups or between pairs of related fields to test whether a significant difference exists. For example, you can compare mean revenues before and after running a promotion or compare revenues from customers who didn't receive the promotion with those who did.

You can compare means in two different ways, depending on your data:

- **Between groups within a field.** To compare independent groups, select a test field and a grouping field. For example, you could exclude a sample of "holdout" customers when sending a promotion and compare mean revenues for the holdout group with all of the others. In this case, you would specify a single test field that indicates the revenue for each customer, with a flag or nominal field that indicates whether they received the offer. The samples are independent in the sense that each record is assigned to one group or another, and there is no way to link a specific member of one group to a specific member of another. You can also specify a nominal field with more than two values to compare the means for multiple groups. When executed, the node calculates a one-way ANOVA test on the selected fields. In cases where there are only two field groups, the one-way ANOVA results are essentially the same as an independent-samples *t* test. See the topic "Comparing Means for Independent Groups" on page 294 for more information.
- **Between pairs of fields.** When comparing means for two related fields, the groups must be paired in some way for the results to be meaningful. For example, you could compare the mean revenues from the same group of customers before and after running a promotion or compare usage rates for a service between husband-wife pairs to see if they are different. Each record contains two separate but related measures that can be compared meaningfully. When executed, the node calculates a paired-samples *t* test on each field pair selected. See the topic "Comparing Means Between Paired Fields" on page 294 for more information.

Comparing Means for Independent Groups

Select **Between groups within a field** in the Means node to compare the mean for two or more independent groups.

Grouping field. Select a numeric flag or nominal field with two or more distinct values that divides records into the groups you want to compare, such as those who received an offer versus those who did not. Regardless of the number of test fields, only one grouping field can be selected.

Test fields. Select one or more numeric fields that contain the measures you want to test. A separate test will be conducted for each field you select. For example, you could test the impact of a given promotion on usage, revenue, and churn.

Comparing Means Between Paired Fields

Select **Between pairs of fields** in the Means node to compare means between separate fields. The fields must be related in some way for the results to be meaningful, such as revenues before and after a promotion. Multiple field pairs can also be selected.

Field one. Select a numeric field that contains the first of the measures you want to compare. In a before-and-after study, this would be the "before" field.

Field two. Select the second field you want to compare.

Add. Adds the selected pair to the Test field pair(s) list.

Repeat field selections as needed to add multiple pairs to the list.

Correlation settings. Enables you to specify options for labeling the strength of correlations. See the topic "Correlation Settings" on page 291 for more information.

Means Node Options

The Options tab enables you to set the threshold *p* values used to label results as important, marginal, or unimportant. You can also edit the label for each ranking. Importance is measured on a percentage scale and can be broadly defined as 1 minus the probability of obtaining a result (such as the difference in means between two fields) as extreme as or more extreme than the observed result by chance alone. For example, a *p* value greater than 0.95 indicates less than a 5% chance that the result could be explained by chance alone.

Importance labels. You can edit the labels used to label each field pair or group in the output. The default labels are *important*, *marginal*, and *unimportant*.

Cutoff values. Specifies the threshold for each rank. Typically *p* values greater than 0.95 would rank as important, while those lower than 0.9 would be unimportant, but these thresholds can be adjusted as needed.

Note: Importance measures are available in a number of nodes. The specific computations depend on the node and on the type of target and input fields used, but the values can still be compared, since all are measured on a percentage scale.

Means Node Output Browser

The Means output browser displays cross-tabulated data and enables you to perform standard operations including selecting and copying the table one row at a time, and sorting by any column, and saving and printing the table. See the topic "Viewing output" on page 272 for more information.

The specific information in the table depends on the type of comparison (groups within a field or separate fields).

Sort by. Enables you to sort the output by a specific column. Click the up or down arrow to change the direction of the sort. Alternatively, you can click on any column heading to sort by that column. (To change the direction of the sort within the column, click again.)

View. You can choose **Simple** or **Advanced** to control the level of detail in the display. The advanced view includes all of the information from the simple view but with additional details provided.

Means Output Comparing Groups within a Field

When comparing groups within a field, the name of the grouping field is displayed above the output table, and means and related statistics are reported separately for each group. The table includes a separate row for each test field.

The following columns are displayed:

- **Field.** Lists the names of the selected test fields.
- **Means by group.** Displays the mean for each category of the grouping field. For example, you might compare those who received a special offer (*New Promotion*) with those who didn't (*Standard*). In the advanced view, the standard deviation, standard error, and count are also displayed.
- **Importance.** Displays the importance value and label. See the topic "Means Node Options" on page 294 for more information.

Advanced Output

In the advanced view, the following additional columns are displayed.

- **F-Test.** This test is based on the ratio of the variance between the groups and the variance within each group. If the means are the same for all groups, you would expect the *F* ratio to be close to 1 since both are estimates of the same population variance. The larger this ratio, the greater the variation between groups and the greater than chance that a significant difference exists.
- **df.** Displays the degrees of freedom.

Means Output Comparing Pairs of Fields

When comparing separate fields, the output table includes a row for each selected field pair.

- **Field One/Two.** Displays the name of the first and second field in each pair. In the advanced view, the standard deviation, standard error, and count are also displayed.
- **Mean One/Two.** Displays the mean for each field, respectively.
- **Correlation.** Measures the strength of relationship between two continuous (numeric range) fields. Values close to +1.0 indicate a strong positive association, and values close to –1.0 indicate a strong negative association. See the topic "Correlation Settings" on page 291 for more information.
- **Mean Difference.** Displays the difference between the two field means.
- **Importance.** Displays the importance value and label. See the topic "Means Node Options" on page 294 for more information.

Advanced Output

Advanced output adds the following columns:

95% Confidence Interval. Lower and upper boundaries of the range within which the true mean is likely to fall in 95% of all possible samples of this size from this population.

T-Test. The *t* statistic is obtained by dividing the mean difference by its standard error. The greater the absolute value of this statistic, the greater the probability that the means are not the same.

df. Displays the degrees of freedom for the statistic.

Report Node

The Report node enables you to create formatted reports containing fixed text, as well as data and other expressions derived from the data. You specify the format of the report by using text templates to define the fixed text and the data output constructions. You can provide custom text formatting using HTML tags in the template and by setting options on the Output tab. Data values and other conditional output are included in the report using CLEM expressions in the template.

Alternatives to the Report Node

The Report node is most typically used to list records or cases output from a stream, such as all records meeting a certain condition. In this regard, it can be thought of as a less-structured alternative to the Table node.

- If you want a report that lists field information or anything else that is defined in the stream rather than the data itself (such as field definitions specified in a Type node), then a script can be used instead.
- To generate a report that includes multiple output objects (such as a collection of models, tables, and graphs generated by one or more streams) and that can be output in multiple formats (including text, HTML, and Microsoft Word/Office), an IBM SPSS Modeler project can be used.
- To produce a list of field names without using scripting, you can use a Table node preceded by a Sample node that discards all records. This produces a table with no rows, which can be transposed on export to produce a list of field names in a single column. (Select **Transpose data** on the Output tab in the Table node to do this.)

Report Node Template Tab

Creating a template. To define the contents of the report, you create a template on the Report node Template tab. The template consists of lines of text, each of which specifies something about the contents of the report, and some special tag lines used to indicate the scope of the content lines. Within each content line, CLEM expressions enclosed in square brackets ([]) are evaluated before the line is sent to the report. There are three possible scopes for a line in the template:

Fixed. Lines that are not marked otherwise are considered fixed. Fixed lines are copied into the report only once, after any expressions that they contain are evaluated. For example, the line

```
This is my report, printed on [@TODAY]
```

would copy a single line to the report, containing the text and the current date.

Global (iterate ALL). Lines contained between the special tags #ALL and # are copied to the report once for each record of input data. CLEM expressions (enclosed in brackets) are evaluated based on the current record for each output line. For example, the lines

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

would include one line for each record indicating the record number and age.

To generate a list of all records:

```
#ALL
[Age]    [Sex]    [Cholesterol]     [BP]
#
```

Conditional (iterate WHERE). Lines contained between the special tags #WHERE <*condition*> and # are copied to the report once for each record where the specified condition is true. The condition is a CLEM expression. (In the WHERE condition, the brackets are optional.) For example, the lines

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

will write one line to the file for each record with a value of *M* for sex. The complete report will contain the fixed, global, and conditional lines defined by applying the template to the input data.

You can specify options for displaying or saving results using the Output tab, common to various types of output nodes. See the topic “Output Node Output Tab” on page 275 for more information.

Outputting Data in HTML or XML Format

You can include HTML or XML tags directly in the template in order to write reports in either of these formats. For example, the following template produces an HTML table.

```
This report is written in HTML.
Only records where Age is above 60 are included.

<HTML>
  <TABLE border="2">
      <TR>
        <TD>Age</TD>
        <TD>BP</TD>
        <TD>Cholesterol</TD>
        <TD>Drug</TD>
      </TR>

  #WHERE Age > 60
    <TR>
          <TD>[Age]</TD>
          <TD>[BP]</TD>
          <TD>[Cholesterol]</TD>
          <TD>[Drug]</TD>
        </TR>
  #
  </TABLE>
</HTML>
```

Report Node Output Browser

The report browser shows you the contents of the generated report. The usual saving, exporting, and printing options are available from the File menu, and the usual editing options are available from the Edit menu. See the topic "Viewing output" on page 272 for more information.

Set Globals Node

The Set Globals node scans the data and computes summary values that can be used in CLEM expressions. For example, you can use a Set Globals node to compute statistics for a field called *age* and then use the overall mean of *age* in CLEM expressions by inserting the function `@GLOBAL_MEAN(age)`.

Set Globals Node Settings Tab

Globals to be created. Select the field or fields for which you want globals to be available. You can select multiple fields. For each field, specify the statistics to compute by making sure that the statistics you want are selected in the columns next to the field name.

- **MEAN.** The average (mean) value for the field across all records.
- **SUM.** The sum of values for the field across all records.
- **MIN.** The minimum value for the field.
- **MAX.** The maximum value for the field.
- **SDEV.** The standard deviation, which is a measure of variability in the values of a field and is calculated as the square root of the variance.

Default operation(s). The options selected here will be used when new fields are added to the Globals list above. To change the default set of statistics, select or deselect statistics as appropriate. You can also use the **Apply** button to apply the default operations to all fields in the list.

Note: Some operations are not applicable to non-numeric fields (for example, Sum for a date/time field). Operations that cannot be used with a selected field are disabled

Clear all globals before executing. Select this option to remove all global values before calculating new values. If this option is not selected, newly calculated values replace older values, but globals that are not recalculated remain available, as well.

Display preview of globals created after execution. If you select this option, the Globals tab of the stream properties dialog box will appear after execution to display the calculated global values.

Simulation Fitting Node

The Simulation Fitting node fits a set of candidate statistical distributions to each field in the data. The fit of each distribution to a field is assessed using a goodness of fit criterion. When a Simulation Fitting node is executed, a Simulation Generate node is built (or an existing node is updated). Each field is assigned its best fitting distribution. The Simulation Generate node can then be used to generate simulated data for each field.

Although the Simulation Fitting node is a terminal node, it does not add a model to the generated models palette, add an output or chart to the outputs tab, or export data.

Note: If the historical data is sparse (that is, there are lots of missing values), it may difficult for the fitting component to find enough valid values to fit distributions to the data. In cases where the data is sparse, before fitting you should either remove the sparse fields if they are not required, or impute the missing values. Using the options on the **Quality** tab of the Data Audit node you can view the number of complete records, identify which fields are sparse, and select an imputation method. If there are an insufficient number of records for distribution fitting, you can use a Balance node to increase the number of records.

Using a Simulation Fitting node to automatically create a Simulation Generate node

The first time the Simulation Fitting node is executed, a Simulation Generate node is created with an update link to the Simulation Fitting node. If the Simulation Fitting node is executed again, a new Simulation Generate node will be created only if the update link has been removed. A Simulation Fitting node can also be used to update a connected Simulation Generate node. The result depends on whether the same fields are present in both nodes, and if the fields are unlocked in the Simulation Generate node. See the topic “Simulation Generate Node” on page 46 for more information.

A Simulation Fitting node can only have an update link to a Simulation Generate node. To define an update link to a Simulation Generate node, follow these steps:

1. Right-click the Simulation Fitting node.
2. From the menu, select **Define Update Link**.
3. Click the Simulation Generate node to which you want to define an update link.

To remove an update link between a Simulation Fitting node and a Simulation Generate node, right-click on the update link, and select **Remove Link**.

Distribution Fitting

A statistical distribution is the theoretical frequency of the occurrence of values that a variable can take. In the Simulation Fitting node, a set of theoretical statistical distributions is compared to each field of data. The distributions that are available for fitting are described in the topic “Distributions” on page 55. The parameters of the theoretical distribution are adjusted to give the best fit to the data according to a measurement of the goodness of fit; either the Anderson-Darling criterion or the Kolmogorov-Smirnov criterion. The results of the distribution fitting by the Simulation Fitting node show which distributions were fitted, the best estimates of the parameters for each distribution, and how well each distribution fits the data. During distribution fitting, correlations between fields with numeric storage types, and contingencies between fields with a categorical distribution, are also calculated. The results of the distribution fitting are used to create a Simulation Generate node.

Before any distributions are fitted to your data, the first 1000 records are examined for missing values. If there are too many missing values, distribution fitting is not possible. If so, you must decide whether either of the following options are appropriate:

- Use an upstream node to remove records with missing values.
- Use an upstream node to impute values for missing values.

Distribution fitting does not exclude user-missing values. If your data have user-missing values and you want those values to be excluded from distribution fitting then you should set those values to system missing.

The role of a field is not taken into account when the distributions are fitted. For example, fields with the role **Target** are treated the same as fields with roles of **Input**, **None**, **Both**, **Partition**, **Split**, **Frequency**, and **ID**.

Fields are treated differently during distribution fitting according to their storage type and measurement level. The treatment of fields during distribution fitting is described in the following table.

Table 44. Distribution fitting according to storage type and measurement level of fields

<table>
<tr><th>Storage type</th><th colspan="5">Measurement Level</th></tr>
<tr><td></td><td>Continuous</td><td>Categorical</td><td>Flag</td><td>Nominal</td><td>Ordinal</td><td>Typeless</td></tr>
<tr><td>String</td><td>Impossible</td><td colspan="3">Categorical, dice and fixed distributions are fitted</td><td rowspan="9">Field is ignored and not passed to the Simulation Generate node.</td></tr>
<tr><td>Integer</td><td rowspan="6">All distributions are fitted. Correlations and contingencies are calculated.</td><td colspan="3" rowspan="6">The categorical distribution is fitted. Correlations are not calculated.</td><td rowspan="6">Binomial, negative binomial and Poisson distributions are fitted, and correlations are calculated.</td></tr>
<tr><td>Real</td></tr>
<tr><td>Time</td></tr>
<tr><td>Date</td></tr>
<tr><td>Timestamp</td></tr>
<tr><td>Unknown</td><td colspan="5">Appropriate storage type is determined from the data.</td></tr>
</table>

Fields with the measurement level ordinal are treated like continuous fields and are included in the correlations table in the Simulation Generate node. If you want a distribution other than binomial, negative binomial or Poisson to be fitted to an ordinal field, you must change the measurement level of the field to continuous. If you have previously defined a label for each value of an ordinal field, and then change the measurement level to continuous, the labels will be lost.

Fields that have single values are not treated differently during distribution fitting to fields with multiple values. Fields with the storage type time, date, or timestamp are treated as numeric.

Fitting Distributions to Split Fields

If your data contains a split field, and you want distribution fitting to be carried out separately for each split, you must transform the data by using an upstream Restructure node. Using the Restructure node, generate a new field for each value of the split field. This restructured data can then be used for distribution fitting in the Simulation Fitting node.

Simulation Fitting Node Settings Tab

Source node name. You can generate the name of the generated (or updated) Simulation Generate node automatically by selecting **Auto**. The automatically generated name is the name specified in the Simulation Fitting node if a custom name has been specified (or Sim Gen if no custom name has been specified in the Simulation Fitting node). Select **Custom** to specify a custom name in the adjoining text field. Unless the text field is edited, the default custom name is Sim Gen.

Fitting Options With these options you can specify how the distributions are fitted to the fields and how the fit of the distributions is assessed.

- **Number of cases to sample.** This specifies the number of cases to use when fitting distributions to the fields in the data set. Select **All cases** to fit distributions to all of the records in the data. If your data set is very large, you might want to consider limiting the number of cases that are used for distribution fitting. Select **Limit to first N cases** to use only the first N cases. Click the arrows to specify the number of cases to use. Alternatively, you can use an upstream node to take a random sample of records for distribution fitting.
- **Goodness of fit criteria (continuous fields only).** For continuous fields, select either the Anderson-Darling test or the Kolmogorov-Smirnoff test of goodness of fit to rank distributions when fitting distributions to the fields. The Anderson-Darling test is selected by default and is especially recommended when you want to ensure the best possible fit in the tail regions. Both statistics are calculated for every candidate distribution, but only the selected statistic is used to order the distributions and determine the best fitting distribution.
- **Bins (empirical distribution only).** For continuous fields, the Empirical distribution is the cumulative distribution function of the historical data. It is the probability of each value, or range of values, and is derived directly from the data. You can specify the number of bins that are used for calculating the Empirical distribution for continuous fields by clicking the arrows. The default is 100 and the maximum is 1000.
- **Weight field (optional).** If your data set contains a weight field, click the field picker icon and select the weight field from the list. The weight field is then excluded from the distribution fitting process. The list shows all fields in the data set that have the measurement level continuous. You can select only one weight field.

Simulation Evaluation Node

The Simulation Evaluation node is a terminal node that evaluates a specified field, provides a distribution of the field, and produces charts of distributions and correlations. This node is primarily used to evaluate continuous fields. It therefore compliments the evaluation chart, which is generated by an Evaluation node and is useful for evaluating discrete fields. Another difference is that the Simulation Evaluation node evaluates a single prediction across several iterations, whereas the Evaluation node evaluates multiple predictions each with a single iteration. Iterations are generated when more than one value is specified for a distribution parameter in the Simulation Generate node. See the topic “Iterations” on page 55 for more information.

The Simulation Evaluation node is designed to be used with data that was obtained from the Simulation Fitting and Simulation Generate nodes. The node can, however, be used with any other node. Any number of processing steps can be placed between the Simulation Generate node and the Simulation Evaluation node.

Important: The Simulation Evaluation node requires a minimum of 1000 records with valid values for the target field.

Simulation Evaluation Node Settings Tab

On the Settings tab of the Simulation Evaluation node, you can specify the role of each field in your data set, and customize the output that is generated by the simulation.

Select an item. Enables you to switch between the three views of the Simulation Evaluation node: Fields, Density Functions, and Outputs.

Fields view

Target field. This is a required field. Click the arrow to select the target field of your data set from the drop-down list. The selected field can have either continuous, ordinal or nominal measurement level, but cannot have date or an unspecified measurement level.

Iteration field (optional). If your data has an iteration field that indicates to which iteration each record in your data belongs, you must select it here. This means that each iteration will be evaluated separately. Only fields with continuous, ordinal or nominal measurement levels can be selected.

Input data is already sorted by iteration. Only enabled if an iteration field is specified in the **Iteration field (optional)** field. Only select this option if you are sure that your input data is already sorted by the iteration field specified in **Iteration field (optional)**.

Maximum number of iterations to plot. Only enabled if an iteration field is specified in the **Iteration field (optional)** field. Click the arrows to specify the number of iterations to plot. Specifying this number avoids trying to plot too many iterations on a single chart, which would make the plot difficult to interpret. The lowest level that the maximum number of iterations can be set to is 2; the highest level is 50. The maximum number of iterations to plot is initially set to 10.

Input fields for correlation tornado. The correlation tornado chart is a bar chart that displays the correlation coefficients between the specified target and each of the specified inputs. Click the field picker icon to select the input fields to be included in the tornado chart, from a list of available simulated inputs. Only input fields with continuous and ordinal measurement levels can be selected. Nominal, typeless, and date input fields are not available in the list and cannot be selected.

Density Functions view

With the options in this view, you can customize the output for probability density functions and cumulative distribution functions for continuous targets, as well as bar charts of predicted values for categorical targets.

Density functions. Density functions are the primary means of probing the set of outcomes from your simulation.

- **Probability density function (PDF).** Select this option to produce a probability density function for the target field. The probability density function displays the distribution of the target values. You can use the probability density function to determine the probability that the target is within a particular region. For categorical targets (targets with a measurement level of nominal or ordinal) a bar chart is generated that displays the percentage of cases that fall in each category of the target.
- **Cumulative distribution function (CDF).** Select this option to produce a cumulative distribution function for the target field. The cumulative distribution function displays the probability that the value of the target is less than or equal to a specified value. It is only available for continuous targets.

Reference lines (continuous). These options are enabled if **Probability density function (PDF)**, **Cumulative distribution function (CDF)**, or both, are selected. With these options, you can add various fixed vertical reference lines to probability density functions and cumulative distribution functions.

- **Mean.** Select this option to add a reference line at the mean value of the target field.
- **Median.** Select this option to add a reference line at the median value of the target field.
- **Standard Deviations.** Select this option to add reference lines at plus and minus a specified number of standard deviations from the mean of the target field. Selecting this option enables the adjoining **Number** field. Click the arrows to specify the number of standard deviations. The minimum number of standard deviations is 1 and the maximum is 10. The number of standard deviations is initially set to 3.

- **Percentiles.** Select this option to add reference lines at two percentile values of the distribution of the target field. Selecting this option enables the adjoining **Bottom** and **Top** text fields. For example, entering a value of 90 in the **Top** text field would add a reference line at the 90th percentile of the target, which is the value below which 90% of the observations fall. Likewise, a value of 10 in the **Bottom** text field represents the tenth percentile of the target, which is the value below which 10% of the observations fall.
- **Custom reference lines.** Select this option to add reference lines at specified values along the horizontal axis. Selecting this option enables the adjoining **Values** table. Every time that you enter a valid number into the **Values** table, a new empty row is appended to the bottom of the table. A *valid* number is a number within the range of values of the target field

Note: When multiple density functions or distribution functions (from multiple iterations) are displayed on a single chart, reference lines (other than custom lines) are separately applied to each function.

Categorical target (PDF only). These options are only enabled if **Probability density function (PDF)** is selected.

- **Category values to report.** For models with categorical target fields, the result of the model is a set of predicted probabilities, one for each category, that the target value falls in each category. The category with the highest probability is taken to be the predicted category and used in generating the bar chart for the probability density function. Select **Predicted category** to generate the bar chart. Select **Predicted probabilities** to generate histograms of the distribution of predicted probabilities for each of the categories of the target field. You can also select **Both** to generate both types of chart.
- **Grouping for sensitivity analysis.** Simulations that include sensitivity analysis iterations generate an independent target field (or predicted target field from a model) for each iteration that is defined by the analysis. There is one iteration for each value of the distribution parameter that is being varied. When iterations are present, the bar chart of the predicted category for a categorical target field is displayed as a clustered bar chart that includes the results for all iterations. Select either **Group categories together** or **Group iterations together**.

Outputs view

Percentile values of target distributions. With these options, you can choose to create a table of percentile values of the target distributions, and specify the percentiles to display.

Create a table of percentile values. For continuous target fields, select this option to obtain a table of specified percentiles of the target distributions. Choose one of the following options to specify the percentiles:

- **Quartiles.** Quartiles are the 25th, 50th, and 75th percentiles of the target field distribution. The observations are divided into four groups of equal size.
- **Intervals.** If you want an equal number of groups other than four, select **Intervals**. Selecting this option enables the adjoining **Number** field. Click the arrows to specify the number of intervals. The minimum number of intervals is 2 and the maximum is 100. The number of intervals is initially set to 10.
- **Custom percentiles.** Select **Custom percentiles** to specify individual percentiles, for example, the 99th percentile. Selecting this option enables the adjoining **Values** table. Every time that you enter a valid number, between 1 and 100, into the **Values** table, a new empty row is appended to the bottom of the table.

Simulation Evaluation node output

When the Simulation Evaluation node is executed, the output is added to the Output manager. The Simulation Evaluation output browser displays the results of executing the Simulation Evaluation node. The usual saving, exporting, and printing options are available from the **File** menu, and the usual editing options are available from the **Edit** menu. See the topic "Viewing output" on page 272 for more information. The **View** menu is only enabled if one of the charts is selected. It is not enabled for the distribution table or information outputs. From the **View** menu you can select **Edit Mode** to change the layout and look of the chart, or **Explore Mode** to explore the data and values represented by the chart. The Static mode fixes the chart reference lines (and sliders) in their current positions so they cannot be

moved. The Static mode is the only mode where you can copy, print, or export the chart with its reference lines. To select this mode, click **Static Mode** on the **View** menu.

The Simulation Evaluation output browser window consists of two panels. On the left of the window, there is a navigation panel that displays thumbnail representations of the charts that were generated when the Simulation Evaluation node was executed. When a thumbnail is selected, the chart output is displayed on the panel on the right of the window.

Navigation Panel

The navigation panel of the output browser contains thumbnails of the charts that are generated from a simulation. The thumbnails that are shown on the navigation panel depend on the measurement level of the target field, and on the options that are selected in the Simulation Evaluation node dialog box. Descriptions of the thumbnails are given in the following table.

Table 45. Navigation panel thumbnails

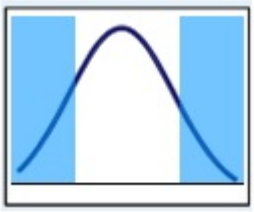
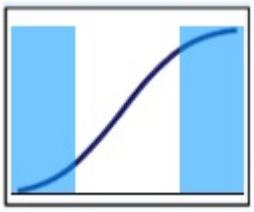
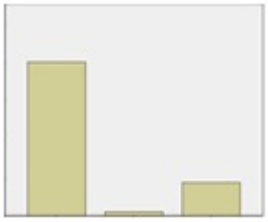
Thumbnail	Description	Comments
	Probability density function	This thumbnail is shown only if the measurement level of the target field is continuous, and **Probability Density Function (PDF)** is selected on the Density Functions view of the Simulation Evaluation node dialog box. If the measurement level of the target field is categorical, this thumbnail is not shown.
	Cumulative distribution function	This thumbnail is shown only if the measurement level of the target field is continuous, and **Cumulative Distribution Function (CDF)** is selected on the Density Functions view of the Simulation Evaluation node dialog box. If the measurement level of the target field is categorical, this thumbnail is not shown.
	Predicted category values	This thumbnail is shown only if the measurement level of the target field is categorical, **Probability Density Function (PDF)** is selected on the Density Functions view of the Simulation Evaluation node dialog box, and either **Predicted category** or **Both** is selected in the **Category values to report** area. If the measurement level of the target field is continuous, this thumbnail is not shown.

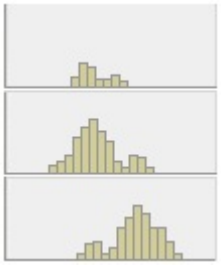
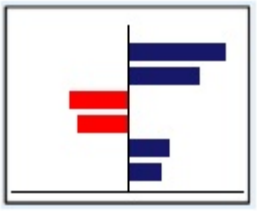
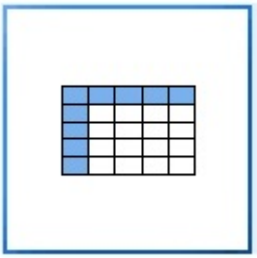
Table 45. Navigation panel thumbnails (continued)		
Thumbnail	**Description**	**Comments**
	Predicted category probabilities	This thumbnail is shown only if the measurement level of the target field is categorical, **Probability Density Function (PDF)** is selected on the Density Functions view of the Simulation Evaluation node dialog box, and either **Predicted probabilities** or **Both** is selected in the **Category values to report** area. If the measurement level of the target field is continuous, this thumbnail is not shown.
	Tornado charts	This thumbnail is shown only if one or more input fields are selected in the **Input fields for correlation tornado** field on the Fields view of the Simulation Evaluation node dialog box.
	Distribution table	This thumbnail is shown only if the measurement level of the target field is continuous, and **Create a table of percentile values** is selected on the Outputs view of the Simulation Evaluation node dialog box. The **View** menu is disabled for this chart. If the measurement level of the target field is categorical, this thumbnail is not shown.
	Information	This thumbnail is always shown. The **View** menu is disabled for this output.

Chart output

The types of output charts that are available depend on the measurement level of the target field, whether an iteration field is used, and the options that are selected in the Simulation Evaluation node dialog box. A number of the charts that are generated from a simulation have interactive features that you can use to customize the display. Interactive features are available by clicking **Chart Options**. All simulation charts are graphboard visualizations.

Probability density function charts for continuous targets. This chart shows both probability and frequency, with the probability scale on the left vertical axis and the frequency scale on the right vertical

axis. The chart has two sliding vertical reference lines that divide the chart into separate regions. The table below the chart displays the percent of the distribution that is in each of the regions. If multiple density functions are displayed on the same chart (because of iterations), the table has a separate row for the probabilities that are associated with each density function, and an extra column that contains the iteration name and a color that is associated with each density function. The iterations are listed in the table in alphabetical order, according to the iteration label. If no iteration label is available, the iteration value is used instead. The table cannot be edited.

Each of the reference lines has a slider (inverted triangle) that you can use to easily move the line. Each slider has a label that indicates its current position. By default, the sliders are positioned at the 5th and 95th percentiles of the distribution. If there are multiple iterations, the sliders are positioned at the 5th and 95th percentiles of the first iteration that is listed in the table. You cannot move the lines to cross through each other.

A number of additional features are available by clicking **Chart Options**. In particular, you can explicitly set the positions of the sliders, add fixed reference lines, and change the chart view from a continuous curve to a histogram. See the topic "Chart Options" on page 306 for more information. Right-click the chart to copy or export the chart.

Cumulative distribution function charts for continuous targets. This chart has the same two movable vertical reference lines and associated table that are described for the probability density function chart. The slider controls and table behave the same as the probability density function when there are multiple iterations. The same colors that are used to identify which density function belongs to each iteration are used for the distribution functions.

This chart also provides access to the **Chart Options** dialog box, which enables you to explicitly set the positions of the sliders, add fixed reference lines, and specify whether the cumulative distribution function is displayed as an increasing function (the default) or a decreasing function. See the topic "Chart Options" on page 306 for more information. Right-click the chart to copy, export, or edit the chart. Selecting **Edit** opens the chart in a floating graphboard editor window.

Predicted category values chart for categorical targets. For categorical target fields, a bar chart displays the predicted values. The predicted values are displayed as the percent of the target field that is predicted to fall in each category. For categorical target fields with sensitivity analysis iterations, results for the predicted target category are displayed as a clustered bar chart that includes the results for all iterations. The chart is clustered by category or by iteration, depending on which option was selected in the **Grouping for sensitivity analysis** area in the Density Functions view of the Simulation Evaluation node dialog box. Right-click the chart to copy, export, or edit the chart. Selecting **Edit** opens the chart in a floating graphboard editor window.

Predicted category probabilities chart for categorical targets. For categorical target fields, a histogram displays the distribution of predicted probabilities for each of the categories of the target. For categorical target fields with sensitivity analysis iterations, the histograms are displayed by category or by iteration, depending on which option was selected in the **Grouping for sensitivity analysis** area in the Density Functions view of the Simulation Evaluation node dialog box. If the histograms are grouped by category, a drop-down list containing the iteration labels enables you to choose which iteration to display. You can also select the iteration to display by right-clicking the chart and selecting the iteration from the **Iteration** submenu. If the histograms are grouped by iteration, a drop-down list containing the category names enables you to choose which category to display. You can also select which category to display by right-clicking the chart and selecting the category from the **Category** submenu.

This chart is only available for a subset of models, and on the model nugget the option to generate all group probabilities must be selected. For example, on the Logistic model nugget, you must select **Append all probabilities**. The following model nuggets support this option:

- Logistic, SVM, Bayes, Neural Net and KNN
- Db2/ISW in-database mining models for logistic regression, decision trees and naïve Bayes

By default, the option to generate all group probabilities is not selected on these model nuggets.

Tornado charts. The tornado chart is a bar chart that shows the sensitivity of the target field to each of the specified inputs. The sensitivity is measured by the correlation of the target with each input. The title

of the chart contains the name of the target field. Each bar on the chart represents the correlation between the target field and an input field. The simulated inputs that are included on the chart are the inputs that are selected in the **Input fields for correlation tornado** field on the Fields view of the Simulation Evaluation node dialog box. Each bar is labeled with the correlation value. Bars are ordered by the absolute value of the correlations, from largest value to smallest. If there are iterations, a separate chart is generated for each iteration. Each chart has a subtitle, which contains the name of the iteration.

Distribution table. This table contains the value of the target field, below which the specified percent of the observations fall. The table contains a row for each percentile value that is specified on the Outputs view of the Simulation Evaluation node dialog box. The percentile values can be quartiles, a different number of equally spaced percentiles, or individually specified percentiles. The distribution table contains a column for each iteration.

Information. This section gives an overall summary of the fields and records that are used in the evaluation. It also shows the input fields and record counts, which are broken down for each iteration.

Chart Options

In the Chart Options dialog box you can customize the display of activated charts of probability density functions and cumulative distribution functions that are generated from a simulation.

View. The **View** drop-down list only applies to the probability density function chart. You can use it to toggle the chart view from a continuous curve to a histogram. This feature is disabled when multiple density functions (from multiple iterations) are displayed on the same chart. When there are multiple density functions, the density functions can only be viewed as continuous curves.

Order. The **Order** drop-down list only applies to the cumulative distribution function chart. It specifies whether the cumulative distribution function is displayed as an ascending function (the default) or a descending function. When displayed as a descending function, the value of the function at a given point on the horizontal axis is the probability that the target field lies to the right of that point.

Slider positions. The **Upper** text field contains the current position of the right sliding reference line. The **Lower** text field contains the current position of the left sliding reference line. You can explicitly set the positions of the sliders by entering values in the **Upper** and **Lower** text fields. The value in the **Lower** text field must be strictly less than the value in the **Upper** text field. You can remove the left reference line by selecting **-Infinity**, effectively setting the position to negative infinity. This action disables the **Lower** text field. You can remove the right reference line by selecting **Infinity**, effectively setting its position to infinity. This action disables the **Upper** text field. You cannot remove both reference lines; selecting **-Infinity** disables the **Infinity** check box, and vice versa.

Reference lines. You can add various fixed vertical reference lines to probability density functions and cumulative distribution functions.

- **Mean.** You can add a reference line at the mean of the target field.
- **Median.** You can add a reference line at the median of the target field.
- **Standard Deviations.** You can add reference lines at plus and minus a specified number of standard deviations from the mean of the target field. You can enter the number of standard deviations to be used in the adjoining text field. The minimum number of standard deviations is 1, and the maximum is 10. The number of standard deviations is initially set to 3.
- **Percentiles.** You can add reference lines at one or two percentile values of the distribution for the target field by entering values into the **Bottom** and **Top** text fields. For example, a value of 95 in the **Top** text field represents the 95th percentile, which is the value below which 95% of the observations fall. Likewise, a value of 5 in the **Bottom** text field represents the fifth percentile, which is the value below which 5% of the observations fall. For the **Bottom** text field, the minimum percentile value is 0, and the maximum is 49. For the **Top** text field, the minimum percentile value is 50, and the maximum is 100.
- **Custom positions.** You can add reference lines at specified values along the horizontal axis. You can remove custom reference lines by deleting the entry from the grid.

When you click **OK**, the sliders, labels above the sliders, reference lines, and the table below the chart are updated to reflect the options that are selected in the Chart Options dialog box. Click **Cancel** to close the

dialog box without making any changes. Reference lines can be removed by deselecting the associated choice in the Chart Options dialog and clicking **OK**.

Note: When multiple density functions or distribution functions are displayed on a single chart (because of results from sensitivity analysis iterations), reference lines (other than custom lines) are separately applied to each function. Only the reference lines for the first iteration are shown. The reference line labels include the iteration label. The iteration label is derived from upstream, typically from a Simulation Generate node. If no iteration label is available, the iteration value is used instead. The **Mean**, **Median**, **Standard Deviations**, and **Percentiles** options are disabled for cumulative distribution functions with multiple iterations.

Extension Output node

If **Output to screen** is selected on the **Output** tab of the Extension Output node dialog, on-screen output is displayed in an output browser window. The output is also added to the Output manager. The output browser window has its own set of menus that allow you to print or save the output, or export it to another format. The **Edit** menu only contains the **Copy** option. The Extension Output node's output browser has two tabs; the **Text Output** tab that displays text output, and the **Graph Output** tab that displays graphs and charts.

If **Output to file** is selected on the **Output** tab of the Extension Output node dialog, the output browser window is not displayed upon successful execution of the Extension Output node.

Extension Output node - Syntax tab

Select your type of syntax – **R** or **Python for Spark**. See the following sections for more information. When your syntax is ready, you can click **Run** to execute the Extension Output node. The output objects are added to the Output manager, or optionally to the file specified in the **Filename** field on the **Output** tab.

R Syntax

R Syntax. You can enter, or paste, custom R scripting syntax for data analysis into this field.

Convert flag fields. Specifies how flag fields are treated. There are two options: **Strings to factor, Integers and Reals to double**, and **Logical values (True, False)**. If you select **Logical values (True, False)** the original values of the flag fields are lost. For example, if a field has values Male and Female, these are changed to True and False.

Convert missing values to the R 'not available' value (NA). When selected, any missing values are converted to the R NA value. The value NA is used by R to identify missing values. Some R functions that you use might have an argument that can be used to control how the function behaves when the data contain NA. For example, the function might allow you to choose to automatically exclude records that contain NA. If this option is not selected, any missing values are passed to R unchanged, and might cause errors when your R script is executed.

Convert date/time fields to R classes with special control for time zones. When selected, variables with date or datetime formats are converted to R date/time objects. You must select one of the following options:

- **R POSIXct.** Variables with date or datetime formats are converted to R `POSIXct` objects.
- **R POSIXlt (list).** Variables with date or datetime formats are converted to R `POSIXlt` objects.

Note: The POSIX formats are advanced options. Use these options only if your R script specifies that datetime fields are treated in ways that require these formats. The POSIX formats do not apply to variables with time formats.

Python Syntax

Python Syntax. You can enter, or paste, custom Python scripting syntax for data analysis into this field. For more information about Python for Spark, see Python for Spark and Scripting with Python for Spark.

Extension Output node - Console Output tab

The **Console Output** tab contains any output that is received when the R script or Python for Spark script on the Syntax tab runs (for example, if using an R script, it shows output received from the R console when the R script in the **R Syntax** field on the **Syntax** tab is executed). This output might include R or Python error messages or warnings that are produced when the R or Python script is executed. The output can be used, primarily, to debug the script. The **Console Output** tab also contains the script from the **R Syntax** or **Python Syntax** field.

Every time the Extension Output script is executed, the content of the **Console Output** tab is overwritten with the output received from the R console or Python for Spark. The output cannot be edited.

Extension Output node - Output tab

Output name. Specifies the name of the output that is produced when the node is executed. When **Auto** is selected, the name of the output is automatically set to "R Output" or "Python Output" depending on the script type. Optionally, you can select **Custom** to specify a different name.

Output to screen. Select this option to generate and display the output in a new window. The output is also added to the Output manager.

Output to file. Select this option to save the output to a file. Doing so enables the **Output Graph** and **Output File** radio buttons.

Output Graph. Only enabled if **Output to file** is selected. Select this option to save any graphs that result from executing the Extension Output node to a file. Specify a filename to use for the generated output in the **Filename** field. Click the ellipses button (**...**) to choose a specific file and location. Specify the file type in the **File type** drop-down list. The following file types are available:

- Output object (`.cou`)
- HTML (`.html`)

Output Text. Only enabled if **Output to file** is selected. Select this option to save any text output that results from executing the Extension Output node to a file. Specify a filename to use for the generated output in the **Filename** field. Click the ellipses button (**...**) to specify a specific file and location. Specify the file type in the **File type** drop-down list. The following file types are available:

- HTML (`.html`)
- Output object (`.cou`)
- Text document (`.txt`)

Extension Output Browser

If **Output to screen** is selected on the **Output** tab of the Extension Output node dialog box, on-screen output is displayed in an output browser window. The output is also added to the Output manager. The output browser window has its own set of menus that allow you to print or save the output, or export it to another format. The **Edit** menu only contains the **Copy** option. The output browser of the Extension Output node has two tabs:

- The **Text Output** tab displays text output
- The **Graph Output** tab displays graphs and charts

If **Output to file** is selected on the **Output** tab of the Extension Output node dialog box instead of **Output to screen**, the output browser window is not displayed upon successful execution of the Extension Output node.

Extension Output Browser - Text Output tab

The **Text Output** tab displays any text output that is generated when the R script or Python for Spark script on the **Syntax** tab of the Extension Output node is executed.

Note: R or Python for Spark error messages or warnings that result from executing your Extension Output script are always displayed on the **Console Output** tab of the Extension Output node.

Extension Output Browser - Graph Output tab

The **Graph Output** tab displays any graphs or charts that are generated when the R script or Python for Spark script on the **Syntax** tab of the Extension Output node is executed. For example, if your R script contains a call to the R `plot` function, the resulting graph is displayed on this tab.

KDE nodes

Kernel Density Estimation (KDE)© uses the Ball Tree or KD Tree algorithms for efficient queries, and walks the line between unsupervised learning, feature engineering, and data modeling. Neighbor-based approaches such as KDE are some of the most popular and useful density estimation techniques. KDE can be performed in any number of dimensions, though in practice high dimensionality can cause a degradation of performance. The KDE Modeling and KDE Simulation nodes in SPSS Modeler expose the core features and commonly used parameters of the KDE library. The nodes are implemented in Python. [1]

To use a KDE node, you must set up an upstream Type node. The KDE node will read input values from the Type node (or the Types tab of an upstream source node).

The **KDE Modeling** node is available on SPSS Modeler's Modeling tab and Python tab. The KDE Modeling node generates a model nugget, and the nugget's scored values are kernel density values from the input data.

The **KDE Simulation** node is available on the Output tab and the Python tab. The KDE Simulation node generates a KDE Gen source node that can create some records that have the same distribution as the input data. The KDE Gen node includes a Settings tab where you can specify how many records the node will create (default is 1) and generate a random seed.

For more information about KDE, including examples, see the KDE documentation available at http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation. [1]

[1] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

KDE Modeling node and KDE Simulation node Fields

The Fields tab specifies which fields are used in the analysis.

Use predefined roles. This option uses the input settings from an upstream Type node (or the Types tab of an upstream source node).

Use custom field assignments. To manually assign inputs, select this option.

Fields. Use the arrow buttons to assign items manually from this list to the Inputs list on the right of the screen. The icons indicate the valid measurement levels for each field. To select all the fields in the list, click the **All** button, or click an individual measurement level button to select all fields with that measurement level.

Inputs. Select one or more fields as inputs for clustering. KDE can only deal with continuous fields.

KDE nodes Build Options

Use the Build Options tab to specify build options for the KDE nodes, including **basic options** for kernel density parameters and cluster labels, and **advanced options** such as tolerance, leaf size, and whether to use a breadth-first approach. For additional information about these options, see the following online resources:

- Kernel Density Estimation Python API Parameter Reference[1]
- Kernel Density Estimation User Guide[2]

Basic

Bandwidth. Specify the bandwidth of the kernel.

Kernel. Select the kernel to use. Available kernels for the KDE Modeling node are **Gaussian**, **Tophat**, **Epanechnikov**, **Eponential**, **Linear**, or **Cosine**. Available kernels for the KDE Simulation node are **Gaussian** or **Tophat**. For details about these available kernels, see the Kernel Density Estimation User Guide.[2]

Algorithm. Select **Auto**, **Ball Tree** or **KD Tree** for the tree algorithm to use. For more information, see Ball Tree[3] and KD Tree.[4]

Metric. Select a distance metric. Available metrics are **Euclidean**, **Braycurtis**, **Chebyshev**, **Canberra**, **Cityblock**, **Dice**, **Hamming**, **Infinity**, **Jaccard**, **L1**, **L2**, **Matching**, **Manhattan**, **P**, **Rogerstanimoto**, **Russellrao**, **Sokalmichener**, **Sokalsneath**, **Kulsinski**, or **Minkowski**. If you select **Minkowski**, set the **P Value** as desired.

The metrics available in this drop-down will vary depending on which algorithm you choose. Also note that the normalization of the density output is correct only for the Euclidean distance metric.

Advanced

Absolute Tolerance. Specify the desired absolute tolerance of the result. A larger tolerance will generally result in faster run time. Default is **0.0**.

Relative Tolerance. Specify the desired relative tolerance of the result. A larger tolerance will generally result in faster run time. Default is **1E-8**.

Leaf Size. Specify the leaf size of the underlying tree. Default is **40**. Changing the leaf size may significantly impact performance and required memory. For more information about the Ball Tree and KD Tree algorithms, see Ball Tree[3] and KD Tree.[4]

Breadth first. Select **True** if you want to use a breadth-first approach or **False** to use a depth-first approach.

The following table shows the relationship between the settings in the SPSS Modeler KDE node dialogs and the Python KDE library parameters.

Table 46. Node properties mapped to Python library parameters

SPSS Modeler setting	Script name (property name)	KDE parameter
Inputs	`inputs`	
Bandwidth	`bandwidth`	`bandwidth`
Kernel	`kernel`	`kernel`
Algorithm	`algorithm`	`algorithm`
Metric	`metric`	`metric`
P Value	`pValue`	`pValue`
Absolute Tolerance	`atol`	`atol`
Relative Tolerance	`rtol`	`Rtol`
Leaf Size	`leafSize`	`leafSize`
Breadth first	`breadthFirst`	`breadthFirst`

[1] "API Reference." *sklearn.neighbors.KernelDensity*. Web. © 2007-2018, scikit-learn developers.

[2] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

[3] "Ball Tree." *Five balltree construction algorithms*. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

[4] "K-D Tree." *Multidimensional binary search trees used for associative searching.* © 1975, Bentley, J.L., Communications of the ACM.

KDE Modeling node and KDE Simulation node Model Options

Model name. You can generate the model name automatically based on the target or ID field (or model type in cases where no such field is specified) or specify a custom name.

IBM SPSS Statistics Helper Applications

If a compatible version of IBM SPSS Statistics is installed and licensed on your computer, you can configure IBM SPSS Modeler to process data with IBM SPSS Statistics functionality using the Statistics Transform, Statistics Model, Statistics Output, or Statistics Export nodes.

For information on product compatibility with the current version of IBM SPSS Modeler, see the corporate Support site at *http://www.ibm.com/support*.

To configure IBM SPSS Modeler to work with IBM SPSS Statistics and other applications, choose:

Tools > **Options** > **Helper Applications**

IBM SPSS Statistics Interactive. Enter the full path and name of the command (for example, `C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe`) to be used when launching IBM SPSS Statistics directly on a data file produced by the Statistics Export node. See "Statistics Export Node" on page 327 for more information.

Connection. If IBM SPSS Statistics Server is located on the same host as IBM SPSS Modeler Server, you can enable a connection between the two applications, which increases efficiency by leaving data on the server during analysis. Select **Server** to enable the **Port** option below. The default setting is **Local**.

Port. Specify the server port for IBM SPSS Statistics Server.

IBM SPSS Statistics Location Utility. To enable IBM SPSS Modeler to use the Statistics Transform, Statistics Model, and Statistics Output nodes, you must have a copy of IBM SPSS Statistics installed and licensed on the computer where the stream is run.

- If running IBM SPSS Modeler in local (stand-alone) mode, the licensed copy of IBM SPSS Statistics must be on the local computer. Click this button to specify the location of the local IBM SPSS Statistics installation you want to use for the licensing.
- In addition, if running in distributed mode against a remote IBM SPSS Modeler Server, you also need to run a utility at the IBM SPSS Modeler Server host to create the *statistics.ini* file, which indicates to IBM SPSS Statistics the installation path for IBM SPSS Modeler Server. To do this, from the command prompt, change to the IBM SPSS Modeler Server *bin* directory and, for Windows, run:

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  Alternatively, for UNIX, run:

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
  ```

If you do not have a licensed copy of IBM SPSS Statistics on your local machine, you can still run the Statistics File node against a IBM SPSS Statistics server, but attempts to run other IBM SPSS Statistics nodes will display an error message.

Comments

If you have trouble running the IBM SPSS Statistics procedure nodes, consider the following tips:

- If field names used in IBM SPSS Modeler are longer than eight characters (for versions prior to IBM SPSS Statistics 12.0), longer than 64 characters (for IBM SPSS Statistics 12.0 and subsequent versions), or contain invalid characters, it is necessary to rename or truncate them before reading them into IBM SPSS Statistics. See "Renaming or Filtering Fields for IBM SPSS Statistics " on page 328 for more information.

- If IBM SPSS Statistics was installed after IBM SPSS Modeler, you may need to specify the IBM SPSS Statistics location, as explained previously.

Chapter 7. Export Nodes

Overview of Export Nodes

Export nodes provide a mechanism for exporting data in various formats to interface with your other software tools.

The following export nodes are available:

The Database export node writes data to an ODBC-compliant relational data source. In order to write to an ODBC data source, the data source must exist and you must have write permission for it.

The Flat File export node outputs data to a delimited text file. It is useful for exporting data that can be read by other analysis or spreadsheet software.

The Statistics Export node outputs data in IBM SPSS Statistics *.sav* or *.zsav* format. The `.sav` or `.zsav` files can be read by IBM SPSS Statistics Base and other products. This is also the format used for cache files in IBM SPSS Modeler.

The Data Collection export node outputs data in the format used by Data Collection market research software. A Data Collection Data Library must be installed to use this node.

The IBM Cognos Export node exports data in a format that can be read by Cognos databases.

The IBM Cognos TM1 Export node exports data in a format that can be read by Cognos TM1 databases.

The SAS export node outputs data in SAS format, to be read into SAS or a SAS-compatible software package. Three SAS file formats are available: SAS for Windows/OS2, SAS for UNIX, or SAS Version 7/8.

The Excel export node outputs data in the Microsoft Excel `.xlsx` file format. Optionally, you can choose to launch Excel automatically and open the exported file when the node is executed.

The XML export node outputs data to a file in XML format. You can optionally create an XML source node to read the exported data back into the stream.

The JSON export node outputs data in JSON format.

Database Export Node

You can use Database nodes to write data to ODBC-compliant relational data sources, which are explained in the description of the Database source node. See the topic “Database source node” on page 14 for more information.

Use the following general steps to write data to a database:

1. Install an ODBC driver and configure a data source to the database you want to use.
2. On the Database node Export tab, specify the data source and table you want to write to. You can create a new table or insert data into an existing one.
3. Specify additional options as needed.

These steps are described in more detail in the next several topics.

Database Node Export Tab

Note: Some of the databases to which you can export might not support column names in tables that are more than 30 characters long. If an error message is displayed that mentions that your table has an incorrect column name, reduce the size of the name to fewer than 30 characters.

Data source. Shows the selected data source. Enter the name or select it from the drop-down list. If you don't see the desired database in the list, select **Add new database connection** and locate your database from the Database Connections dialog box. See “Adding a database connection” on page 15 for more information.

Table name. Enter the name of the table to which you want to send the data. If you select the **Insert into table** option, you can select an existing table in the database by clicking the **Select** button.

Create table. Select this option to create a new database table or to overwrite an existing database table.

Insert into table. Select this option to insert the data as new rows in an existing database table.

Merge table. (Where available) Select this option to update selected database columns with values from corresponding source data fields. Selecting this option enables the **Merge** button, which displays a dialog from where you can map source data fields to database columns.

Drop existing table. Select this option to delete any existing table with the same name when creating a new table.

Delete existing rows. Select this option to delete existing rows from the table before exporting, when inserting into a table.

Note: If either of the two options above are selected, you will receive an **Overwrite warning** message when you execute the node. To suppress the warnings, deselect **Warn when a node overwrites a database table** on the Notifications tab of the User Options dialog box.

Default string size. Fields you have marked as typeless in an upstream Type node are written to the database as string fields. Specify the size of strings to be used for typeless fields.

Click **Schema** to open a dialog box where you can set various export options (for databases that support this feature), set SQL data types for your fields, and specify the primary key for purposes of database indexing. See “Database export schema options” on page 316 for more information.

Click **Indexes** to specify options for indexing the exported table in order to improve database performance. See “Database Export Index Options” on page 318 for more information.

Click **Advanced** to specify bulk loading and database commit options. See “Database export advanced options” on page 319 for more information.

Quote table and column names. Select options used when sending a CREATE TABLE statement to the database. Tables or columns with spaces or nonstandard characters must be quoted.

- **As needed.** Select to allow IBM SPSS Modeler to automatically determine when quoting is needed on an individual basis.
- **Always.** Select to always enclose table and column names in quotes.
- **Never.** Select to disable the use of quotes.

Generate an import node for this data. Select to generate a Database source node for the data as exported to the specified data source and table. Upon execution, this node is added to the stream canvas.

Database Export Merge Options

This dialog enables you to map fields from the source data onto columns in the target database table. Where a source data field is mapped to a database column, the column value is replaced with the source data value when the stream is run. Unmapped source fields are left unchanged in the database.

Map Fields. This is where you specify the mapping between source data fields and database columns. Source data fields with the same name as columns in the database are mapped automatically.

- **Map.** Maps a source data field selected in the field list on the left of the button to a database column selected in the list on the right. You can map more than one field at a time, but the number of entries selected in both lists must be the same.
- **Unmap.** Removes the mapping for one or more selected database columns. This button is activated when you select a field or database column in the table on the right of the dialog.
- **Add.** Adds one or more source data fields selected in the field list on the left of the button to the list on the right ready for mapping. This button is activated when you select a field in the list on the left and no field with that name exists in the list on the right. Clicking this button maps the selected field to a new database column with the same name. The word **<NEW>** is displayed after the database column name to indicate that this is a new field.

Merge Rows. You use a key field, such as *Transaction ID*, to merge records with the same value in the key field. This is equivalent to a database "equi-join." Key values must be those of primary keys; that is, they must be unique, and cannot contain null values.

- **Possible keys.** Lists all fields found in all input data sources. Select one or more fields from this list and use the arrow button to add them as key fields for merging records. Any map field with a corresponding mapped database column is available as a key, except that fields added as new database columns (shown with **<NEW>** after the name) are not available.
- **Keys for merge.** Lists all fields used to merge records from all input data sources based on values of the key fields. To remove a key from the list, select one and use the arrow button to return it to the Possible Keys list. When more than one key field is selected, the option below is enabled.
- **Only include records which exist in database.** Performs a partial join; if the record is in the database and the stream, the mapped fields will be updated.
- **Add records to database.** Performs an outer join; all records in the stream will be merged (if the same record exists in the database) or added (if the record does not yet exist in the database).

To map a source data field to a new database column

1. Click the source field name in the list on the left, under **Map Fields**.
2. Click the **Add** button to complete the mapping.

To map a source data field to an existing database column

1. Click the source field name in the list on the left, under **Map Fields**.
2. Click the column name under **Database Column** on the right.
3. Click the **Map** button to complete the mapping.

To remove a mapping

1. In the list on the right, under Field, click the name of the field for which you want to remove the mapping.
2. Click the **Unmap** button.

To deselect a field in any of the lists

Hold down the CTRL key and click the field name.

Database export schema options

On the database export Schema dialog box, you can set options for database export (for databases that support these options), set SQL data types for your fields, specify which fields are primary keys, and customize the CREATE TABLE statement generated upon export.

The dialog box has several parts:

- The section at the top (if displayed) contains options for export to a database that supports these options. This section is not displayed if you are not connected to such a database.
- The text field in the center displays the template used to generate the CREATE TABLE command, which by default follows the format:

  ```
  CREATE TABLE <table-name> <(table columns)>
  ```

- The table in the lower portion enables you to specify the SQL data type for each field and to indicate which fields are primary keys as discussed below. The dialog box automatically generates the values of the <table-name> and <(table columns)> parameters based on the specifications in the table.

Setting database export options

If this section is displayed, you can specify a number of settings for export to the database. The database types that support this feature are as follows.

- SQL Server Enterprise and Developer editions. See the topic "Options for SQL Server" on page 317 for more information.
- Oracle Enterprise or Personal editions. See the topic "Options for Oracle" on page 317 for more information.

Customizing CREATE TABLE statements

Using the text field portion of this dialog box, you can add extra database-specific options to the CREATE TABLE statement.

1. Select the **Customize CREATE TABLE command** check box to activate the text window.
2. Add any database-specific options to the statement. Be sure to retain the text <table-name> and (<table-columns>) parameters because these are substituted for the real table name and column definitions by IBM SPSS Modeler.

Setting SQL data types

By default, IBM SPSS Modeler enables the database server to assign SQL data types automatically. To override the automatic type for a field, find the row corresponding to the field and select the desired type from the drop-down list in the *Type* column of the schema table. You can use Shift-click to select more than one row.

For types that take a length, precision, or scale argument (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC, and NUMBER), you should specify a length rather than allow the database server to assign an automatic length. For example, specifying a sensible value, such as VARCHAR(25), for length ensures that the storage type in IBM SPSS Modeler will be overwritten if that is your intention. To override the automatic assignment, select **Specify** from the Type drop-down list and replace the type definition with the desired SQL type definition statement.

The easiest way to do this is to first select the type that is closest to the desired type definition and then select **Specify** to edit that definition. For example, to set the SQL data type to VARCHAR(25), first set the type to **VARCHAR(length)** from the Type drop-down list, and then select **Specify** and replace the text length with the value 25.

Primary keys

If one or more columns in the exported table must have a unique value or combination of values for every row, you can indicate this by selecting the **Primary Key** check box for each field that applies. Most databases will not allow the table to be modified in a manner that invalidates a primary key constraint and will automatically create an index over the primary key to help enforce this restriction. (Optionally, you can create indexes for other fields in the Indexes dialog box. See the topic "Database Export Index Options" on page 318 for more information.)

Options for SQL Server

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Row.** Enables row-level compression (for example, the equivalent of `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW);` in SQL).
- **Page.** Enables page-level compression (for example, `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE);` in SQL).

Options for Oracle

Oracle settings - Basic option

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Default.** Enables default compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS;` in SQL). In this case, it has the same effect as the **Basic** option.
- **Basic.** Enables basic compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS BASIC;` in SQL).

Oracle settings - Advanced option

Use compression. If selected, creates tables for export with compression.

Compression for. Choose the level of compression.

- **Default.** Enables default compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS;` in SQL). In this case, it has the same effect as the **Basic** option.
- **Basic.** Enables basic compression (for example, `CREATE TABLE MYTABLE(…) COMPRESS BASIC;` in SQL).
- **OLTP.** Enables OLTP compression (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP;` in SQL).
- **Query Low/High.** (Exadata servers only) Enables hybrid columnar compression for query (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW;` or `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH;` in SQL). Compression for query is useful in data warehousing environments; `HIGH` provides a higher compression ratio than `LOW`.
- **Archive Low/High.** (Exadata servers only) Enables hybrid columnar compression for archive (for example, `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW;` or `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH;` in SQL). Compression for archive is useful for

compressing data that will be stored for long periods of time; HIGH provides a higher compression ratio than LOW.

Database Export Index Options

The Indexes dialog box enables you to create indexes on database tables exported from IBM SPSS Modeler. You can specify the field sets you want to include and customize the CREATE INDEX command, as needed.

The dialog box has two parts:

- The text field at the top displays a template that can be used to generate one or more CREATE INDEX commands, which by default follows the format:

  ```
  CREATE INDEX <index-name> ON <table-name>
  ```

- The table in the lower portion of the dialog box enables you to add specifications for each index you want to create. For each index, specify the index name and the fields or columns to include. The dialog box automatically generates the values of the <index-name> and <table-name> parameters accordingly.

  For example, the generated SQL for a single index on the fields *empid* and *deptid* might look like this:

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  You can add multiple rows to create multiple indexes. A separate CREATE INDEX command is generated for each row.

Customizing the CREATE INDEX Command

Optionally, you can customize the CREATE INDEX command for all indexes or for a specific index only. This gives you the flexibility to accommodate specific database requirements or options and to apply customizations to all indexes or only specific ones, as needed.

- Select **Customize CREATE INDEX command** at the top of the dialog box to modify the template used for all indexes added subsequently. Note that changes will not automatically apply to indexes that have already been added to the table.
- Select one or more rows in the table and then click **Update selected indexes** at the top of the dialog box to apply the current customizations to all selected rows.
- Select the **Customize** check box in each row to modify the command template for that index only.

Note that the values of the <index-name> and <table-name> parameters are generated automatically by the dialog box based on the table specifications and cannot be edited directly.

BITMAP KEYWORD. If you are using an Oracle database, you can customize the template to create a bitmap index rather than a standard index, as follows:

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

Bitmap indexes may be useful for indexing columns with a small number of distinct values. The resulting SQL might look this:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE keyword. Most databases support the UNIQUE keyword in the CREATE INDEX command. This enforces a uniqueness constraint similar to a primary key constraint on the underlying table.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Note that for fields actually designated as primary keys, this specification is not necessary. Most databases will automatically create an index for any fields specified as primary key fields within the CREATE TABLE command, so explicitly creating indexes on these fields is not necessary. See the topic "Database export schema options" on page 316 for more information.

FILLFACTOR keyword. Some physical parameters for the index can be fine-tuned. For example, SQL Server enables the user to trade off the index size (after initial creation) against the costs of maintenance as future changes are made to the table.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Other Comments

- If an index already exists with the specified name, index creation will fail. Any failures will initially be treated as warnings, allowing subsequent indexes to be created and then re-reported as an error in the message log after all indexes have been attempted.
- For best performance, indexes should be created after data has been loaded into the table. Indexes must contain at least one column.
- Before executing the node, you can preview the generated SQL in the message log.
- For temporary tables written to the database (that is, when node caching is enabled) the options to specify primary keys and indexes are not available. However the system may create indexes on the temporary table as appropriate, depending on how the data is used in downstream nodes. For example, if cached data is subsequently joined by a *DEPT* column, it would make sense to index the cached tabled on this column.

Indexes and Query Optimization

In some database management systems, once a database table has been created, loaded, and indexed, a further step is required before the optimizer is able to utilize the indexes to speed up query execution on the new table. For example, in Oracle, the cost-based query optimizer requires that a table be analyzed before its indexes can be used in query optimization. The internal ODBC properties file for Oracle (not user-visible) contains an option to make this happen, as follows:

```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

This step is executed whenever a table is created in Oracle (regardless of whether primary keys or indexes are defined). If necessary, the ODBC properties file for additional databases can be customized in a similar way - contact Support for assistance.

Database export advanced options

When you click the Advanced button from the Database export node dialog box, a new dialog box is displayed in which you can specify technical details for exporting results to a database.

Use batch commit. Select to turn off row-by-row commits to the database.

Batch size. Specifies the number of records to send to the database before committing to memory. Lowering this number provides greater data integrity at the cost of slower transfer speeds. You may want to fine-tune this number for optimal performance with your database.

Use bulk loading. Specifies a method for bulk loading data to the database directly from IBM SPSS Modeler. Some experimentation may be required to select which bulk load options are appropriate for a particular scenario.

- **Via ODBC.** Select to use the ODBC API to execute multiple-row inserts with greater efficiency than normal export to the database. Choose from row-wise or column-wise binding in the options below.
- **Via external loader.** Select to use a custom bulk loader program specific to your database. Selecting this option activates a variety of options below.

Advanced ODBC Options. These options are available only when **Via ODBC** is selected. Note that this functionality may not be supported by all ODBC drivers.

- **Row-wise.** Select row-wise binding to use the `SQLBulkOperations` call for loading data into the database. Row-wise binding typically improves speed compared to the use of parameterized inserts that insert data on a record-by-record basis.
- **Column-wise.** Select to use column-wise binding for loading data into the database. Column-wise binding improves performance by binding each database column (in a parameterized `INSERT`

statement) to an array of *N* values. Executing the INSERT statement once causes *N* rows to be inserted into the database. This method can dramatically increase performance.

External Loader Options. When **Via external loader** is specified, a variety of options are displayed for exporting the dataset to a file and specifying and executing a custom loader program to load the data from that file into the database. IBM SPSS Modeler can interface with external loaders for many popular database systems. Several scripts have been included with the software and are available along with technical documentation under the `scripts` subdirectory. Note that in order to use this functionality, Python 2.7 must be installed on the same machine as IBM SPSS Modeler or IBM SPSS Modeler Server, and the `python_exe_path` parameter must be set in the `options.cfg` file. See the topic "Bulk loader programming" on page 320 for more information.

- **Use delimiter.** Specifies which delimiter character should be used in the exported file. Select **Tab** to delimit with tab and **Space** to delimit with spaces. Select **Other** to specify another character, such as a comma (,).
- **Specify data file.** Select to enter the path to use for the data file written during bulk loading. By default, a temporary file is created in the temp directory on the server.
- **Specify loader program.** Select to specify a bulk loading program. By default, the software searches the `scripts` subdirectory of the IBM SPSS Modeler installation for a Python script to execute for a given database. Several scripts have been included with the software and are available along with technical documentation under the `scripts` subdirectory.
- **Generate log.** Select to generate a log file to the specified directory. The log file contains error information and is useful if the bulk load operation fails.
- **Check table size.** Select to perform table checking that ensures that the increase in table size corresponds to the number of rows exported from IBM SPSS Modeler.
- **Extra loader options.** Specifies additional arguments to the loader program. Use double quotes for arguments containing spaces.

Double quotes are included in optional arguments by escaping with a backslash. For example, the option specified as `-comment "This is a \"comment\""` includes both the `-comment` flag and the comment itself rendered as `This is a "comment"`.

A single backslash can be included by escaping with another backslash. For example, the option specified as `-specialdir "C:\\Test Scripts\\"` includes the flag `-specialdir` and the directory rendered as `C:\Test Scripts\`.

Bulk loader programming

The Database export node includes options for bulk loading on the Advanced Options dialog box. Bulk loader programs can be used to load data from a text file into a database.

The option **Use bulk loading - via external loader** configures IBM SPSS Modeler to do three things:

- Create any required database tables.
- Export the data to a text file.
- Invoke a bulk loader program to load the data from this file into the database table.

Typically the bulk loader program is not the database load utility itself (for example, Oracle's sqlldr utility) but a small script or program which forms the correct arguments, creates any database-specific auxiliary files (such as a control file) and then invokes the database load utility. The information in the following sections will help you edit an existing bulk loader.

Alternatively you can write your own program for bulk loading. See the topic "Developing bulk loader programs" on page 324 for more information. Note that this is not covered under a standard Technical Support agreement, and you should contact an IBM Services representative for assistance.

Scripts for bulk loading

IBM SPSS Modeler ships with a number of bulk loader programs for different databases that are implemented using Python scripts. When you run a stream containing a Database export node with the

Via external loader option selected, IBM SPSS Modeler creates the database table (if required) via ODBC, exports the data to a temporary file on the host running IBM SPSS Modeler Server, then invokes the bulk load script. This script in turn executes utilities provided by the DBMS vendor to upload data from the temporary files to the database.

Note: The IBM SPSS Modeler installation does not include a Python runtime interpreter, so a separate installation of Python is required. See the topic "Database export advanced options" on page 319 for more information.

Scripts are provided (in the *\scripts* folder of the IBM SPSS Modeler installation directory) for the databases listed in the following table.

Table 47. Bulk loader scripts provided

Database	Script name	Further information
IBM Db2	`db2_loader.py`	See the topic "Bulk loading data to IBM Db2 databases" on page 321 for more information.
IBM Netezza	`netezza_loader.py`	See the topic "Bulk loading data to IBM Netezza databases" on page 322 for more information.
Oracle	`oracle_loader.py`	See the topic "Bulk loading data to Oracle databases" on page 322 for more information.
SQL Server	`mssql_loader.py`	See the topic "Bulk loading data to SQL Server databases" on page 323 for more information.
Teradata	`teradata_loader.py`	See the topic "Bulk loading data to Teradata databases" on page 323 for more information.

Bulk loading data to IBM Db2 databases

The following points may help you to configure for bulk loading from IBM SPSS Modeler to an IBM Db2 database using the External Loader option in the DB Export Advanced Options dialog box.

Ensure that the Db2 command line processor (CLP) utility is installed

The script `db2_loader.py` invokes the Db2 LOAD command. Ensure that the command line processor (db2 on UNIX, `db2cmd` on Windows) is installed on the server on which `db2_loader.py` is to be executed (typically, the host running IBM SPSS Modeler Server).

Check whether the local database alias name is the same as the actual database name

The Db2 local database alias is the name used by Db2 client software to refer to a database in a local or remote Db2 instance. If the local database alias is different from the name of the remote database, supply the extra loader option:

```
-alias <local_database_alias>
```

For example, the remote database is named STARS on host GALAXY but the Db2 local database alias on the host running IBM SPSS Modeler Server is STARS_GALAXY. Use the extra loader option

```
-alias STARS_GALAXY
```

Non-ASCII character data encoding

If you are bulk loading data that is not in ASCII format, you should ensure that the codepage variable in the configuration section of `db2_loader.py` is correctly set up on your system.

Blank strings

Blank strings are exported to the database as NULL values.

Bulk loading data to IBM Netezza databases

The following points may help you to configure for bulk loading from IBM SPSS Modeler to an IBM Netezza database using the External Loader option in the DB Export Advanced Options dialog box.

Ensure that the Netezza nzload utility is installed

The script *netezza_loader.py* invokes the Netezza utility *nzload*. Ensure that *nzload* is installed and correctly configured on the server on which *netezza_loader.py* is to be executed.

Exporting non-ASCII data

If your export contains data that is not in ASCII format, you may need to add `-encoding UTF8` to the **Extra loader options** field in the DB Export Advanced Options dialog box. This should ensure that non-ASCII data is correctly uploaded.

Date, Time and Timestamp format data

In the stream properties, set the date format to **DD-MM-YYYY** and the time format to **HH:MM:SS**.

Blank strings

Blank strings are exported to the database as NULL values.

Different order of columns in stream and target table when inserting data into an existing table

If the order of columns in the stream is different from that in the target table, data values will be inserted into the wrong columns. Use a Field Reorder node to ensure that the order of columns in the stream matches the order in the target table. See the topic "Field Reorder Node" on page 161 for more information.

Tracking nzload progress

When running IBM SPSS Modeler in local mode, add `-sts` to the **Extra loader options** field in the DB Export Advanced Options dialog box to see status messages every 10000 rows in the command window opened by the *nzload* utility.

Bulk loading data to Oracle databases

The following points may help you to configure for bulk loading from IBM SPSS Modeler to an Oracle database using the External Loader option in the DB Export Advanced Options dialog box.

Ensure that the Oracle sqlldr utility is installed

The script *oracle_loader.py* invokes the Oracle utility *sqlldr*. Note that *sqlldr* is not automatically included in Oracle Client. Ensure that *sqlldr* is installed on the server on which *oracle_loader.py* is to be executed.

Specify the database SID or service name

If you are exporting data to a non-local Oracle server or your local Oracle server has multiple databases, you will need to specify the following in the **Extra loader options** field in the DB Export Advanced Options dialog box to pass in the SID or service name:

```
-database <SID>
```

Editing the configuration section in oracle_loader.py

On UNIX (and optionally, on Windows) systems, edit the configuration section at the start of the *oracle_loader.py* script. Here, values for ORACLE_SID, NLS_LANG, TNS_ADMIN and ORACLE_HOME environment variables can be specified if appropriate, as well as the full path of the *sqlldr* utility.

Date, Time and Timestamp format data

In the stream properties, you should normally set the date format to **YYYY-MM-DD** and the time format to **HH:MM:SS**.

If you need to use different date and time formats from the above, consult your Oracle documentation and edit the *oracle_loader.py* script file.

Non-ASCII character data encoding

If you are bulk loading data that is not in ASCII format, you should ensure that the environment variable NLS_LANG is correctly set up on your system. This is read by the Oracle loader utility *sqlldr*. For example, the correct value for NLS_LANG for Shift-JIS on Windows is Japanese_Japan.JA16SJIS. For more details on NLS_LANG, check your Oracle documentation.

Blank strings

Blank strings are exported to the database as NULL values.

Bulk loading data to SQL Server databases

The following points may help you to configure for bulk loading from IBM SPSS Modeler to a SQL Server database using the External Loader option in the DB Export Advanced Options dialog box.

Ensure that the SQL Server bcp.exe utility is installed
The script *mssql_loader.py* invokes the SQL Server utility *bcp.exe*. Ensure that *bcp.exe* is installed on the server where *mssql_loader.py* is to be executed.

Using spaces as delimiter does not work
Avoid choosing space as the delimiter in the DB Export Advanced Options dialog box.

Check table size option recommended
We recommend that you enable the **Check table size** option in the DB Export Advanced Options dialog box. Failures in the bulk load process are not always detected, and enabling this option performs an extra check that the correct number of rows has been loaded.

Blank strings
Blank strings are exported to the database as NULL values.

Specify the fully qualified SQL server named instance
There may be occasions when SPSS Modeler cannot access SQL Server due to an unqualified hostname and displays the following error:

```
Error encountered while executing external bulk loader. The log file may provide more details.
```

To correct this error, add the following string, including the double-quotes, to the **Extra loader options** field:

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

Bulk loading data to Teradata databases

The following points may help you to configure for bulk loading from IBM SPSS Modeler to a Teradata database using the External Loader option in the DB Export Advanced Options dialog box.

Ensure that the Teradata fastload utility is installed

The script *teradata_loader.py* invokes the Teradata utility *fastload*. Ensure that *fastload* is installed and correctly configured on the server on which *teradata_loader.py* is to be run.

Data can be bulk loaded only to empty tables

You can only use empty tables as targets for a bulk load. If a target table contains any data prior to the bulk load, the operation will fail.

Date, Time and Timestamp format data

In the stream properties, set the date format to **YYYY-MM-DD** and the time format to **HH:MM:SS**.

Blank strings

Blank strings are exported to the database as NULL values.

Teradata process ID (tdpid)

By default, *fastload* exports data to the Teradata system with `tdpid=dbc`. Usually, there will be an entry in the HOSTS file which associates `dbccop1` with the IP address of the Teradata server. To use a different server, specify the following in the **Extra loader options** field in the DB Export Advanced Options dialog box to pass the `tdpid` of this server:

```
-tdpid <id>
```

Spaces in table and column names

If table or column names contain spaces, bulk load operation will fail. If possible, rename the table or column names to remove spaces.

Developing bulk loader programs

This topic explains how to develop a bulk loader program that can be run from IBM SPSS Modeler to load data from a text file into a database. Note that this is not covered under a standard Technical Support agreement, and you should contact an IBM Services representative for assistance.

Using Python to build bulk loader programs

By default, IBM SPSS Modeler searches for a default bulk loader program based on the database type. See Table 47 on page 321.

You can use the script *test_loader.py* to assist in developing batch loader programs. See the topic "Testing bulk loader programs" on page 326 for more information.

Objects passed to the bulk loader program

IBM SPSS Modeler writes two files that are passed to the bulk loader program.

- **Data file.** This contains the data to be loaded, in text format.
- **Schema file.** This is an XML file that describes the names and types of the columns, and provides information on how the data file is formatted (for example, what character is used as the delimiter between fields).

In addition, IBM SPSS Modeler passes other information such as table name, user name and password as arguments when invoking the bulk load program.

Note: To signal successful completion to IBM SPSS Modeler, the bulk loader program should delete the schema file.

Arguments passed to the bulk loader program

The arguments passed to the program are listed in the following table.

Table 48. Arguments passed to bulk loader

Argument	Description
`schemafile`	Path of the schema file.
`data file`	Path of the data file.
`servername`	Name of the DBMS server; may be blank.
`databasename`	Name of the database within the DBMS server; may be blank.
`username`	User name to log into database.
`password`	Password to log into database.
`tablename`	Name of the table to load.
`ownername`	Name of the table owner (also known as schema name).

Table 48. Arguments passed to bulk loader (continued)	
Argument	**Description**
logfilename	Name of the logfile (if blank, no log file is created).
rowcount	Number of rows in the dataset.

Any options specified in the **Extra loader options** field on the DB Export Advanced Options dialog box are passed to the bulk loader program after these standard arguments.

Format of the data file

Data is written to the data file in text format, with each field being separated by a delimiter character that is specified on the DB Export Advanced Options dialog box. Here is an example of how a tab-delimited data file might appear.

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

The file is written in the local encoding used by IBM SPSS Modeler Server (or IBM SPSS Modeler if not connected to IBM SPSS Modeler Server). Some formatting is controlled via the IBM SPSS Modeler stream settings.

Format of the schema file

The schema file is an XML file which describes the data file. Here is an example to accompany the preceding data file.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
    <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD"
time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
        <column name="Age" encoded_name="416765" type="integer"/>
        <column name="Sex" encoded_name="536578" type="char" size="1"/>
        <column name="BP" encoded_name="4250" type="char" size="6"/>
        <column name="Cholesterol" encoded_name="43686F6C65737465726F6C"
type="char" size="6"/>
        <column name="Na" encoded_name="4E61" type="real"/>
        <column name="K" encoded_name="4B" type="real"/>
        <column name="Drug" encoded_name="44727567" type="char" size="5"/>
    </table>
</DBSCHEMA>
```

The following two tables list the attributes of the <table> and <column> elements of the schema file.

Table 49. Attributes of the <table> element	
Attribute	**Description**
delimiter	The field delimiter character (TAB is represented as \t).
commit_every	The batch size interval (as on the DB Export Advanced Options dialog box).
date_format	The format used for representing dates.
time_format	The format used for representing times.
append_existing	true if the table to be loaded already contains data; false otherwise.
delete_datafile	true if the bulk load program should delete the data file on completion of loading.

Table 50. Attributes of the <column> element	
Attribute	**Description**
`name`	The column name.
`encoded_name`	The column name converted to the same encoding as the data file and output as a series of two-digit hexadecimal numbers.
`type`	The data type of the column: one of `integer`, `real`, `char`, `time`, `date`, and `datetime`.
`size`	For the `char` data type, the maximum width of the column in characters.

Testing bulk loader programs

You can test bulk loading using a test script *test_loader.py* included in the *\scripts* folder of the IBM SPSS Modeler installation directory. Doing so is useful when trying to develop, debug, or troubleshoot bulk load programs or scripts for use with IBM SPSS Modeler.

To use the test script, proceed as follows.

1. Run the *test_loader.py* script to copy the schema and data files to the files *schema.xml* and *data.txt*, and create a Windows batch file (*test.bat*).
2. Edit the *test.bat* file to select the bulk loader program or script to test.
3. Run *test.bat* from a command shell to test the chosen bulk load program or script.

Note: Running *test.bat* does not actually load data into the database.

Flat File Export Node

The Flat File export node enables you to write data to a delimited text file. This is useful for exporting data that can be read by other analysis or spreadsheet software.

If your data contains geospatial information you can export it as a flat file and, if you generate a Variable File source node for use within the same stream, all storage, measurement, and geospatial metadata is retained in the new source node. However if you export the data and then import it in a different stream, you must take some extra steps to set the geospatial metadata in the new source node. For more information, see the topic "Variable File Node" on page 24.

Note: You cannot write files in the old cache format, because IBM SPSS Modeler no longer uses that format for cache files. IBM SPSS Modeler cache files are now saved in IBM SPSS Statistics *.sav* format, which you can write using a Statistics export node. For more information, see the topic "Statistics Export Node" on page 327.

Flat File Export Tab

Export file. Specifies the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

Write mode. If **Overwrite** is selected, any existing data in the specified file will be overwritten. If **Append** is selected, output will be added to the end of the existing file, preserving any data it contains.

- **Include field names.** If this option is selected, field names will be written to the first line of the output file. This option is available only for the **Overwrite** write mode.

New line after each record. If this option is selected, each record will be written on a new line in the output file.

Field separator. Specifies the character to insert between field values in the generated text file. Options are **Comma**, **Tab**, **Space**, and **Other**. If you select **Other**, enter the desired delimiter character(s) in the text box.

Symbol quotes. Specifies the type of quoting to use for values of symbolic fields. Options are **None** (values are not quoted), **Single (')**, **Double (")**, and **Other**. If you select **Other**, enter the desired quoting character(s) in the text box.

Encoding. Specifies the text-encoding method used. You can choose between the system default, stream default, or UTF-8.

- The system default is specified in the Windows Control Panel or, if running in distributed mode, on the server computer.
- The stream default is specified in the Stream Properties dialog box.

Decimal symbol. Specifies how decimals should be represented in the data.

- **Stream default.** The decimal separator defined by the current stream's default setting will be used. This will normally be the decimal separator defined by the computer's locale settings.
- **Period (.).** The period character will be used as the decimal separator.
- **Comma (,).** The comma character will be used as the decimal separator.

Generate an import node for this data. Select this option to automatically generate a Variable File source node that will read the exported data file. See the topic "Variable File Node" on page 24 for more information.

Statistics Export Node

The Statistics Export node enables you to export data in IBM SPSS Statistics *.sav* format. IBM SPSS Statistics *.sav* files can be read by IBM SPSS Statistics Base and other modules. This is also the format used for IBM SPSS Modeler cache files.

Mapping IBM SPSS Modeler field names to IBM SPSS Statistics variable names can sometimes cause errors because IBM SPSS Statistics variable names are limited to 64 characters and cannot include certain characters, such as spaces, dollar signs ($), and dashes (–). There are two ways to adjust for these restrictions:

- You can rename fields conforming to IBM SPSS Statistics variable name requirements by clicking the Filter tab. See the topic "Renaming or Filtering Fields for IBM SPSS Statistics " on page 328 for more information.
- Choose to export both field names and labels from IBM SPSS Modeler.

Note: IBM SPSS Modeler writes *.sav* files in Unicode UTF-8 format. IBM SPSS Statistics only supports files in Unicode UTF-8 format from release 16.0 onwards. To prevent the possibility of data corruption *.sav* files saved with Unicode encoding should not be used in releases of IBM SPSS Statistics prior to 16.0. For more information, see the IBM SPSS Statistics help.

Multiple response sets. Any multiple response sets defined in the stream will automatically be preserved when the file is exported. You can view and edit multiple response sets from any node with a Filter tab. See the topic "Editing Multiple Response Sets" on page 135 for more information.

Statistics Export Node - Export Tab

Export file Specifies the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

File type Select if the file is to be saved in normal *.sav* or compressed.*zsav* format.

Encrypt file with password To protect the file with a password, select this box; you are the prompted to enter and confirm the **Password** in a separate dialog box.

Note: Password protected files can only be opened by SPSS Modeler version 16 or greater, or by SPSS Statistics version 21 or greater.

Export field names Specifies a method of handling variable names and labels upon export from SPSS Modeler to an SPSS Statistics *.sav* or *.zsav* file.

- **Names and variable labels** Select to export both SPSS Modeler field names and field labels. Names are exported as SPSS Statistics variable names, while labels are exported as SPSS Statistics variable labels.
- **Names as variable labels** Select to use the SPSS Modeler field names as variable labels in SPSS Statistics. SPSS Modeler allows characters in field names that are invalid in SPSS Statistics variable names. To prevent possibly creating invalid SPSS Statistics names, select **Names as variable labels** instead, or use the Filter tab to adjust field names.

Launch Application If SPSS Statistics is installed on your computer, you can select this option to invoke the application directly on the saved data file. Options for launching the application must be specified in the Helper Applications dialog box. See the topic " IBM SPSS Statistics Helper Applications" on page 311 for more information. To simply create an SPSS Statistics *.sav* or *.zsav* file without opening an external program, deselect this option.

Note: When running SPSS Modeler and SPSS Statistics together in Server (distributed) mode, writing the data out and launching a SPSS Statistics session does not automatically open a SPSS Statistics client showing the data set read into the active data set. The workaround is to manually open the data file in the SPSS Statistics client once it is launched.

Generate an import node for this data Select this option to automatically generate a Statistics File source node that will read the exported data file. See the topic "Statistics File Node" on page 28 for more information.

Renaming or Filtering Fields for IBM SPSS Statistics

Before exporting or deploying data from IBM SPSS Modeler to external applications such as IBM SPSS Statistics, it may be necessary to rename or adjust field names. The Statistics Transform, Statistics Output, and Statistics Export dialog boxes contain a Filter tab to facilitate this process.

A basic description of Filter tab functionality is discussed elsewhere. See the topic "Setting filtering options" on page 134 for more information. This topic provides tips for reading data into IBM SPSS Statistics.

To adjust field names to conform to IBM SPSS Statistics naming conventions:

1. On the Filter tab, click the Filter Options Menu toolbar button (the first one on the toolbar).
2. Select Rename For IBM SPSS Statistics.
3. On the Rename For IBM SPSS Statistics dialog, you can choose to replace invalid characters in filenames with either a **Hash (#)** character or an **Underscore (_)**.

Rename multi response sets. Select this option if you want to adjust the names of multiple response sets, which can be imported into IBM SPSS Modeler using a Statistics File source node. They are used to record data that can have more than one value for each case, such as in survey responses.

Data Collection Export Node

The Data Collection export node saves data in the format used by Data Collection market research software, based on the Data Collection Data Model. This format distinguishes case data—the actual responses to questions gathered during a survey—from the metadata that describes how the case data is collected and organized. Metadata consists of information such as question texts, variable names and descriptions, multiple-response sets, translations of the various texts, and definitions of the structure of the case data. See the topic " Data Collection Node" on page 29 for more information.

Metadata file. Specifies the name of the questionnaire definition file (*.mdd*) where the exported metadata will be saved. A default questionnaire is created based on field type information. For example, a nominal (set) field could be represented as a single question with the field description used as the question text and a separate check box for each defined value.

Merge metadata. Specifies whether the metadata will overwrite existing versions or be merged with existing metadata. If the merge option is selected, a new version is created each time the stream is run. This makes it possible to track versions of a questionnaire as it undergoes changes. Each version can be regarded as a snapshot of the metadata used to collect a particular set of case data.

Enable system variables. Specifies whether system variables are included in the exported `.mdd` file. These include variables such as *Respondent.Serial*, *Respondent.Origin*, and *DataCollection.StartTime*.

Case data settings. Specifies the IBM SPSS Statistics data (*.sav*) file where case data is exported. Note that all the restrictions on variable and value names apply here, so for example you may need to switch to the Filter tab and use the "Rename for IBM SPSS Statistics" option on the Filter options menu to correct invalid characters in field names.

Generate an import node for this data. Select this option to automatically generate a Data Collection source node that will read the exported data file.

Multiple response sets. Any multiple response sets defined in the stream will automatically be preserved when the file is exported. You can view and edit multiple response sets from any node with a Filter tab. See the topic "Editing Multiple Response Sets" on page 135 for more information.

Analytic Server Export node

The Analytic Server Export node enables you to write data from your analysis to an existing Analytic Server data source. This can, for example, be text files on Hadoop Distributed File System (HDFS) or a database.

Typically, a stream with an Analytic Server Export node also begins with Analytic Server Source nodes, and is submitted to the Analytic Server and executed on HDFS. Alternatively, a stream with "local" data sources can end with an Analytic Server Export node in order to upload relatively small datasets (no more than 100,000 records) for use with Analytic Server.

If you want to use your own Analytic Server connection instead of the default connection defined by your administrator, deselect **Use default Analytic Server** and select your connection.

Data source. Select a data source containing the data you wish to use. A data source contains the files and metadata associated with that source. Click **Select** to display a list of available data sources. See the topic "Selecting a data source" on page 10 for more information.

If you need to create a new data source or edit an existing one, click **Launch Data Source Editor...**.

Mode. Select **Append** to add to the existing data source, or **Overwrite** to replace the contents of the data source.

Generate an Import node for this data. Select to generate a source node for the data as exported to the specified data source. This node is added to the stream canvas.

Note that using multiple Analytic Server connections can be useful in controlling the flow of data. For example, when using the Analytic Server Source and Export nodes, you may want to use different Analytic Server connections in different branches of a stream so that when each branch runs it uses its own Analytic Server and no data will be pulled to the IBM SPSS Modeler Server. Note that if a branch contains more than one Analytic Server connection, the data will be pulled from the Analytic Servers to the IBM SPSS Modeler Server.

IBM Cognos export node

The IBM Cognos Export node enables you to export data from an IBM SPSS Modeler stream to Cognos Analytics, in UTF-8 format. In this way, Cognos can make use of transformed or scored data from IBM SPSS Modeler. For example, you could use Cognos Report Studio to create a report based on the exported data, including the predictions and confidence values. The report could then be saved on the Cognos server and distributed to Cognos users.

Note: You can export only relational data, not OLAP data.

To export data to Cognos, you need to specify the following:

- Cognos connection - the connection to the Cognos Analytics server
- ODBC connection - the connection to the Cognos data server that the Cognos server uses

Within the Cognos connection you specify a Cognos data source to use. This data source must use the same login as the ODBC data source.

You export the actual stream data to the data server, and the package metadata to the Cognos server.

As with any other export node, you can also use the Publish tab of the node dialog box to publish the stream for deployment using IBM SPSS Modeler Solution Publisher.

Note: The Cognos source node only supports Cognos CQM packages. DQM packages are not supported.

Cognos connection

This is where you specify the connection to the Cognos Analytics server that you want to use for the export. The procedure involves exporting the metadata to a new package on the Cognos server, while the stream data is exported to the Cognos data server.

Connection. Click the **Edit** button to display a dialog box where you can define the URL and other details of the Cognos server to which you want to export the data. If you are already logged on to a Cognos server through IBM SPSS Modeler, you can also edit the details of the current connection. See "Cognos connections" on page 36 for more information.

Data source. The name of the Cognos data source (typically a database) to which you are exporting the data. The drop-down list shows all the Cognos data sources that you can access from the current connection. Click the **Refresh** button to update the list.

Folder. The path and name of the folder on the Cognos server where the export package is to be created.

Package name. The name of the package in the specified folder that is to contain the exported metadata. This must be a new package with a single query subject; you cannot export to an existing package.

Mode. Specifies how you want to perform the export:

- **Publish package now.** (default) Performs the export operation as soon as you click **Run**.
- **Export action script.** Creates an XML script that you can run later (for example, using Framework Manager) to perform the export. Type the path and file name for the script in the **File** field, or use the **Edit** button to specify the name and location of the script file.

Generate an import node for this data. Select to generate a source node for the data as exported to the specified data source and table. When you click **Run**, this node is added to the stream canvas.

ODBC connection

Here you specify the connection to the Cognos data server (that is, the database) to which the stream data is to be exported.

Note: You must ensure that the data source you specify here points to the same one specified on the **Cognos connections** panel. You must also ensure that the Cognos connection data source uses the same login as the ODBC data source.

Data source. Shows the selected data source. Enter the name or select it from the drop-down list. If you don't see the desired database in the list, select **Add new database connection** and locate your database from the Database Connections dialog box. See "Adding a database connection" on page 15 for more information.

Table name. Enter the name of the table to which you want to send the data. If you select the **Insert into table** option, you can select an existing table in the database by clicking the **Select** button.

Create table. Select this option to create a new database table or to overwrite an existing database table.

Insert into table. Select this option to insert the data as new rows in an existing database table.

Merge table. (Where available) Select this option to update selected database columns with values from corresponding source data fields. Selecting this option enables the **Merge** button, which displays a dialog from where you can map source data fields to database columns.

Drop existing table. Select this option to delete any existing table with the same name when creating a new table.

Delete existing rows. Select this option to delete existing rows from the table before exporting, when inserting into a table.

Note: If either of the two options above are selected, you will receive an **Overwrite warning** message when you execute the node. To suppress the warnings, deselect **Warn when a node overwrites a database table** on the Notifications tab of the User Options dialog box.

Default string size. Fields you have marked as typeless in an upstream Type node are written to the database as string fields. Specify the size of strings to be used for typeless fields.

Click **Schema** to open a dialog box where you can set various export options (for databases that support this feature), set SQL data types for your fields, and specify the primary key for purposes of database indexing. See "Database export schema options" on page 316 for more information.

Click **Indexes** to specify options for indexing the exported table in order to improve database performance. See "Database Export Index Options" on page 318 for more information.

Click **Advanced** to specify bulk loading and database commit options. See "Database export advanced options" on page 319 for more information.

Quote table and column names. Select options used when sending a CREATE TABLE statement to the database. Tables or columns with spaces or nonstandard characters must be quoted.

- **As needed.** Select to allow IBM SPSS Modeler to automatically determine when quoting is needed on an individual basis.
- **Always.** Select to always enclose table and column names in quotes.
- **Never.** Select to disable the use of quotes.

Generate an import node for this data. Select to generate a source node for the data as exported to the specified data source and table. When you click **Run**, this node is added to the stream canvas.

IBM Cognos TM1 Export Node

The IBM Cognos Export node enables you to export data from an SPSS Modeler stream to Cognos TM1. In this way, Cognos Analytics can make use of transformed or scored data from SPSS Modeler.

Note: You can export only measures, not context dimension data; alternatively, you can add new elements to the cube.

To export data to Cognos Analytics, you need to specify the following:

- The connection to the Cognos TM1 server.
- The Cube into which the data will be exported.
- The mapping from the SPSS data names to the equivalent TM1 dimensions and measures.

Note: The TM1 user needs the following permissions: Write privilege of cubes, Read privilege of dimensions, and Write privilege of dimension elements. In addition, IBM Cognos TM1 10.2 Fix Pack 3, or later, is required before SPSS Modeler can import and export Cognos TM1 data. Existing streams that were based on previous versions will still function.

Administrator credentials are not required for this node. But if you are still using the old legacy pre-17.1 TM1 node, administrator credentials are still required.

SPSS Modeler only supports working with Cognos TM1 via IntegratedSecurityMode 1, 4, and 5.

As with any other export node, you can also use the Publish tab of the node dialog box to publish the stream for deployment using IBM SPSS Modeler Solution Publisher.

Note: Before you can use the TM1 Source or Export nodes in SPSS Modeler, you must verify some settings in the `tm1s.cfg` file; this is the TM1 server configuration file in the root directory of the TM1 server.

- HTTPPortNumber - set a valid port number; typically 1-65535. Note that this is not the port number you subsequently specify in the connection in the node; it is an internal port used by TM1 that is disabled by default. If necessary, contact your TM1 administrator to confirm the valid setting for this port.
- UseSSL - if you set this to *True*, HTTPS is used as the transport protocol. In this case you must import the TM1 certification to the SPSS Modeler Server JRE.

Connecting to an IBM Cognos TM1 cube to export data

The first step in exporting data to an IBM Cognos TM1 database is to select the relevant TM1 administration host , and associated server and cube, on the **Connection** tab of the IBM Cognos TM1 dialog box.

Note: Only actual "null" values will be discarded when exporting data to TM1. Zero (0) values will be exported as a valid value. Also note that only fields with a storage type of *string* can be mapped to dimensions on the Mapping tab. Before exporting to TM1, you must use IBM SPSS Modeler client to convert non-string data types to string.

Admin Host Type the URL of the administration host where the TM1 server you want to connect to is installed. The administration host is defined as a single URL for all TM1 servers. From this URL, all IBM Cognos TM1 servers installed and running on your environment can be discovered and accessed.

TM1 Server When you have established the connection to the administration host, select the server that contains the data you want to import and click **Login**. If you have not previously connected to this server you are prompted to enter your **User name** and **Password**; alternatively, you can search for previously entered login details that you have saved as a **Stored Credential**.

Select a TM1 cube to export Displays the name of the cubes within the TM1 server into which you can export data.

To choose the data to export, select the cube and click the right arrow to move the cube into the **Export to cube** field. When you have selected the cube, use the Mapping tab to map the TM1 dimensions and measures to the relevant SPSS fields or to a fixed value (*Select* operation).

Mapping IBM Cognos TM1 data for export

After you select the TM1 administration host and associated TM1 server and cube, use the Mapping tab of the IBM Cognos TM1 Export dialog box to map TM1 dimensions and measures to SPSS fields or set TM1 dimensions to a fixed value.

Note: Only fields with a storage type of *string* can be mapped to dimensions. Before exporting to TM1, you must use IBM SPSS Modeler client to convert non-string data types to string.

Fields Lists the data field names from the SPSS data file that are available for export.

TM1 Dimensions Shows the TM1 cube selected on the Connection tab, along with its regular dimensions, the measure dimension, and the elements of the selected measure dimension. Select the name of the TM1 dimension, or measure, to map to the SPSS data field.

The following options are available on the Mapping tab.

Select a measure dimension From the list of dimensions for the selected cube, select one to be the measure dimension.

When you select a dimension, except the measure dimension, and click **Select**, a dialog is displayed that shows the leaf elements of the selected dimension. You can only select leaf elements. Selected elements are labeled with an **S**.

Map Maps the selected SPSS data field to the selected TM1 dimension or measure (a regular dimension or a specific measure or element from the measure dimension). Mapped fields are labeled with an **M**.

Unmap Unmaps the selected SPSS data field from the selected TM1 dimension or measure. Note that you can only unmap a single mapping at a time. The unmapped SPSS data field is moved back to the left column.

Create New Creates a new measure in the TM1 measure dimension. A dialog is displayed in which you enter the new **TM1 measure name**. This option is only available for measure dimensions, not for regular dimensions.

For more information about TM1, see the IBM Cognos TM1 documentation at http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html.

SAS Export Node

This feature is available in SPSS Modeler Professional and SPSS Modeler Premium.

The SAS export node enables you to write data in SAS format to be read into SAS or a SAS-compatible software package. You can export in three SAS file formats: SAS for Windows/OS2, SAS for UNIX, or SAS.

SAS Export Node Export Tab

Export file. Specify the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

Export. Specify the export file format. Options are **SAS for Windows/OS2**, **SAS for UNIX**, or **SAS Version 7/8/9**.

Export field names. Select options for exporting field names and labels from IBM SPSS Modeler for use with SAS.

- **Names and variable labels.** Select to export both IBM SPSS Modeler field names and field labels. Names are exported as SAS variable names, while labels are exported as SAS variable labels.
- **Names as variable labels.** Select to use the IBM SPSS Modeler field names as variable labels in SAS. IBM SPSS Modeler allows characters in field names that are invalid in SAS variable names. To prevent possibly creating invalid SAS names, select **Names and variable labels** instead.

Generate an import node for this data. Select this option to automatically generate a SAS source node that will read the exported data file. See the topic "SAS Source Node" on page 39 for more information.

Note: The maximum allowed length of a string is 255 bytes. If a string is more than 255 bytes, it will be truncated when exporting.

Excel Export Node

The Excel export node outputs data in Microsoft Excel `.xlsx` format. Optionally, you can choose to automatically launch Excel and open the exported file when the node is executed.

Excel Node Export Tab

File name. Enter a filename or click the file chooser button to browse to the file's location. The default filename is *excelexp.xlsx*.

File type. The Excel .xlsx file format is supported.

Create new file. Creates a new Excel file.

Insert into existing file. Content is replaced beginning at the cell designated by the **Start in cell** field. Other cells in the spreadsheet are left with their original content.

Include field names. Specifies whether field names should be included in the first row of the worksheet.

Start in cell. The cell location used for the first export record (or first field name if **Include field names** is checked). Data are filled to the right and down from this initial cell.

Choose worksheet. Specifies the worksheet to which you want to export the data. You can identify the worksheet either by index or by name:

- **By index.** If you are creating a new file, specify a number from 0 to 9 to identify the worksheet to which you want to export, beginning with 0 for the first worksheet, 1 for the second worksheet, and so on. You can use values of 10 or higher only if a worksheet already exists at this position.
- **By name.** If you are creating a new file, specify the name used for the worksheet. If you are inserting into an existing file, the data are inserted into this worksheet if it exists, otherwise a new worksheet with this name is created.

Launch Excel. Specifies whether Excel is automatically launched on the exported file when the node is executed. Note that when running in distributed mode against IBM SPSS Modeler Server, the output is saved to the server file system, and Excel is launched on the Client with a copy of the exported file.

Generate an import node for this data. Select this option to automatically generate an Excel source node that will read the exported data file. See the topic "Excel Source Node" on page 39 for more information.

Extension Export node

With the Extension Export node, you can run R or Python for Spark scripts to export data.

Extension Export node - Syntax tab

Select your type of syntax – **R** or **Python for Spark**. See the following sections for more information. When your syntax is ready, you can click **Run** to execute the Extension Export node.

R Syntax

R Syntax. You can enter, or paste, custom R scripting syntax for data analysis into this field.

Convert flag fields. Specifies how flag fields are treated. There are two options: **Strings to factor, Integers and Reals to double**, and **Logical values (True, False)**. If you select **Logical values (True, False)** the original values of the flag fields are lost. For example, if a field has values Male and Female, these are changed to True and False.

Convert missing values to the R 'not available' value (NA). When selected, any missing values are converted to the R NA value. The value NA is used by R to identify missing values. Some R functions that you use might have an argument that can be used to control how the function behaves when the data contain NA. For example, the function might allow you to choose to automatically exclude records that contain NA. If this option is not selected, any missing values are passed to R unchanged, and might cause errors when your R script is executed.

Convert date/time fields to R classes with special control for time zones. When selected, variables with date or datetime formats are converted to R date/time objects. You must select one of the following options:

- **R POSIXct.** Variables with date or datetime formats are converted to R `POSIXct` objects.
- **R POSIXlt (list).** Variables with date or datetime formats are converted to R `POSIXlt` objects.

Note: The POSIX formats are advanced options. Use these options only if your R script specifies that datetime fields are treated in ways that require these formats. The POSIX formats do not apply to variables with time formats.

Python Syntax

Python Syntax. You can enter, or paste, custom Python scripting syntax for data analysis into this field. For more information about Python for Spark, see Python for Spark and Scripting with Python for Spark.

Extension Export node - Console Output tab

The **Console Output** tab contains any output that is received when the R script or Python for Spark script on the Syntax tab runs (for example, if using an R script, it shows output received from the R console when the R script in the **R Syntax** field on the **Syntax** tab is executed). This output might include R or

Python error messages or warnings that are produced when the R or Python script is executed. The output can be used, primarily, to debug the script. The **Console Output** tab also contains the script from the **R Syntax** or **Python Syntax** field.

Every time the Extension Export script is executed, the content of the **Console Output** tab is overwritten with the output received from the R console or Python for Spark. The output cannot be edited.

XML Export Node

The XML Export node enables you to output data in XML format, using UTF-8 encoding. You can optionally create an XML source node to read the exported data back into the stream.

XML export file. The full path and file name of the XML file to which you want to export the data.

Use XML schema. Select this check box if you want to use a schema or DTD to control the structure of the exported data. Doing so activates the **Map** button, described below.

If you do not use a schema or DTD, the following default structure is used for the exported data:

```
<records>
     <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
      :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

Spaces in a field name are replaced with underscores; for example, "My Field" becomes <My_Field>.

Map. If you have chosen to use an XML schema, this button opens a dialog where you can specify which part of the XML structure should be used to start each new record. See the topic "XML Mapping Records Options" on page 336 for more information.

Mapped fields. Indicates the number of fields that have been mapped.

Generate an import node for this data. Select this option to automatically generate an XML source node that will read the exported data file back into the stream. See the topic "XML Source Node" on page 40 for more information.

Writing XML Data

When an XML element is specified, the field value is placed inside the element tag:

```
<element>value</element>
```

When an attribute is mapped, the field value is placed as the value for the attribute:

```
<element attribute="value">
```

If a field is mapped to an element above the <records> element, the field is written only once, and will be a constant for all records. The value for this element will come from the first record.

If a null value is to be written, it is done by specifying empty content. For elements, this is:

```
<element></element>
```

For attributes, it is:

```
<element attribute="">
```

XML Mapping Records Options

The Records tab enables you to specify which part of the XML structure to use to start each new record. In order to map correctly onto a schema, you need to specify the record delimiter.

XML structure. A hierarchical tree showing the structure of the XML schema specified on the previous screen.

Records (XPath expression). To set the record delimiter, select an element in the XML structure and click the right-arrow button. Each time this element is encountered in the source data, a new record is created in the output file.

Note: If you select the root element in the XML structure, only a single record can be written, and all other records are skipped.

XML Mapping Fields Options

The Fields tab is used to map fields in the data set to elements or attributes in the XML structure when a schema file is used.

Field names that match an element or attribute name are automatically mapped as long as the element or attribute name is unique. Thus if there is both an element and an attribute with the name `field1`, there is no automatic mapping. If there is only one item in the structure named `field1`, a field with that name in the stream is automatically mapped.

Fields. The list of fields in the model. Select one or more fields as the source part of the mapping. You can use the buttons at the bottom of the list to select all fields, or all fields with a particular measurement level.

XML structure. Select an element in the XML structure as the map target. To create the mapping, click Map. The mapping is then displayed. The number of fields that have been mapped in this way is displayed below this list.

To remove a mapping, select the item in the XML structure list and click **Unmap**.

Display attributes. Displays or hides the attributes, if any, of the XML elements in the XML structure.

XML Mapping Preview

On the Preview tab, click **Update** to see a preview of the XML that will be written.

If the mapping is incorrect, return to the Records or Fields tab to correct the errors and click **Update** again to see the result.

JSON Export node

Use the JSON export node to output data in JSON format, using UTF-8 encoding. Optionally, you can also create a JSON source node to read the exported data back into the stream.

When SPSS Modeler writes data to a JSON export file, it performs the following translations.

Table 51. JSON data export translation

SPSS Modeler Data Storage	JSON Value
String	string
Integer	number(int)
Real	number(real)

Table 51. JSON data export translation (continued)

SPSS Modeler Data Storage	JSON Value
Date	string
Time	string
Timestamp	string
List	Not supported. Fields of list will be excluded.
Missing values	null

JSON export file. The full path and file name of the JSON file where data will be exported.

JSON string format. Specify the format of the JSON string. Select **Records** if you want the JSON export node to output a collection of name and value pairs. Or select **Values** if you only want to export values (no names).

JSON string format. Specify the format of the JSON string. Select Records the JSON export node will output a collection of name and value pairs. Or select Values if you only want to export values (no names).

Generate an import node for this data. Select this option to automatically generate a JSON source node that will read the exported data file back into the stream. For more information, see "JSON Source node" on page 59.

Common Export node tabs

The following options can be specified for all export nodes by clicking the corresponding tab:

- **Publish tab.** Used to publish results of a stream.
- **Annotations tab.** Used for all nodes, this tab offers options to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

Publishing streams

Publishing streams is done directly from IBM SPSS Modeler using any of the standard export nodes: Database, Flat File, Statistics Export, Extension Export, Data Collection Export, SAS Export, Excel, and XML Export nodes. The type of export node determines the format of the results to be written each time the published stream is executed using the IBM SPSS Modeler Solution Publisher Runtime or external application. For example, if you want to write your results to a database each time the published stream is run, use a Database export node.

To publish a stream

1. Open or build a stream in the normal manner and attach an export node at the end.
2. On the Publish tab in the export node, specify a root name for the published files (that is, the file name to which the `.pim`, `.par`, and `.xml` extensions will be appended).
3. Click **Publish** to publish the stream, or select **Publish the stream** to automatically publish the stream each time the node is executed.

Published name. Specify the root name for the published image and parameter files.

- The **image file** (`*.pim`) provides all of the information needed for the Runtime to execute the published stream exactly as it was at the time of export. If you are confident that you will not need to change any of the settings for the stream (such as the input data source or the output data file), you can deploy the image file only.
- The **parameter file** (`*.par`) contains configurable information about data sources, output files, and execution options. If you want to be able to control the input or output of the stream without republishing the stream, you will need the parameter file as well as the image file.

- The **metadata file** (`*.xml`) describes the inputs and outputs of the image and their data models. It is designed for use by applications which embed the runtime library and which need to know the structure of the input and output data.

  Note: This file is only produced if you select the **Publish metadata** option.

Publish parameters. If required, you can include stream parameters in the `*.par` file. You can change these stream parameter values when you execute the image either by editing the `*.par` file or through the runtime API.

This option enables the **Parameters** button. The Publish Parameters dialog box is displayed when you click the button.

Choose the parameters you want to include in the published image by selecting the relevant option in the **Publish** column.

On stream execution. Specifies whether the stream is automatically published when the node is executed.

- **Export data.** Executes the export node in the standard manner, without publishing the stream. (Basically, the node executes in IBM SPSS Modeler the same way it would if IBM SPSS Modeler Solution Publisher were not available.) If you select this option, the stream will not be published unless you do so explicitly by clicking **Publish** in the export node dialog box. Alternatively, you can publish the current stream using the Publish tool on the toolbar or by using a script.
- **Publish the stream.** Publishes the stream for deployment using IBM SPSS Modeler Solution Publisher. Select this option if you want to automatically publish the stream every time it is executed.

Note:

- If you plan to run the published stream with new or updated data, it is important to note that the order of fields in the input file must be the same as the order of fields in the source node input file specified in the published stream.
- When publishing to external applications, consider filtering extraneous fields or renaming fields to conform with input requirements. Both can be accomplished using a Filter node prior to the export node.

Chapter 8. IBM SPSS Statistics Nodes

IBM SPSS Statistics Nodes - Overview

To complement IBM SPSS Modeler and its data mining abilities, IBM SPSS Statistics provides you with the ability to carry out further statistical analysis and data management.

If you have a compatible, licensed copy of IBM SPSS Statistics installed, you can connect to it from IBM SPSS Modeler and carry out complex, multistep data manipulation and analysis not otherwise supported by IBM SPSS Modeler. For the advanced user there is also the option to further modify the analysis by using command syntax. See the Release Notes for information on version compatibility.

If available, the IBM SPSS Statistics nodes are shown on a dedicated part of the nodes palette.

Note: We recommend that you instantiate your data in a Type node before using the IBM SPSS Statistics Transform, Model, or Output nodes. This is also a requirement when using the AUTORECODE syntax command.

The IBM SPSS Statistics palette contains the following nodes:

The Statistics File node reads data from the *.sav* or *.zsav* file format used by IBM SPSS Statistics, as well as cache files saved in IBM SPSS Modeler, which also use the same format.

The Statistics Transform node runs a selection of IBM SPSS Statistics syntax commands against data sources in IBM SPSS Modeler. This node requires a licensed copy of IBM SPSS Statistics.

The Statistics Model node enables you to analyze and work with your data by running IBM SPSS Statistics procedures that produce PMML. This node requires a licensed copy of IBM SPSS Statistics.

The Statistics Output node allows you to call an IBM SPSS Statistics procedure to analyze your IBM SPSS Modeler data. A wide variety of IBM SPSS Statistics analytical procedures is available. This node requires a licensed copy of IBM SPSS Statistics.

The Statistics Export node outputs data in IBM SPSS Statistics *.sav* or *.zsav* format. The `.sav` or `.zsav` files can be read by IBM SPSS Statistics Base and other products. This is also the format used for cache files in IBM SPSS Modeler.

Note: If your copy of SPSS Statistics is licensed for a single user only and you run a stream with two or more branches, each of which contains an SPSS Statistics node, you may receive a licensing error. This occurs when the SPSS Statistics session for one branch has not terminated before the session for another branch attempts to start. If possible, redesign the stream so that multiple branches with SPSS Statistics nodes do not execute in parallel.

Statistics File Node

You can use the Statistics File node to read data directly from a saved IBM SPSS Statistics file (*.sav* or *.zsav*). This format is now used to replace the cache file from earlier versions of IBM SPSS Modeler. If you would like to import a saved cache file, you should use the IBM SPSS Statistics File node.

Import file. Specify the name of the file. You can enter a filename or click the ellipsis button (**...**) to select a file. The file path is shown once you have selected a file.

File is password encrypted. Select this box if you know the file is password protected; you are the prompted to enter the **Password**. If the file is password protected, and you do not enter the password, a warning message is displayed if you attempt to change to another tab, refresh the data, preview the node contents, or try to execute a stream containing the node.

Note: Password protected files can only be opened by IBM SPSS Modeler version 16 or greater.

Variable names. Select a method of handling variable names and labels upon import from an IBM SPSS Statistics *.sav* or *.zsav* file. Metadata that you choose to include here persists throughout your work in IBM SPSS Modeler and may be exported again for use in IBM SPSS Statistics.

- **Read names and labels.** Select to read both variable names and labels into IBM SPSS Modeler. By default, this option is selected and variable names are displayed in the Type node. Labels may be displayed in charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box. By default, the display of labels in output is disabled.
- **Read labels as names.** Select to read the descriptive variable labels from the IBM SPSS Statistics *.sav* or *.zsav* file rather than the short field names, and use these labels as variable names in IBM SPSS Modeler.

Values. Select a method of handling values and labels upon import from an IBM SPSS Statistics *.sav* or *.zsav* file. Metadata that you choose to include here persists throughout your work in IBM SPSS Modeler and may be exported again for use in IBM SPSS Statistics.

- **Read data and labels.** Select to read both actual values and value labels into IBM SPSS Modeler. By default, this option is selected and values themselves are displayed in the Type node. Value labels may be displayed in the Expression Builder, charts, model browsers, and other types of output, depending on the options specified in the stream properties dialog box.
- **Read labels as data.** Select if you want to use the value labels from the *.sav* or *.zsav* file rather than the numerical or symbolic codes used to represent the values. For example, selecting this option for data with a gender field whose values of 1 and 2 actually represent *male* and *female*, respectively, will convert the field to a string and import *male* and *female* as the actual values.

  It is important to consider missing values in your IBM SPSS Statistics data before selecting this option. For example, if a numeric field uses labels only for missing values (0 = *No Answer*, –99 = *Unknown*), then selecting the option above will import only the value labels *No Answer* and *Unknown* and will convert the field to a string. In such cases, you should import the values themselves and set missing values in a Type node.

Use field format information to determine storage. If this box is cleared, field values that are formatted in the *.sav* file as integers (i.e., fields specified as F*n*.0 in the Variable View in IBM SPSS Statistics) are imported using integer storage. All other field values except strings are imported as real numbers.

If this box is selected (default), all field values except strings are imported as real numbers, whether formatted in the *.sav* file as integers or not.

Multiple response sets. Any multiple response sets defined in the IBM SPSS Statistics file will automatically be preserved when the file is imported. You can view and edit multiple response sets from any node with a Filter tab. See the topic "Editing Multiple Response Sets" on page 135 for more information.

Statistics Transform Node

The Statistics Transform node enables you to complete data transformations using IBM SPSS Statistics command syntax. This makes it possible to complete a number of transformations not supported by IBM SPSS Modeler and allows automation of complex, multistep transformations, including the creation of a number of fields from a single node. It resembles the Statistics Output node, except that the data are returned to IBM SPSS Modeler for further analysis, whereas, in the Output node the data are returned as the requested output objects, such as graphs or tables.

You must have a compatible version of IBM SPSS Statistics installed and licensed on your computer to use this node. See " IBM SPSS Statistics Helper Applications" on page 311 for more information. See the online Release Notes for compatibility information.

If necessary, you can use the Filter tab to filter or rename fields so they conform to IBM SPSS Statistics naming standards. See "Renaming or Filtering Fields for IBM SPSS Statistics " on page 328 for more information.

Syntax Reference. For details about specific IBM SPSS Statistics procedures, see the *IBM SPSS Statistics Command Syntax Reference* guide, included with your copy of the IBM SPSS Statistics software. To view the guide from the Syntax tab, choose the **Syntax editor** option and click the **Launch Statistics Syntax Help** button.

Note: Not all IBM SPSS Statistics syntax is supported by this node. See the topic "Allowable Syntax" on page 341 for more information.

Statistics Transform Node - Syntax Tab

IBM SPSS Statistics dialog option

If you are unfamiliar with IBM SPSS Statistics syntax for a procedure, the simplest way to create syntax in IBM SPSS Modeler is to choose the **IBM SPSS Statistics dialog** option, select the dialog box for the procedure, complete the dialog box and click OK. Doing so places the syntax onto the Syntax tab of the IBM SPSS Statistics node you are using in IBM SPSS Modeler. You can then run the stream to obtain the output from the procedure.

IBM SPSS Statistics Syntax editor option

Check. After you have entered your syntax commands in the upper part of the dialog box, use this button to validate your entries. Any incorrect syntax is identified in the bottom part of the dialog box.

To ensure that the checking process does not take too long, when you validate the syntax, it is checked against a representative sample of your data to ensure that the entries are valid instead of checking against the entire dataset.

Allowable Syntax

If you have a lot of legacy syntax from IBM SPSS Statistics or are familiar with the data preparation features of IBM SPSS Statistics, you can use the Statistics Transform node to run many of your existing transformations. As a guideline, the node enables you to transform data in predictable ways--for example, by running looped commands or by changing, adding, sorting, filtering, or selecting data.

Examples of the commands that can be carried out are:

- Compute random numbers according to a binomial distribution:

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- Recode a variable into a new variable:

```
RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO
AgeRecoded
```

- Replace missing values:

```
RMV Age_1=SMEAN(Age)
```

The IBM SPSS Statistics syntax that is supported by the Statistics Transform node is listed below.

Command Name

ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
COUNT
CREATE
DATE
DEFINE-!ENDDEFINE
DELETE VARIABLES
DO IF
DO REPEAT
ELSE
ELSE IF
END CASE
END FILE
END IF
END INPUT PROGRAM
END LOOP
END REPEAT
EXECUTE
FILE HANDLE
FILE LABEL
FILE TYPE-END FILE TYPE
FILTER
FORMATS
IF
INCLUDE
INPUT PROGRAM-END INPUT PROGRAM
INSERT

Command Name

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

STRING

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Statistics Model Node

The Statistics Model node enables you to analyze and work with your data by running IBM SPSS Statistics procedures that produce PMML. The model nuggets you create can then be used in the usual way within IBM SPSS Modeler streams for scoring, and so on.

You must have a compatible version of IBM SPSS Statistics installed and licensed on your computer to use this node. See " IBM SPSS Statistics Helper Applications" on page 311 for more information. See the online Release Notes for compatibility information.

The IBM SPSS Statistics analytical procedures that are available depend on the type of license you have.

Statistics Model Node - Model Tab

Model name. You can generate the model name automatically based on the target or ID field (or model type in cases where no such field is specified) or specify a custom name.

Select a dialog. Click to display a list of available IBM SPSS Statistics procedures that you can select and run. The list shows only those procedures that produce PMML and for which you are licensed, and does not include user-written procedures.

1. Click on the required procedure; the relevant IBM SPSS Statistics dialog is displayed.
2. In the IBM SPSS Statistics dialog, enter the details for the procedure.
3. Click **OK** to return to the Statistics Model node; the IBM SPSS Statistics syntax is displayed in the Model tab.
4. To return to the IBM SPSS Statistics dialog at any time, for example to modify your query, click the IBM SPSS Statistics dialog display button to the right of the procedure selection button.

Statistics Model Node - Model Nugget Summary

When you run the Statistics Model node, it executes the associated IBM SPSS Statistics procedure and creates a model nugget that you can use in IBM SPSS Modeler streams for scoring.

The Summary tab of the model nugget displays information about the fields, build settings, and model estimation process. Results are presented in a tree view that can be expanded or collapsed by clicking specific items.

The **View Model** button displays the results in a modified form of the IBM SPSS Statistics Output Viewer. For more information about this viewer, see the IBM SPSS Statistics documentation.

The usual exporting and printing options are available from the File menu. See the topic “Viewing output” on page 272 for more information.

Statistics Output Node

The Statistics Output node enables you to call an IBM SPSS Statistics procedure to analyze your IBM SPSS Modeler data. You can view the results in a browser window or save results in the IBM SPSS Statistics output file format. A wide variety of IBM SPSS Statistics analytical procedures is accessible from IBM SPSS Modeler.

You must have a compatible version of IBM SPSS Statistics installed and licensed on your computer to use this node. See “ IBM SPSS Statistics Helper Applications” on page 311 for more information. See the online Release Notes for compatibility information.

If necessary, you can use the Filter tab to filter or rename fields so they conform to IBM SPSS Statistics naming standards. See “Renaming or Filtering Fields for IBM SPSS Statistics ” on page 328 for more information.

Syntax Reference. For details about specific IBM SPSS Statistics procedures, see the *IBM SPSS Statistics Command Syntax Reference* guide, included with your copy of the IBM SPSS Statistics software. To view the guide from the Syntax tab, choose the **Syntax editor** option and click the **Launch Statistics Syntax Help** button.

Statistics Output Node - Syntax Tab

Use this tab to create syntax for the SPSS Statistics procedure you want to use to analyze your data. Syntax is composed of two parts: a **statement** and associated **options**. The statement specifies the analysis or operation to be performed and the fields to be used. The options specify everything else, including which statistics to display, derived fields to save, and so on.

SPSS Statistics dialog option

If you are unfamiliar with IBM SPSS Statistics syntax for a procedure, the simplest way to create syntax in IBM SPSS Modeler is to choose the **IBM SPSS Statistics dialog** option, select the dialog box for the procedure, complete the dialog box and click OK. Doing so places the syntax onto the Syntax tab of the IBM SPSS Statistics node you are using in IBM SPSS Modeler. You can then run the stream to obtain the output from the procedure.

You can optionally generate a Statistics File source node for importing the resulting data. This is useful, for example, if a procedure writes fields such as scores to the active dataset in addition to displaying output.

Note:

- When generating output in languages other than English, it is advisable to specify the language in the syntax.
- The Output Style option is not supported in the Statistics Output node.

To create the syntax

1. Click the **Select a dialog** button.
2. Choose one of the options:
   - **Analyze** Lists the contents of the SPSS Statistics Analyze menu; choose the procedure you want to use.
   - **Other** If shown, lists dialogs created by the Custom Dialog Builder in SPSS Statistics, as well as any other SPSS Statistics dialogs that do not appear on the Analyze menu and for which you have a licence. If there are no applicable dialogs, this option is not shown.

Note: The Automatic Data Preparation dialogs are not shown.

If you have an SPSS Statistics custom dialog that creates new fields, these fields cannot be used in SPSS Modeler because the Statistics Output node is a terminal node.

Optionally check the **Generate an import node for the resulting data** box to create a Statistics File source node that can be used to import the resulting data into another stream. The node is placed on the screen canvas, with the data contained in the `.sav` file specified in the **File** field (default location is the SPSS Modeler installation directory).

Syntax editor option

To save syntax that has been created for a frequently used procedure:

1. Click the **File Options** button (the first one on the toolbar).
2. Choose **Save** or **Save As** from the menu.
3. Save the file as an `.sps` file.

To use previously created syntax files, replacing the current contents, if any, of the Syntax editor:

1. Click the File Options button (the first one on the toolbar).
2. Choose **Open** from the menu.
3. Select an `.sps` file to paste its contents into the Output node Syntax tab.

To insert previously saved syntax without replacing the current contents:

1. Click the File Options button (the first one on the toolbar).
2. Choose **Insert** from the menu
3. Select an `.sps` file to paste its contents into the Output node at the point specified by the cursor.

Optionally check the **Generate an import node for the resulting data** box to create a Statistics File source node that can be used to import the resulting data into another stream. The node is placed on the

screen canvas, with the data contained in the `.sav` file specified in the **File** field (default location is the SPSS Modeler installation directory).

When you click **Run**, the results are shown in the SPSS Statistics Output Viewer. For more information about the viewer, see the SPSS Statistics documentation.

Note: Syntax for the following items (and their corresponding SPSS Statistics dialog box options) is not supported. They have no impact on the output.

- `OUTPUT ACTIVATE`
- `OUTPUT CLOSE`
- `OUTPUT DISPLAY`
- `OUTPUT EXPORT`
- `OUTPUT MODIFY`
- `OUTPUT NAME`
- `OUTPUT NEW`
- `OUTPUT OPEN`
- `OUTPUT SAVE`

Statistics Output Node - Output Tab

The Output tab lets you specify the format and location of the output. You can choose to display the results on the screen or send them to one of the available file types.

Output name. Specifies the name of the output produced when the node is executed. **Auto** chooses a name based on the node that generates the output. Optionally, you can select **Custom** to specify a different name.

Output to screen (the default)**.** Creates an output object to view online. The output object will appear on the Outputs tab of the manager window when the output node is executed.

Output to file. Saves the output to a file when you run the node. If you choose this option, enter a filename in the **Filename** field (or navigate to a directory and specify a filename using the File Chooser button) and select a file type.

File type. Choose the type of file to which you want to send the output.

- **HTML document (*.html).** Writes the output in HTML format.
- **IBM SPSS Statistics Viewer File (*.spv).** Writes the output in a format that can be read by the IBM SPSS Statistics Output Viewer.
- **IBM SPSS Statistics Web Reports File (*.spw).** Writes the output in IBM SPSS Statistics Web Reports format, which can be published to an IBM SPSS Collaboration and Deployment Services repository and subsequently viewed in a Web browser. See the topic "Publish to Web" on page 273 for more information.

Note: If you select **Output to screen** the IBM SPSS Statistics OMS directive `VIEWER=NO` has no effect; also, the scripting APIs (*Basic* and *Python SpssClient* module) are not available in IBM SPSS Modeler.

Statistics Export Node

The Statistics Export node enables you to export data in IBM SPSS Statistics *.sav* format. IBM SPSS Statistics *.sav* files can be read by IBM SPSS Statistics Base and other modules. This is also the format used for IBM SPSS Modeler cache files.

Mapping IBM SPSS Modeler field names to IBM SPSS Statistics variable names can sometimes cause errors because IBM SPSS Statistics variable names are limited to 64 characters and cannot include certain characters, such as spaces, dollar signs ($), and dashes (–). There are two ways to adjust for these restrictions:

- You can rename fields conforming to IBM SPSS Statistics variable name requirements by clicking the Filter tab. See the topic “Renaming or Filtering Fields for IBM SPSS Statistics ” on page 328 for more information.
- Choose to export both field names and labels from IBM SPSS Modeler.

Note: IBM SPSS Modeler writes *.sav* files in Unicode UTF-8 format. IBM SPSS Statistics only supports files in Unicode UTF-8 format from release 16.0 onwards. To prevent the possibility of data corruption *.sav* files saved with Unicode encoding should not be used in releases of IBM SPSS Statistics prior to 16.0. For more information, see the IBM SPSS Statistics help.

Multiple response sets. Any multiple response sets defined in the stream will automatically be preserved when the file is exported. You can view and edit multiple response sets from any node with a Filter tab. See the topic “Editing Multiple Response Sets” on page 135 for more information.

Statistics Export Node - Export Tab

Export file Specifies the name of the file. Enter a filename or click the file chooser button to browse to the file's location.

File type Select if the file is to be saved in normal *.sav* or compressed*.zsav* format.

Encrypt file with password To protect the file with a password, select this box; you are the prompted to enter and confirm the **Password** in a separate dialog box.

Note: Password protected files can only be opened by SPSS Modeler version 16 or greater, or by SPSS Statistics version 21 or greater.

Export field names Specifies a method of handling variable names and labels upon export from SPSS Modeler to an SPSS Statistics *.sav* or *.zsav* file.

- **Names and variable labels** Select to export both SPSS Modeler field names and field labels. Names are exported as SPSS Statistics variable names, while labels are exported as SPSS Statistics variable labels.
- **Names as variable labels** Select to use the SPSS Modeler field names as variable labels in SPSS Statistics. SPSS Modeler allows characters in field names that are invalid in SPSS Statistics variable names. To prevent possibly creating invalid SPSS Statistics names, select **Names as variable labels** instead, or use the Filter tab to adjust field names.

Launch Application If SPSS Statistics is installed on your computer, you can select this option to invoke the application directly on the saved data file. Options for launching the application must be specified in the Helper Applications dialog box. See the topic “ IBM SPSS Statistics Helper Applications” on page 311 for more information. To simply create an SPSS Statistics *.sav* or *.zsav* file without opening an external program, deselect this option.

Note: When running SPSS Modeler and SPSS Statistics together in Server (distributed) mode, writing the data out and launching a SPSS Statistics session does not automatically open a SPSS Statistics client showing the data set read into the active data set. The workaround is to manually open the data file in the SPSS Statistics client once it is launched.

Generate an import node for this data Select this option to automatically generate a Statistics File source node that will read the exported data file. See the topic “Statistics File Node” on page 28 for more information.

Renaming or Filtering Fields for IBM SPSS Statistics

Before exporting or deploying data from IBM SPSS Modeler to external applications such as IBM SPSS Statistics, it may be necessary to rename or adjust field names. The Statistics Transform, Statistics Output, and Statistics Export dialog boxes contain a Filter tab to facilitate this process.

A basic description of Filter tab functionality is discussed elsewhere. See the topic “Setting filtering options” on page 134 for more information. This topic provides tips for reading data into IBM SPSS Statistics.

To adjust field names to conform to IBM SPSS Statistics naming conventions:

1. On the Filter tab, click the Filter Options Menu toolbar button (the first one on the toolbar).
2. Select Rename For IBM SPSS Statistics.
3. On the Rename For IBM SPSS Statistics dialog, you can choose to replace invalid characters in filenames with either a **Hash (#)** character or an **Underscore (_)**.

Rename multi response sets. Select this option if you want to adjust the names of multiple response sets, which can be imported into IBM SPSS Modeler using a Statistics File source node. They are used to record data that can have more than one value for each case, such as in survey responses.

Chapter 9. SuperNodes

Overview of SuperNodes

One of the reasons that the IBM SPSS Modeler visual programming interface is so easy to learn is that each node has a clearly defined function. However, for complex processing, a long sequence of nodes may be necessary. Eventually, this may clutter the stream canvas and make it difficult to follow stream diagrams. There are two ways to avoid the clutter of a long and complex stream:

- You can split a processing sequence into several streams that feed one into the other. The first stream, for example, creates a data file that the second uses as input. The second creates a file that the third uses as input, and so on. You can manage these multiple streams by saving them in a **project**. A project provides organization for multiple streams and their output. However, a project file contains only a reference to the objects it contains, and you will still have multiple stream files to manage.
- You can create a **SuperNode** as a more streamlined alternative when working with complex stream processes.

SuperNodes group multiple nodes into a single node by encapsulating sections of a data stream. This provides numerous benefits to the data miner:

- Streams are neater and more manageable.
- Nodes can be combined into a business-specific SuperNode.
- SuperNodes can be exported to libraries for reuse in multiple data mining projects.

Types of SuperNodes

SuperNodes are represented in the data stream by a star icon. The icon is shaded to represent the type of SuperNode and the direction in which the stream must flow to or from it.

There are three types of SuperNodes:

- Source SuperNodes
- Process SuperNodes
- Terminal SuperNodes

Source SuperNodes

Source SuperNodes contain a data source just like a normal source node and can be used anywhere that a normal source node can be used. The left side of a source SuperNode is shaded to indicate that it is "closed" on the left and that data must flow downstream *from* a SuperNode.

Source SuperNodes have only one connection point on the right, showing that data leaves the SuperNode and flows to the stream.

Process SuperNodes

Process SuperNodes contain only process nodes and are unshaded to show that data can flow both *in* and *out* of this type of SuperNode.

Process SuperNodes have connection points on both the left and right, showing that data enters the SuperNode and leaves to flow back to the stream. Although SuperNodes can contain additional stream fragments and even extra streams, both connection points must flow through a single path connecting the *From Stream* and *To Stream* points.

Note: Process SuperNodes are also sometimes referred to as *Manipulation SuperNodes*.

Terminal SuperNodes

Terminal SuperNodes contain one or more terminal nodes (plot, table, and so on) and can be used in the same manner as a terminal node. A terminal SuperNode is shaded on the right side to indicate that it is "closed" on the right and that data can flow only *into* a terminal SuperNode.

Terminal SuperNodes have only one connection point on the left, showing that data enters the SuperNode from the stream and terminates inside the SuperNode.

Terminal SuperNodes can also contain scripts that are used to specify the order of execution for all terminal nodes inside the SuperNode. See the topic “SuperNodes and scripting” on page 355 for more information.

Creating SuperNodes

Creating a SuperNode "shrinks" the data stream by encapsulating several nodes into one node. Once you have created or loaded a stream on the canvas, there are several ways to create a SuperNode.

Multiple Selection

The simplest way to create a SuperNode is by selecting all of the nodes that you want to encapsulate:

1. Use the mouse to select multiple nodes on the stream canvas. You can also use Shift-click to select a stream or section of a stream.

   Note: Nodes that you select must be from a continuous or forked stream. You cannot select nodes that are not adjacent or connected in some way.
2. Then, using one of the following three methods, encapsulate the selected nodes:
   - Click the SuperNode icon (shaped like a star) on the toolbar.
   - Right-click the SuperNode, and from the context menu choose:

     Create SuperNode > **From Selection**
   - From the SuperNode menu, choose:

     Create SuperNode > **From Selection**

All three of these options encapsulate the nodes into a SuperNode shaded to reflect its type--source, process, or terminal--based on its contents.

Single Selection

You can also create a SuperNode by selecting a single node and using menu options to determine the start and end of the SuperNode or encapsulating everything downstream of the selected node.

1. Click the node that determines the start of encapsulation.
2. From the SuperNode menu, choose:

   Create SuperNode > **From Here**

SuperNodes can also be created more interactively by selecting the start and end of the stream section to encapsulate nodes:

1. Click on the first or last node that you want to include in the SuperNode.
2. From the SuperNode menu, choose:

   Create SuperNode > **Select...**
3. Alternatively, you can use the context menu options by right-clicking the desired node.
4. The cursor becomes a SuperNode icon, indicating that you must select another point in the stream. Move either upstream or downstream to the "other end" of the SuperNode fragment and click on a node. This action will replace all nodes in between with the SuperNode star icon.

Note: Nodes that you select must be from a continuous or forked stream. You cannot select nodes that are not adjacent or connected in some way.

Nesting SuperNodes

SuperNodes can be nested within other SuperNodes. The same rules for each type of SuperNode (source, process, and terminal) apply to nested SuperNodes. For example, a process SuperNode with nesting must have a continuous data flow through all nested SuperNodes in order for it to remain a process SuperNode. If one of the nested SuperNodes is terminal, then data would no longer flow through the hierarchy.

Terminal and source SuperNodes can contain other types of nested SuperNodes, but the same basic rules for creating SuperNodes apply.

Locking SuperNodes

Once you have created a SuperNode, you can lock it with a password to prevent it from being amended. For example, you might do this if you are creating streams, or parts of streams, as fixed-value templates for use by others in your organization who have less experience with setting up IBM SPSS Modeler enquiries.

When a SuperNode is locked users can still enter values on the Parameters tab for any parameters that have been defined, and a locked SuperNode can be executed without entering the password.

Note: Locking and unlocking cannot be performed using scripts.

Locking and unlocking a SuperNode

Warning: Lost passwords cannot be recovered.

You can lock or unlock a SuperNode from any of the three tabs.

1. Click **Lock Node**.
2. Enter and confirm the password.
3. Click **OK**.

A password protected SuperNode is identified on the stream canvas by a small padlock symbol to the top-left of the SuperNode icon.

Unlocking a SuperNode

1. To permanently remove the password protection, click **Unlock Node**. You are prompted for the password.
2. Enter the password and click **OK**. The SuperNode is no longer password protected and the padlock symbol no longer shows next to the icon in the stream.

For a stream saved in a version of SPSS Modeler between 16 and 17.0 that contains a locked SuperNode, when opening the stream in a different environment such as in IBM SPSS Collaboration and Deployment Services or on a Mac when the SPSS Modeler-installed JRE is different, it must first be opened, unlocked, and resaved using version 17.1 or later on the old environment where it was last saved.

Sometimes an incorrect password error will be displayed when unlocking a SuperNode in stream older than version 18. To work around this, reopen and unlock the node using the exact IBM SPSS Modeler version (or a more current version) on the same platform with the same system local settings as when it was last saved. Then open it in version 18 or later. lock the node, and save the stream again.

Editing a locked SuperNode

If you attempt to either define parameters or zoom in to display a locked SuperNode you are prompted to enter the password.

Enter the password and click **OK**.

You are now able to edit the parameter definitions and zoom in and out as often as you require until you close the stream that the SuperNode is in.

Note that this does not remove the password protection, it only allows you access to work with the SuperNode. See the topic "Locking and unlocking a SuperNode" on page 351 for more information.

Editing SuperNodes

Once you have created a SuperNode, you can examine it more closely by zooming in to it; if the SuperNode is locked, you will be prompted to enter the password. See the topic "Editing a locked SuperNode" on page 351 for more information.

To view the contents of a SuperNode, you can use the zoom-in icon from the IBM SPSS Modeler toolbar, or the following method:

1. Right-click a SuperNode.
2. From the context menu, choose **Zoom In**.

The contents of the selected SuperNode will be displayed in a slightly different IBM SPSS Modeler environment, with connectors showing the flow of data through the stream or stream fragment. At this level on the stream canvas, there are several tasks that you can perform:

- Modify the SuperNode type—source, process, or terminal.
- Create parameters or edit the values of a parameter. Parameters are used in scripting and CLEM expressions.
- Specify caching options for the SuperNode and its subnodes.
- Create or modify a SuperNode script (terminal SuperNodes only).

Modifying SuperNode types

In some circumstances, it is useful to alter the type of a SuperNode. This option is available only when you are zoomed in to a SuperNode, and it applies only to the SuperNode at that level. The three types of SuperNodes are explained in the following table.

Table 52. Types of SuperNode

Type of SuperNode	Description
Source SuperNode	One connection going out
Process SuperNode	Two connections: one coming in and one going out
Terminal SuperNode	One connection coming in

To change the type of a SuperNode

1. Be sure that you are zoomed in to the SuperNode.
2. From the SuperNode menu, choose **SuperNode Type**, and then choose the type.

Annotating and renaming SuperNodes

You can rename a SuperNode as it appears in the stream, and write annotations used in a project or report. To access these properties:

- Right-click a SuperNode (zoomed out) and choose **Rename and Annotate**.
- Alternatively, from the SuperNode menu, choose **Rename and Annotate**. This option is available in both zoomed-in and zoomed-out modes.

In both cases, a dialog box opens with the Annotations tab selected. Use the options here to customize the name displayed on the stream canvas and provide documentation regarding SuperNode operations.

Using Comments with SuperNodes

If you create a SuperNode from a commented node or nugget, you must include the comment in the selection to create the SuperNode if you want the comment to appear in the SuperNode. If you omit the comment from the selection, the comment will remain on the stream when the SuperNode is created.

When you expand a SuperNode that included comments, the comments are reinstated to where they were before the SuperNode was created.

When you expand a SuperNode that included commented objects, but the comments were not included in the SuperNode, the objects are reinstated to where they were, but the comments are not reattached.

SuperNode parameters

In IBM SPSS Modeler, you have the ability to set user-defined variables, such as `Minvalue`, whose values can be specified when used in scripting or CLEM expressions. These variables are called **parameters**. You can set parameters for streams, sessions, and SuperNodes. Any parameters set for a SuperNode are available when building CLEM expressions in that SuperNode or any nested nodes. Parameters set for nested SuperNodes are not available to their parent SuperNode.

There are two steps to creating and setting parameters for SuperNodes:

1. Define parameters for the SuperNode.
2. Then, specify the value for each parameter of the SuperNode.

These parameters can then be used in CLEM expressions for any encapsulated nodes.

Defining SuperNode Parameters

Parameters for a SuperNode can be defined in both zoomed-out and zoomed-in modes. The parameters defined apply to all encapsulated nodes. To define the parameters of a SuperNode, you first need to access the Parameters tab of the SuperNode dialog box. Use one of the following methods to open the dialog box:

- Double-click a SuperNode in the stream.
- From the SuperNode menu, choose **Set Parameters**.
- Alternatively, when zoomed in to a SuperNode, choose **Set Parameters** from the context menu.

Once you have opened the dialog box, the Parameters tab is visible with any previously defined parameters.

To Define a New Parameter

Click the **Define Parameters** button to open the dialog box.

Name. Parameter names are listed here. You can create a new parameter by entering a name in this field. For example, to create a parameter for the minimum temperature, you could type `minvalue`. Do not include the $P- prefix that denotes a parameter in CLEM expressions. This name is also used for display in the CLEM Expression Builder.

Long name. Lists the descriptive name for each parameter created.

Storage. Select a storage type from the list. Storage indicates how the data values are stored in the parameter. For example, when working with values containing leading zeros that you want to preserve (such as 008), you should select **String** as the storage type. Otherwise, the zeros will be stripped from the value. Available storage types are string, integer, real, time, date, and timestamp. For date parameters, note that values must be specified using ISO standard notation as shown in the next paragraph.

Value. Lists the current value for each parameter. Adjust the parameter as required. Note that for date parameters, values must be specified in ISO standard notation (that is, `YYYY-MM-DD`). Dates specified in other formats are not accepted.

Type (optional). If you plan to deploy the stream to an external application, select a measurement level from the list. Otherwise, it is advisable to leave the *Type* column as is. If you want to specify value constraints for the parameter, such as upper and lower bounds for a numeric range, select **Specify** from the list.

Note that long name, storage, and type options can be set for parameters through the user interface only. These options cannot be set using scripts.

Click the arrows at the right to move the selected parameter further up or down the list of available parameters. Use the delete button (marked with an *X*) to remove the selected parameter.

Setting Values for SuperNode Parameters

Once you have defined parameters for a SuperNode, you can specify values using the parameters in a CLEM expression or script.

To Specify the Parameters of a SuperNode

1. Double-click the SuperNode icon to open the SuperNode dialog box.
2. Alternatively, from the SuperNode menu, choose **Set Parameters**.
3. Click the **Parameters** tab. *Note*: The fields in this dialog box are the fields defined by clicking the **Define Parameters** button on this tab.
4. Enter a value in the text box for each parameter that you have created. For example, you can set the value *minvalue* to a particular threshold of interest. This parameter can then be used in numerous operations, such as selecting records above or below this threshold for further exploration.

Using SuperNode Parameters to Access Node Properties

SuperNode parameters can also be used to define node properties (also known as **slot parameters**) for encapsulated nodes. For example, suppose you want to specify that a SuperNode train an encapsulated Neural Net node for a certain length of time using a random sample of the data available. Using parameters, you can specify values for the length of time and percentage sample.

Suppose your example SuperNode contains a Sample node called *Sample* and a Neural Net node called *Train*. You can use the node dialog boxes to specify the Sample node's **Sample** setting as **Random %** and the Neural Net node's **Stop on** setting to **Time**. Once these options are specified, you can access the node properties with parameters and specify specific values for the SuperNode. In the SuperNode dialog box, click **Define Parameters** and create the parameters shown in the following table.

Table 53. Parameters to create

Parameter	Value	Long name
Train.time	5	Time to train (minutes)
Sample.random	10	Percentage random sample

Note: The parameter names, such as *Sample.random*, use correct syntax for referring to node properties, where *Sample* represents the name of the node and *random* is a node property.

Once you have defined these parameters, you can easily modify values for the two Sample and Neural Net node properties without reopening each dialog box. Instead, simply select **Set Parameters** from the SuperNode menu to access the Parameters tab of the SuperNode dialog box, where you can specify new values for **Random %** and **Time**. This is particularly useful when exploring the data during numerous iterations of model building.

SuperNodes and caching

From within a SuperNode, all nodes except terminal nodes can be cached. Caching is controlled by right-clicking a node and choosing one of several options from the Cache context menu. This menu option is available both from outside a SuperNode and for the nodes encapsulated within a SuperNode.

There are several guidelines for SuperNode caches:

- If any of the nodes encapsulated in a SuperNode have caching enabled, the SuperNode will also.
- Disabling the cache on a SuperNode disables the cache for *all* encapsulated nodes.
- Enabling caching on a SuperNode actually enables the cache on the last cacheable subnode. In other words, if the last subnode is a Select node, the cache will be enabled for that Select node. If the last subnode is a terminal node (which does not allow caching), the next node upstream that supports caching will be enabled.
- Once you have set caches for the subnodes of a SuperNode, any activities upstream from the cached node, such as adding or editing nodes, will flush the caches.

SuperNodes and scripting

You can use the SPSS Modeler scripting language to write simple programs that manipulate and execute the contents of a terminal SuperNode. For instance, you might want to specify the order of execution for a complex stream. As an example, if a SuperNode contains a Set Globals node that needs to be executed before a Plot node, you can create a script that executes the Set Globals node first. Values calculated by this node, such as the average or standard deviation, can then be used when the Plot node is executed.

The Script tab of the SuperNode dialog box is available only for terminal SuperNodes.

To open the Scripting dialog box for a Terminal SuperNode:

- Right-click the SuperNode canvas and choose **SuperNode Script**.
- Alternatively, in both zoomed-in and zoomed-out modes, you can choose **SuperNode Script** from the SuperNode menu.

Note: SuperNode scripts are executed only with the stream and SuperNode when you have selected **Run this script** in the dialog box.

Specific options for scripting and its use within SPSS Modeler are discussed in the *Scripting and Automation Guide*, which is available as a PDF file as part of your product download.

Saving and loading SuperNodes

One of the advantages of SuperNodes is that they can be saved and reused in other streams. When saving and loading SuperNodes, note that they use an `.slb` extension.

To save a SuperNode

1. Zoom in on the SuperNode.
2. From the SuperNode menu, choose **Save SuperNode**.
3. Specify a filename and directory in the dialog box.
4. Select whether to add the saved SuperNode to the current project.
5. Click **Save**.

To load a SuperNode

1. From the Insert menu in the IBM SPSS Modeler window, choose **SuperNode**.
2. Select a SuperNode file (`.slb`) from the current directory or browse to a different one.
3. Click **Load**.

Note: Imported SuperNodes have the default values for all of their parameters. To change the parameters, double-click a SuperNode on the stream canvas.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Terms and conditions for product documentation

Permissions for the use of these publications are granted subject to the following terms and conditions.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Personal use

You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Glossary

C

Covariance

An unstandardized measure of association between two variables, equal to the cross-product deviation divided by N-1.

K

Kurtosis

A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution.

M

Maximum

The largest value of a numeric variable.

Mean

A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Median

The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Minimum

The smallest value of a numeric variable.

Mode

The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode.

R

Range

The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

S

Skewness

A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

standard deviation

A measure of dispersion around the mean, equal to the square root of the variance. The standard deviation is measured in the same units as the original variable.

Standard Deviation

A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

Standard Error

A measure of how much the value of a test statistic varies from sample to sample. It is the standard deviation of the sampling distribution for a statistic. For example, the standard error of the mean is the standard deviation of the sample means.

Standard Error of Kurtosis

The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Standard Error of Mean

A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Standard Error of Skewness

The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Sum

The sum or total of the values, across all cases with nonmissing values.

U

Unique

Evaluates all effects simultaneously, adjusting each effect for all other effects of any type.

V

Valid

Valid cases having neither the system-missing value, nor a value defined as user-missing.

Variance

A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

Index

Special Characters

-auto- settings 255
.dbf files 59
.par file 337
.pim file 337
.sav files 28, 340
.sd2 (SAS) files 39
.shp files 59
.slb files 355
.ssd (SAS) files 39
.tpt (SAS) files 39
.zsav files 28, 340

Numerics

1-in-n sampling 66
2-D dot plot 176
3-D area chart
    description 176
3-D bar chart 176
3-D density 176
3-D graphs 170
3-D histogram 176
3-D pie chart 176
3-D scatterplot 176

A

adding
    records 70
adjusted propensity scores
    balancing data 69
ADO databases
    importing 30
Aggregate node
    approximation for median 72
    approximation for quartile 72
    optimization settings 72
    overview 70
    parallel processing 71
    performance 71
    setting options 71
aggregating records 157
aggregating time series data 163
analysis browser
    interpreting 281
Analysis node
    analysis tab 280
    output tab 275
Analytic Server export 329
Analytic Server source 9
animation in graphs 168
Anonymize node
    creating anonymized values 147
    overview 145
    setting options 146
anonymizing field names 135
ANOVA
    Means node 294
anti-join 75
Append node
    field matching 81
    overview 81
    setting options 81
    tagging fields 80
application examples 3
area chart
    3-D 176
ARIMA models
    transfer functions 92
ascending order 74
assigning data types 107
association plots 219
attributes
    in maps 197
audit
    Data Audit node 282
    initial data audit 282
Auto Data Prep node 109
auto-typing 122, 127
automated data preparation
    action details 119
    action summary 117
    derive node generation 121
    exclude fields 112
    exclusion of unused fields 111
    field analysis 116
    field details 118
    field processing summary 115
    field settings 111
    fields 110
    fields table 117
    links between views 115
    model view 115
    name fields 114
    normalize continuous target 121
    objectives 109
    predictive power 117
    prepare dates and times 111
    reset views 115
    unused field exclusion 111
automatic data preparation
    construction 113
    feature selection 113
    input preparation 112
    normalize continuous target 112
    preparing inputs 112
    preparing targets 112
    selection of features 113
    target preparation 112
automatic date recognition 24, 27
automatic recode 143, 144

B

balance factors 69
Balance node
  generating from graphs 252
  overview 69
  setting options 69
bands in graphs 246
bar chart
  3-D 176
  example 184, 185
  of counts 176
  on a map 176
baseline
  evaluation chart options 230
best line
  evaluation chart options 230
biased data 69
binned scatterplot
  hex bins 176
Binning node
  equal counts 149
  equal sums 149
  fixed-width bins 148
  mean/standard deviation bins 151
  optimal 151
  overview 147
  previewing bins 152
  ranks 150
  setting options 148
BITMAP indexes
  database tables 318
blank handling
  Binning node 148
  filling values 142
blank rows
  Excel files 39
blank values
  in Matrix tables 277
blanks
  in Matrix tables 277
boxplot
  example 188
bubble plot 176
bulk loading 319, 320
business rule
  evaluation chart options 231

C

cache
  SuperNodes 354
cache file node 28, 340
case data
  Data Collection source node 29, 30
categorical data 126
cell ranges
  Excel files 39
Chart Options 306
chart output 304
charts
  saving output 275
checking types 131
chi-square
chi-square *(continued)*
  Matrix node 279
choropleth
  example 192
choropleth map 176
clear values 60
CLEM expressions 63
cluster 254, 262
clustered samples 65–67
Codes variables
  Data Collection source node 30
coercing values 131
Cognos, see IBM Cognos 36
coincidence matrix
  Analysis node 280
collection measurement level 130, 138
Collection node
  appearance tab 218
  options tab 217, 218
  using the graph 219
collection type 130
collection values for derive formula 138
collision modifiers 262
color graph overlay 168
color map
  example 192
column order
  table browser 275, 277
column-wise binding 319
combining data
  from multiple files 74
comma-delimited files
  exporting 274, 333
  saving 275
comment characters
  in variable files 24
comments
  using with SuperNodes 352
commit size 319
composite records
  custom settings 85
compute durations
  automated data preparation 111
concatenating records 81
conditions
  ranked 78
  specifying a series 141
  specifying for a merge 78
confidence intervals
  Means node 295
contiguous data sampling 66
contiguous keys 72
continuous data 126, 129
control characters 9
convert measurement levels 126
converting sets to flags 157
coordinate systems
  transforming 261
coordinates map 176
copying type attributes 132
copying visualizations 263
correlations
  absolute value 291
  descriptive labels 291

correlations *(continued)*
Means node 295
probability 291
significance 291
statistics output 292
costs
evaluation charts 230
count value for aggregation 71
counts
Binning node 149
statistics output 292
CPLEX Optimization node
overview 104
setting options 105
CREATE INDEX command 318
creating
new fields 136, 137
CRISP-DM
data understanding 5
CRISP-DM process model
data preparation 107
cross-tabulation
Matrix node 277, 278
CSV data
importing 30
currency display format 133
cut points
Binning node 147
cyclical time elements
automated data preparation 111

D

DAT files
exporting 274, 333
saving 275
data
aggregating 70
anonymizing 145
audit 282
exploring 282
preparation 63
storage 142, 143
storage type 127
understanding 63
unsupported control characters 9
data audit browser
Edit menu 284
File menu 284
generating graphs 288
generating nodes 288
Data Audit node
output tab 275
settings tab 283
Data Collection export node 328
Data Collection source node
database connection settings 32
label types 32
language 32
log files 30
metadata files 30
multiple response sets 32
Data Collection survey data
importing 29, 30
data quality
Data Audit browser 286
data sources
database connections 15
data types
instantiation 126
database
bulk loading 319, 320
database connections
defining 15
preset values 18
Database export node
data source 314
export tab 314
indexing tables 318
mapping source data fields to database columns 315
merge options 315
schema 316
table name 314
Database source node
potential issues 17
query editor 21
selecting tables and views 20
SQL queries 15
date recognition 24, 27
date/time 122
dates
setting formats 133
decile bins 149
decimal places
display formats 133
decimal symbol
Flat File export node 326
number display formats 133
decreasing data 64, 65
degrees of freedom
Matrix node 279
Means node 295
deleting
map files 195
output objects 272
visualization stylesheets 195
visualization templates 195
delimited text data 24
delimiters 24, 319
density
3-D 176
density definition in Space-Time-Boxes 99
depth of list 8
Derive node
collection values 138
conditional 141
converting field storage 142
count 141
deriving a geospatial field 139
deriving a list field 139
flag 140
formula 138
formula value 138
generating from a Binning node 152
generating from automated data preparation 121
generating from bins 147
generating from graphs 252
generating from Web graph links 223

Derive node *(continued)*
geospatial values 138
multiple derive 137
nominal 141
overview 136
recoding values 142
setting options 137
state 141
descending order 74
directed layout for web graphs 221
direction of fields 131
discarding
fields 133
disguising data for use in a model 145
display formats
currency 133
decimal places 133
grouping symbol 133
numbers 133
scientific 133
Distinct node
composite settings 84, 85
optimization settings 84
overview 82
sorting records 82
distribution 215
Distribution node
appearance tab 213
plot tab 213
using the graph 213
using the table 213
documentation 3
dodge 254, 262
dot plot
2-D 176
example 187
dummy coding 157
duplicate
fields 74, 134
records 82
duration computation
automated data preparation 111

E

E-Plot node
appearance tab 242
options tab 242
plot tab 242
using the graph 243
editing graphs
size of graphic elements 258
editing visualizations
adding 3-D effects 261
automatic settings 255
axes 259
categories 260
collapsing categories 260
colors and patterns 256
combining categories 260
dashing 256
excluding categories 260
legend position 263
margins 258
editing visualizations *(continued)*
number formats 258
padding 258
panels 261
point aspect ratio 257
point rotation 257
point shape 257
rules 255
scales 259
selection 255
sorting categories 260
text 256
transforming coordinate systems 261
transparency 256
transpose 261
employee_data.sav data file 341
encapsulating nodes 350
Ensemble node
combining scores 154
output fields 154
EOL characters 24
equal counts
Binning node 149
ESRI files 196
ESRI server 59
evaluating models 280
Evaluation node
appearance tab 232, 237
business rule 231
hit condition 231
options tab 231
plot tab 230
reading results 232
score expression 231
using the graph 233
events
creating 211
examples
Applications Guide 3
overview 4
Excel
launching from IBM SPSS Modeler 333
Excel export node 333
Excel files
exporting 333
Excel Import node
generating from output 333
Excel source node 39
execution
specifying the order of 355
expected values
Matrix node 278
exploring data
Data Audit node 282
exploring graphs
graph bands 246
magic wand 251
marking elements 251
regions 249
export decimal places 133
export nodes
Analytic Server export 329
exporting
data from IBM Cognos TM1 332

exporting *(continued)*
map files 195
output 274
SuperNodes 355
visualization stylesheets 195
visualization templates 195
exporting data
DAT files 333
flat file format 326
geospatial 326
IBM Cognos Export node 36, 329, 330
IBM Cognos TM1 Export node 331
SAS format 333
text 333
to a database 314
to Excel 333
to IBM SPSS Statistics 327, 346
to JSON 336
XML format 335
Expression Builder 63
extension
derived field 137
Extension Export node
console output tab 334
Extension Import node
console output tab 58
Extension Output browser 308, 309
Extension Output node
console output tab 308
output tab 308
syntax tab 307
Extension Transform node
console output tab 97

F

F statistic
Means node 295
false values 130
features
in maps 197
field attributes 132
field derivation formula 138
field names
anonymizing 135
data export 314, 326, 327, 333, 347
field operations nodes
generating from a data audit 288
Time Intervals node 163
Field Reorder node
automatic sorting 162
custom ordering 162
setting options 162
field storage
converting 142
field types
in visualizations 173
fields
anonymizing data 145
deriving multiple fields 137
field and value labels 127
reordering 161
selecting multiple 138
transposing 159
Filler node
overview 142
FILLFACTOR keyword
indexing database tables 318
Filter node
multiple response sets 135
overview 133
setting options 134
filtering fields
for IBM SPSS Statistics 328, 347
First quartile
time series aggregation 163
Fixed File node
automatic date recognition 27
overview 27
setting options 27
fixed-field text data 27
flag type 122, 130
Flat File export node
export tab 326
flat files 24
flow map 176
format files 39
formats
data 6
fractional ranks 150
free-field text data 24
frequencies
Binning node 149

G

gains charts 226, 232
generating flags 157, 159
generating nodes from graphs
Balance nodes 252
Derive nodes 252
Filter nodes 252
Reclassify nodes 252
Select nodes 252
geographic coordinate system 164
Geospatial
setting import options 59
geospatial data
deriving 139
exporting 326
importing 24
in variable files 26
lists in variable files 26
merging 78
ranked condition merge 78
restrictions 124
geospatial data in maps 234
geospatial data preparation
Reproject node 165
geospatial data reprojection 164
geospatial measurement level 130, 138
geospatial measurement levels 8, 122, 124
Geospatial source node
.dbf files 59
.shp files 59
Map service 59
geospatial type 130
geospatial values for derive formula 138

global values 297
graph nodes
animation 168
Collection 217
Distribution 213
E-Plot 242
Evaluation 226
Graphboard 171
Histogram 215
map visualization 234
Multiplot 208
overlays 168
panels 168
Plot 202
Time Plot 211
Web 219
graph output tab 309
graph types
graphboard 176
graphboard
graph types 176
Graphboard node
appearance tab 194
graphic elements
changing 262
collision modifiers 262
converting 262
graphs
3-D 170
annotations tab 170
axis labels 264
bands 246
collections 217
copying 266
default color scheme 265
deleting regions 249
distributions 213
e-plots 242
evaluation charts 226
exploring 245
exporting 266
footnote 264
from Graphboard 171
generating from a data audit 288
generating nodes 252
histograms 215
map visualization 234
multiplot 208
output tabs 169
plots 202
printing 266
regions 249
rotating a 3-D image 170
saving 266
saving edited layouts 265
saving layout changes 265
saving output 275
size of graphic elements 258
stylesheet 265
time series 211
title 264
webs 219
grouping symbol
number display formats 133
grouping values 213

H

handling missing values 107
hassubstring function 140
HDATA format
Data Collection source node 29
heat map
example 190
helper applications 311
hex binned scatterplot 176
hexadecimal control characters 9
high-low chart 176
high-low-close chart 176
histogram
3-D 176
example 186
Histogram node
appearance tab 216
plot tab 215
using the graph 216
History node
overview 161
hits
evaluation chart options 231
HTML
saving output 275
HTML output
Report node 296
view in browser 274

I

IBM Cognos BI source node
importing reports 35
IBM Cognos Export node 36, 329, 330
IBM Cognos source node
icons 34
importing data 34
IBM Cognos TM1 export node
exporting data 332
mapping export data 332
IBM Cognos TM1 Export node 331
IBM Cognos TM1 source node
importing data 37
IBM SPSS Collaboration and Deployment Services Repository
using as location for visualization templates, stylesheets, and maps 195
IBM SPSS Modeler
documentation 3
IBM SPSS Modeler Server 1
IBM SPSS Modeler Solution Publisher 337
IBM SPSS Statistics
launching from IBM SPSS Modeler 311, 327, 344, 347
license location 311
valid field names 328, 347
IBM SPSS Statistics data files
importing survey data 30
IBM SPSS Statistics models
about 343
advanced nugget details 344

IBM SPSS Statistics models *(continued)*
model nugget 344
model options 344
IBM SPSS Statistics nodes 339
IBM SPSS Statistics Output node
Output tab 346
icons, IBM Cognos 34
if-then-else statements 141
imbalanced data 69
importance
comparing means 294
Means node 295
importing
data from IBM Cognos 34
data from IBM Cognos TM1 37
map files 195
reports from IBM Cognos BI 35
SuperNodes 355
visualization stylesheets 195
visualization templates 195
In2data databases
importing 30
incomplete records 76
indexing database tables 318
inner join 75
instantiation
source node 61
integer ranges 129
intervals
time series data 163
interventions
creating 211

J

jitter 254, 262
jittering 206, 242
joining datasets 81
joins
partial outer 77
JSON export node 336
JSON files
exporting 336
JSON source node 59

K

KDE Modeling node 309
KDE node
inputs 309
key fields 71, 157
key method 74
key value for aggregation 71

L

label fields
labeling records in output 131
label types
Data Collection source node 32
labels
exporting 327, 333, 347
importing 28, 39, 340
labels *(continued)*
specifying 127, 129, 130
lagged data 161
language
Data Collection source node 32
large databases
performing a data audit 282
layer options on maps 235
layers in geospatial maps 234
legend
position 263
lift charts 226, 232
line chart
on a map 176
line plots 202, 208, 242
LineString 124
links
Web node 221
list
depth 8
deriving 139
geospatial measurement levels 124
maximum length 127
list storage format 8
list storage type 26
lists
collection data type 130
geospatial data type 130
lists in variable files 26
locally weighted least squares regression
E-Plot node 242
Plot node 205
locking SuperNodes 351
LOESS smoother
E-Plot node 242
Plot node 205
log transformation
Time Series Modeler 92
lowess smoother See LOESS smoother
E-Plot node 242
Plot node 205

M

magic wand in graphs 251
main dataset 81
managers
outputs tab 272
map
color 176
overlay 176
with arrows 176
with bar charts 176
with line charts 176
with pie charts 176
with points 176
Map Conversion Utility 196, 197
map data reprojection 164
map files
deleting 195
exporting 195
importing 195
location 195
renaming 195

map files *(continued)*
selecting in Graphboard Template Chooser 175
map service
Geospatial source node 59
map shapefiles
concepts 197
editing pre-installed SMZ maps 196
types 197
using with Graphboard Template Chooser 196
map visualization
example 192
Map visualization node
change layer options 235
plot tab 234
map visualizations
creating 183
mapping
data for export to IBM Cognos TM1 332
mapping fields 315
maps
converting ESRI shapefiles 196
deleting features 201
deleting individual elements 201
distributing 202
feature labels 199
merging features 200
moving features 200
projection 201
smoothing 198
thinning 198
market research data
Data Collection source node 29, 32
importing 30, 33
marking elements 249, 251
matrix browser
Generate menu 279
Matrix node
appearance tab 278
column percentages 278
cross-tabulation 278
highlighting 278
output browser 279
output tab 275
row percentages 278
settings tab 277
sorting rows and columns 278
matrix output
saving as text 275
Max function
time series aggregation 163
maximum
Set Globals node 297
statistics output 292
maximum value for aggregation 71
MDD documents
importing 30
mean
Binning node 151
Set Globals node 297
statistics output 292
Mean function
time series aggregation 163
mean value for aggregation 71
mean value for records 70
mean/standard deviation
used to bin fields 151
means
comparing 293, 294
Means node
importance 294
independent groups 294
output browser 294, 295
output tab 275
paired fields 294
measurement level
changing in visualizations 171
collection 8, 130, 138
defined 122
geospatial 8, 124, 130, 138
in visualizations 173
restrictions in geospatial data 124
median
statistics output 292
median approximation 72
median value for aggregation 71, 72
member (SAS import)
setting 39
Merge node
filtering fields 79
optimization settings 80
overview 74
setting options 76, 78
tagging fields 80
merge options, database export 315
metadata
Data Collection source node 29, 30
Microsoft Excel source node 39
Min function
time series aggregation 163
minimum
Set Globals node 297
statistics output 292
minimum value for aggregation 71
missing values
in Aggregate nodes 70
in Matrix tables 277
mode
statistics output 292
Mode function
time series aggregation 163
model evaluation 226
model options
Statistics Model node 344
model view
in automated data preparation 115
modeling roles
specifying for fields 131
models
anonymizing data for 145
modifying data values 136
MultiLineString 124
multiple category sets 135
multiple derive 137
multiple dichotomy sets 135
multiple fields
selecting 138
multiple inputs 74
multiple response sets

multiple response sets *(continued)*
Data Collection source node 29, 30, 32, 33
defining 135
deleting 135
IBM SPSS Statistics source node 28, 340
in visualizations 173
multiple category sets 135
multiple dichotomy sets 135
Multiplot node
appearance tab 210
plot tab 208
using the graph 210
MultiPoint 124
MultiPolygon 124

N

natural log transformation
Time Series Modeler 92
natural order
altering 161
navigation 303
network layout for web graphs 221
node properties 354
nominal data 129
nonrandom samples 65, 66
normalize continuous target 112, 121
normalize values
graph nodes 208, 211
null values
in Matrix tables 277
nulls
in Matrix tables 277
number display formats 133

O

ODBC
bulk loading via 319, 320
connection for IBM Cognos Export node 330
Database source node 14
ODBC export node. See Database export node 314
one-way ANOVA
Means node 294
opening
output objects 272
optimal binning 151
options
IBM SPSS Statistics 311
Oracle 14
order merging 74
order of execution
specifying 355
order of input data 80
ordering data 74, 162
ordinal data 129
outer join 75
output
exporting 274
generating new nodes from 272
HTML 274
printing 272
saving 272
output elements 304
output files
saving 275
output formats 275
output manager 272
output nodes
output tab 275
publish to web 273
overlap map 176
overlays for graphs 168
overwriting database tables 314

P

p value
importance 294
palettes
displaying 255
hiding 255
moving 255
panel graph overlay 168
paneling 168
parallel coordinates graph 176
parallel processing
Aggregate node 71
merging 80
sorting 74
parameters
in IBM Cognos 36
node properties 354
setting for SuperNodes 353
SuperNodes 353, 354
partial joins 75, 77
partition fields 131, 155, 156
Partition node 155, 156
partitioning data
Analysis node 280
evaluation charts 230
path graph 176
Pearson chi-square
Matrix node 279
Pearson correlations
Means node 295
statistics output 292
percentile bins 149
performance
Aggregate node 71
Binning nodes 152
Derive nodes 152
merging 80
sampling data 65
sorting 74
performance evaluation statistic 280
periodicity
time series data 163
Time Series Modeler 92
pie chart
3-D 176
example 189
on a map 176
using counts 176
Plot node
appearance tab 207
options tab 206

Plot node *(continued)*
plot tab 205
using the graph 207
plotting associations 219
Point 124
point plots 202, 208, 242
polar coordinates 261
Polygon 124
potential issues
Database source node 17
preset values, database connection 18
primary key fields
Database export node 316
printing output 272
profit charts 226, 232
projected coordinate system 164
propensity scores
balancing data 69
properties
node 354
publish to web 273
publishing streams
IBM SPSS Modeler Solution Publisher 337
Python
bulk loading scripts 319, 320
python nodes 95, 238, 240, 241, 309, 311

Q

quality browser
generating filter nodes 287
generating select nodes 288
quality report
Data Audit browser 286
Quancept data
importing 30
Quantum data
importing 30
Quanvert databases
importing 30
quartile approximation 72
quartile bins 149
quartile value for aggregation 71, 72
queries
Database source node 14, 15
query banding
Teradata 19
query editor
Database source node 21
quintile bins 149
quotation marks
importing text files 24
quotes
for database export 314

R

random seed value
sampling records 156
range
statistics output 292
ranges
missing values 127
rank cases 150
ranked conditions
specifying for a merge 78
real ranges 129
recency
setting relative date 73
Reclassify node
generating from a distribution 213
overview 143, 147
recode 143, 144, 147
record
counts 71
labels 131
length 27
record operations nodes 63
records
merging 74
transposing 159
regions in graphs 249
renaming
fields for export 328, 347
map files 195
visualization stylesheets 195
visualization templates 195
renaming output objects 272
replacing field values 142
report browser 297
Report node
output tab 275
template tab 296
reports
saving output 275
Reproject node 165
Reprojection node 164
residuals
Matrix node 278
response charts 226, 232
restrictions in geospatial data 124
Restructure node
with Aggregate node 157
restructuring data 157
revenue
evaluation charts 230
RFM Aggregate node
overview 73
setting options 73
RFM Analysis node
binning values 154
overview 152
settings 153
ribbon chart 176
ROI
charts 226, 232
roles
specifying for fields 131
rotating 3-D graphs 170
row-wise binding 319

S

Sample node
clustered samples 65–67
nonrandom samples 65, 66
random samples 65, 66

Sample node *(continued)*
Sample Sizes for Strata 68
sampling frame 65
stratified samples 65–68
systematic samples 65, 66
weighted samples 67
sampling data 68
sampling frame 65
SAS
setting import options 39
SAS export node 333
SAS source node
.sd2 (SAS) files 39
.ssd (SAS) files 39
.tpt (SAS) files 39
transport files 39
saving
output 272
output objects 272, 275
scale factors 69
scatterplot
3-D 176
binned 176
hex binned 176
scatterplot matrix
example 191, 192
scatterplot matrix (SPLOM) 176
scatterplots 202, 208, 242
schema
Database export node 316
scientific display format 133
scoring
evaluation chart options 231
scripting
SuperNodes 355
searching
table browser 277
seed value
sampling and records 156
Select node
generating from graphs 252
generating from Web graph links 223
overview 64
selecting rows (cases) 64
selecting values 246, 249, 251
Set Globals node
settings tab 297
set random seed
sampling records 156
Set to Flag node 157
set type 122
sets
converting to flags 157
transforming 144, 145
shape graph overlay 168
shapefiles 196
significance
correlation strength 291
simulated data
Simulation Generate node 46
Simulation evaluation node
settings tab 301
Simulation Evaluation node
output settings 301
Simulation fitting node
distribution fitting 298
settings tab 300
Simulation Fitting node
distribution fitting 298
output settings 300
Simulation generate node
setting options 47
Simulation Generate node
overview 46
size graph overlay 168
smoother
E-Plot node 242
Plot node 205
SMOTE node 95
SMZ files
creating 196
deleting 195
editing pre-installed SMZ files 196
exporting 195
importing 195
overview 196
pre-installed 196
renaming 195
Sort node
optimization settings 74
overview 74
sorting
Distinct node 82
fields 161
presorted fields 74, 84
records 74
source nodes
Analytic Server source 9
Database source node 14
Excel source node 39
Fixed File node 27
Geospatial source node 59
IBM Cognos source node 33, 36
IBM Cognos TM1 source node 37
instantiating types 61
JSON source node 59
overview 5
SAS source node 39
Simulation Generate node 46, 47
Statistics File node 28, 340
The Weather Company source 38
TWC source 38
User Input node 42
Variable File node 24
XML source node 40
SourceFile variables
Data Collection source node 30
Space-Time-Boxes node
defining density 99
overview 97
SPLOM
example 191, 192
SQL queries
Database source node 14, 15, 21
square root transformation
Time Series Modeler 92
stack 254, 262
stacked bar chart

stacked bar chart *(continued)*
example 185
standard deviation
Binning node 151
Set Globals node 297
statistics output 292
standard deviation for aggregation 71
standard error of mean
statistics output 292
statistics
Data Audit node 282
editing in visualizations 262
Matrix node 277
statistics browser
Generate menu 292
generating filter nodes 293
interpreting 292
Statistics Export node
Export tab 327, 347
Statistics File node 28, 340
Statistics node
correlation labels 291
correlations 291
output tab 275
settings tab 291
statistics 291
Statistics Output node
Syntax tab 344
Statistics Transform node
allowable syntax 341
setting options 341
Syntax tab 341
storage
converting 142, 143
storage formats 6
storage types
list 26
stratified samples 65–68
stream parameters 21
streaming TCM node 99–104
Streaming Time Series models
aggregation and distribution options 88
ARIMA 90
build options 90
data specification options 86
estimation period 89
Exponential Smoothing 90
field options 86
general build options 90
missing value options 88
model options 93
observation options 86
time interval options 87
Streaming Time Series node
overview 85
stylesheets
deleting 195
exporting 195
importing 195
renaming 195
sum
Set Globals node 297
statistics output 292
Sum function
Sum function *(continued)*
time series aggregation 163
summary data 70
summary statistics
Data Audit node 282
summed values 71
SuperNode parameters 353, 354
SuperNodes
creating 350
creating caches for 354
editing 352
loading 355
locking 351
nesting 351
password protection 351
process SuperNodes 349
saving 355
scripting 355
setting parameters 353
source SuperNodes 349
terminal SuperNodes 350
types of 349
unlocking 351
using comments with 352
zooming in 352
supervised binning 151
surface graph 176
survey data
Data Collection source node 29
importing 30, 32, 33
Surveycraft data
importing 30
syntax tab
Statistics Output node 344
synthetic data
User Input node 42
system missing values
in Matrix tables 277
System variables
Data Collection source node 30
systematic samples 65, 66

T

t test
independent samples 294
Means node 294, 295
paired samples 294
t-SNE model nuggets 241
t-SNE node 238, 240
table browser
Generate menu 277
reordering columns 275, 277
searching 277
selecting cells 275, 277
Table node
output settings 275
output tab 275
settings tab 275
tables
joining 75
saving as text 275
saving output 275
tabular output

tabular output *(continued)*
reordering columns 275
selecting cells 275
tags 74, 80
templates
deleting 195
exporting 195
importing 195
renaming 195
Report node 296
temporal causal models
streaming TCM node 99
Teradata
query banding 19
test samples
partitioning data 155, 156
text
data 24, 27
delimited 24
text files
exporting 333
text output tab 308
The Weather Company 38
The Weather Company source 38
Third quartile
time series aggregation 163
thresholds
viewing bin thresholds 152
ties
Binning node 149
tiles
Binning node 149
time formats 133
Time Intervals node
overview 163
Time Plot node
appearance tab 212
plot tab 211
using the graph 212
time series 161
time series data
aggregating 163
Time Series models
ARIMA 92
transfer function order 92
transformation 92
timestamp 122
training samples
balancing 69
partitioning data 155, 156
transfer functions
delay 92
denominator orders 92
difference orders 92
numerator orders 92
seasonal orders 92
Transform node 289
transformations
reclassify 143, 147
recode 143, 147
transparency in graphs 168
transport files
SAS source node 39
Transpose node
Transpose node *(continued)*
field names 159
numeric fields 159
string fields 159
transposing data 159
Triple-S data
importing 30
True if any true function
time series aggregation 163
true values 130
truncating field names 134, 135
TWC source 38
type 6
type attributes 132
Type node
blank handling 127
clearing values 60
collection data type 130
continuous data 129
copying types 132
flag field type 130
geospatial data type 130
nominal data 129
ordinal data 129
overview 121
setting modeling role 131
setting options 122, 124, 126

U

unbiased data 69
undefined values 76
UNIQUE keyword
indexing database tables 318
unique records 82
unlocking SuperNodes 351
unsupported control characters 9
unused field exclusion
automated data preparation 111
usage type 6, 122
User Input node
overview 42
setting options 42
user-missing values
in Matrix tables 277

V

validation samples
partitioning data 155, 156
value labels
Statistics File node 28, 340
values
field and value labels 127
reading 127
specifying 127
values for derive formula 138
Variable File node
automatic date recognition 24
geospatial metadata 26
importing geospatial data 26
setting options 24
variable labels

variable labels *(continued)*
Statistics Export node 327, 346
Statistics File node 28, 340
variable names
data export 314, 326, 327, 333, 347
variable types
in visualizations 173
variance
statistics output 292
variance value for aggregation 71
VDATA format
Data Collection source node 29
viewing
HTML output in browser 274
vingtile bins 149
visualization
graphs and charts 167
visualization stylesheets
applying 265
deleting 195
exporting 195
importing 195
location 195
renaming 195
visualization templates
deleting 195
exporting 195
importing 195
location 195
renaming 195
visualizations
axes 259
categories 260
colors and patterns 256
copying 263
dashings 256
edit mode 254
editing 254
legend position 263
margins 258
number formats 258
padding 258
panels 260, 261
point aspect ratio 257
point rotation 257
point shape 257
scales 259
text 256
transforming coordinate systems 261
transparency 256
transpose 260, 261

W

Web node
adjusting points 223
adjusting thresholds 225
appearance tab 222
change layout 223
defining links 221
links slider 223
options tab 221
plot tab 220
slider 223
Web node *(continued)*
using the graph 223
web summary 226
weighted samples 67
weights
evaluation charts 230
worksheets
importing from Excel 39

X

XLSX files
exporting 333
XML export node 335
XML output
Report node 296
XML source node 40
XPath syntax 40

Z

zooming 352