Noeuds source, exécution et de sortie de IBM SPSS Modeler 15

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 537.

Cette version s'applique à IBM SPSS Modeler 15 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Matériel sous licence - Propriété d'IBM

© Copyright IBM Corporation 1994, 2012.

Droits limités pour les utilisateurs au sein d'administrations américaines : utilisation, copie ou divulgation soumise au GSA ADP Schedule Contract avec IBM Corp.

Préface

IBM® SPSS® Modeler est le puissant utilitaire de Data mining de IBM Corp.. SPSS Modeler aide les entreprises et les organismes à améliorer leurs relations avec les clients et les citoyens grâce à une compréhension approfondie des données. A l'aide des connaissances plus précises obtenues par le biais de SPSS Modeler, les entreprises et les organismes peuvent conserver les clients rentables, identifier les opportunités de vente croisée, attirer de nouveaux clients, détecter les éventuelles fraudes, réduire les risques et améliorer les services gouvernementaux.

L'interface visuelle de SPSS Modeler met à contribution les compétences professionnelles de l'utilisateur, ce qui permet d'obtenir des modèles prédictifs plus efficaces et de trouver des solutions plus rapidement. SPSS Modeler dispose de nombreuses techniques de modélisation, telles que les algorithmes de prévision, de classification, de segmentation et de détection d'association. Une fois les modèles créés, l'utilisateur peut utiliser IBM® SPSS® Modeler Solution Publisher pour les remettre aux responsables, où qu'ils se trouvent dans l'entreprise, ou pour les transférer vers une base de données.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics fournit des informations complètes, cohérentes et précises que les preneurs de décision utilisent avec confiance pour améliorer la performance du marché. Un portefeuille étendu d'outils de business intelligence, d'analyses prédictives, de performance financière et de gestion de stratégie, et des applications analytiques offre des connaissances claires, immédiates et appliquables pour améliorer l'efficacité actuelle ainsi que la capacité de prévoir les résultats futurs. Combinées avec de riches solutions industrielles, des pratiques éprouvées et des services professionnels, les organisations de toutes tailles peuvent atteindre la productivité la plus élevée, automatiser des décisions en toute tranquilité et fournir de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les organisations à prévoir des évènements futurs et à agir en conséquence pour mener à de meilleurs résultats Des clients dans le domaine commercial, gouvernemental et académique à travers le monde font confiance à la technologie IBM SPSS et considèrent qu'elle représente un avantage compétitif pour attirer, retenir et ajouter des clients, tout en réduisant la fraude et en atténuant les risques. En incorporant le logiciel IBM SPSS dans leur opérations quotidiennes, les organisations deviennent des entreprises prédictives – capables de diriger et d'automatiser les décisions pour atteindre les buts qu'ils se sont fixés et obtenir des avantages compétitifs sensibles. Pour informations supplémentaires ou pour joindre un revendeur, visitez le site *http://www.ibm.com/spss*.

Assistance technique

L'assistance technique est à la disposition des clients pour la maintenance des produits. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour joindre l'assistance technique, consultez le site Web de IBM Corp. à l'adresse *http://www.ibm.com/support*. Lorsque vous contactez l'assistance technique, n'oubliez pas de préparer vos identifiants, le nom de votre société et votre contrat d'assistance.

© Copyright IBM Corporation 1994, 2012.

Contenu

© Copyright IBM Corporation 1994, 2012.

Chapitre
1

A propos de IBM SPSS Modeler

À propos de IBM SPSS Modeler

IBM® SPSS® Modeler est un ensemble d'outils de data mining qui vous permet de développer rapidement, grâce à vos compétences professionnelles, des modèles prédictifs et de les déployer dans des applications professionnelles afin de faciliter la prise de décision. Conçu autour d'un modèle confirmé, le modèle CRISP-DM, SPSS Modeler prend en charge l'intégralité du processus de Data mining, des données à l'obtention de meilleurs résultats commerciaux.

SPSS Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques. Les méthodes disponibles dans la palette Modélisation vous permettent d'extraire de nouvelles informations de vos données et de développer des modèles prédictifs. Chaque méthode possède ses propres avantages et est donc plus adaptée à certains types de problème spécifiques.

Il est possible d'acquérir SPSS Modeler comme produit autonome ou de l'utiliser en tant que client en combinaison avec SPSS Modeler Server. Plusieurs autres options sont également disponibles, telles que décrites dans les sections suivantes. Pour plus d'informations, consultez *http://www.ibm.com/software/analytics/spss/products/modeler/*.

Produits IBM SPSS Modeler

La famille des produits IBM® SPSS® Modeler et les logiciels associés sont composés des éléments suivants.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adaptateurs pour IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler est une version complète du produit que vous installez et exécutez sur votre ordinateur personnel. Pour obtenir de meilleures performances lors du traitement d'ensembles de données volumineux, vous pouvez exécuter SPSS Modeler en mode local, comme produit autonome, ou l'utiliser en mode réparti, en association avec IBM® SPSS® Modeler Server.

© Copyright IBM Corporation 1994, 2012.

Avec SPSS Modeler, vous pouvez créer des modèles prédictifs précis rapidement et de manière intuitive, sans aucune programmation. L'interface visuelle unique vous permet de visualiser facilement le processus de Data mining. Grâce aux analyses avancées intégrées au produit, vous pouvez découvrir des motifs et tendances masqués dans vos données. Vous pouvez modéliser les résultats et comprendre les facteurs qui les influencent, afin d'exploiter les opportunités commerciales et de réduire les risques.

SPSS Modeler est disponible en deux éditions : SPSS Modeler Professional et SPSS Modeler Premium. Pour plus d'informations, reportez-vous à la section Éditions de IBM SPSS Modeler sur p. 3.

IBM SPSS Modeler Server

Grâce à une architecture client/serveur, SPSS Modeler adresse les demandes d'opérations très consommatrices de ressources à un logiciel serveur puissant. Il offre ainsi des performances accrues sur des ensembles de données plus volumineux.

SPSS Modeler Server est un produit avec licence distincte qui s'exécute en permanence en mode d'analyse réparti sur un hôte de serveur en combinaison avec une ou plusieurs installations de IBM® SPSS® Modeler. Ainsi, SPSS Modeler Server fournit des performances supérieures sur de grands ensembles de données car les opérations nécessitant beaucoup de mémoire peuvent être effectuées sur le serveur sans télécharger de données sur l'ordinateur client. IBM® SPSS® Modeler Server prend également en charge l'optimisation SQL et propose des fonctionnalités de modélisation dans la base de données pour des performances et une automatisation améliorées.

IBM SPSS Modeler Administration Console

Le Modeler Administration Console est une application graphique permettant de gérer de nombreuses options de SPSS Modeler Server qui peuvent également être configurées au moyen d'un fichier d'options. Cette application offre une interface utilisateur sous forme de console permettant de surveiller et de configurer les installations SPSS Modeler Server ; elle est disponible gratuitement pour les clients actuels de SPSS Modeler Server. L'application ne peut être installée que sur des ordinateurs Windows ; en revanche, elle peut administrer un serveur installé sur n'importe quelle plate-forme prise en charge.

IBM SPSS Modeler Batch

Alors que le Data mining est généralement un processus interactif, il est également possible d'exécuter SPSS Modeler à partir d'une ligne de commande sans recourir à l'interface utilisateur graphique. Par exemple, vous pouvez avoir des tâches longue durée ou répétitives à exécuter sans intervention de l'utilisateur. SPSS Modeler Batch est une version spécifique du produit qui prend en charge toutes les fonctions d'analyse de SPSS Modeler sans avoir besoin d'accéder à l'interface utilisateur standard. Une licence SPSS Modeler Server est nécessaire pour utiliser SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher est un outil qui permet de créer une version « packagée » d'un flux SPSS Modeler qui peut être exécutée par un moteur Runtime externe ou intégrée dans une application externe. Ainsi, vous pouvez publier et déployer des flux SPSS Modeler complets dans des environnements où SPSS Modeler n'est pas installé. SPSS Modeler Solution Publisher est fourni avec le service IBM SPSS Collaboration and Deployment Services - Scoring et nécessite une licence distincte. Avec cette licence, vous recevez SPSS Modeler Solution Publisher Runtime qui vous permet d'exécuter les flux publiés.

IBM SPSS Modeler Server Adaptateurs pour IBM SPSS Collaboration and Deployment Services

Différents adaptateurs pour IBM® SPSS® Collaboration and Deployment Services sont disponibles et permettent à SPSS Modeler et SPSS Modeler Server d'interagir avec un référentiel IBM SPSS Collaboration and Deployment Services. Ainsi, un flux SPSS Modeler déployé sur le référentiel peut être partagé par différents utilisateurs ou peut être accessible depuis l'application client léger IBM SPSS Modeler Advantage. Installez l'adaptateur sur le système qui héberge le référentiel.

Éditions de IBM SPSS Modeler

SPSS Modeler est disponible dans les éditions suivantes.

SPSS Modeler Professional

SPSS Modeler Professional offre tous les outils nécessaires à l'utilisation de la plupart des types de données structurées, tels que les comportements et interactions suivis dans les systèmes CRM, les caractéristiques sociodémographiques, les comportements d'achat et les données de vente.

SPSS Modeler Premium

SPSS Modeler Premium est un produit avec licence distincte qui étend le champ d'applications de SPSS Modeler Professional afin de pouvoir traiter des données spécialisées telles que celles utilisées pour les analyses d'entités ou les réseaux sociaux ainsi que des données de texte non structurées. SPSS Modeler Premium comprend les composants suivants :

IBM® SPSS® Modeler Entity Analytics ajoute une dimension entièrement nouvelle aux analyses prédictives IBM® SPSS® Modeler. Alors que les analyses prédictives essaient de prévoir les comportements futurs à partir de données passées, les analyses d'entités se concentrent sur l'amélioration de la cohérence des données actuelles en résolvant les conflits d'identités dans les enregistrements eux-mêmes. Une identité peut être celle d'un individu, d'une organisation, d'un objet ou d'une autre entité pour laquelle une ambiguïté peut exister. La résolution d'identité peut être vitale dans de nombreux domaines, y compris la gestion de la relation client, la détection de la fraude, le blanchiment d'argent et la sécurité nationale et internationale.

IBM SPSS Modeler Social Network Analysis transforme les informations sur les relations en champs qui caractérisent le comportement social des individus et des groupes. Grâce aux données qui décrivent les relations qui sous-tendent les réseaux sociaux, IBM® SPSS® Modeler Social Network Analysis identifie les chefs sociaux qui influencent le comportement des autres individus du réseau. De plus, il est possible de déterminer les individus qui sont le plus influencés par les autres participants du réseau. En combinant ces résultats avec d'autres mesures, il est possible de créer des profils détaillés des individus sur lesquels baser vos modèles prédictifs. Les modèles qui contiennent ces informations sociales seront plus efficaces que les modèles qui en sont dépourvus.

Text Analytics for IBM® SPSS® Modeler utilise des technologies linguistiques avancées et le traitement du langage naturel pour traiter rapidement une large variété de données textuelles non structurées, en extraire les concepts clés et les organiser pour les regrouper dans des catégories. Les concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

Documentation de IBM SPSS Modeler

Une documentation au format d'aide en ligne est disponible dans le menu Aide de SPSS Modeler. Vous y trouverez la documentation de SPSS Modeler, SPSS Modeler Server et de SPSS Modeler Solution Publisher, ainsi que le Guide des applications et d'autres documentations utiles.

La documentation complète de chaque produit (y compris les instructions d'installation) au format PDF est disponible dans le dossier *\Documentation* de chaque DVD de produit. Ces documents d'installation peuvent également être téléchargés sur Internet à l'adresse *http://www-01.ibm.com/support/docview.wss?uid=swg27023172*.

La documentation dans les deux formats est également disponible depuis le Centre d'informations SPSS Modeler à l'adresse *http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/*.

Documentation de SPSS Modeler Professional

La suite de documentation SPSS Modeler Professional (à l'exception des instructions d'installation) est la suivante.

- **Guide de l'utilisateur IBM SPSS Modeler.** Introduction générale à SPSS Modeler : création de flux de données, traitement des valeurs manquantes, création d'expressions CLEM, utilisation des projets et des rapports et regroupement des flux pour le déploiement dans IBM SPSS Collaboration and Deployment Services, des applications prédictives ou IBM SPSS Modeler Advantage.
- **Noeuds de Source, d'exécution et de sortie IBM SPSS Modeler.** Descriptions de tous les noeuds utilisés pour lire, traiter et renvoyer les données de sortie dans différents formats. En pratique, cela signifie tous les noeuds autres que les noeuds de modélisation.

- **IBM SPSS Modeler Noeuds de modélisation.** Description de tous les noeuds utilisés pour créer des modèles de Data mining. IBM® SPSS® Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques.
- **Guide des Algorithmes IBM SPSS Modeler.** Descriptions des fondements mathématiques des méthodes de modélisation utilisées dans SPSS Modeler. Ce guide est disponible au format PDF uniquement.
- **Guide des applications IBM SPSS Modeler.** Les exemples de ce guide fournissent des introductions brèves et ciblées aux méthodes et techniques de modélisation. Un version en ligne de ce guide est également disponible dans le menu Aide. Pour plus d'informations, reportez-vous à la section Exemples d'application sur p. 6.
- **Génération de scripts et automatisation IBM SPSS Modeler.** Informations sur l'automatisation du système via la génération de scripts, y compris les propriétés permettant de manipuler les noeuds et les flux.
- **IBM SPSS Modeler Guide de déploiement.** Informations sur l'exécution des scénarios et des flux SPSS Modeler comme étapes des tâches d'exécution sous IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Guide du développeur.** CLEF permet d'intégrer des programmes tiers tels que des programmes de traitement de données ou des algorithmes de modélisation en tant que noeuds dans SPSS Modeler.
- **Guide d'exploration de base de données IBM SPSS Modeler.** Informations sur la manière de tirer parti de la puissance de votre base de données pour améliorer les performances et étendre la gamme des fonctions analytiques via des algorithmes tiers.
- **Guide des performances et d'administration IBM SPSS Modeler Server.** Informations sur le mode de configuration et d'administration de IBM® SPSS® Modeler Server.
- **Guide de l'utilisateur de IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer SPSS Modeler Server. La console est implémentée en tant que plug-in à l'application Deployment Manager.
- **Guide IBM SPSS Modeler Solution Publisher.** SPSS Modeler Solution Publisher est un module complémentaire qui permet aux entreprises de publier des flux destinés à être utilisés en dehors de l'environnement SPSS Modeler.
- **Guide CRISP-DM IBM SPSS Modeler** Guide détaillé sur l'utilisation de la méthodologie CRISP-DM pour le Data mining avec SPSS Modeler
- **Guide de l'utilisateur IBM SPSS Modeler Batch.** Guide complet sur l'utilisation de IBM SPSS Modeler en mode par lots, avec des détails sur l'exécution en mode par lots et les arguments de ligne de commande. Ce guide est disponible au format PDF uniquement.

Documentation de SPSS Modeler Premium

La suite de documentation SPSS Modeler Premium (à l'exception des instructions d'installation) est la suivante.

- **IBM SPSS Modeler Entity Analytics Guide de l'utilisateur.** Informations sur l'utilisation des analyses d'entités avec SPSS Modeler, notamment l'installation et la configuration du référentiel, les nœuds d'analyses d'entités et les tâches administratives.
- **IBM SPSS Modeler Social Network Analysis Guide de l'utilisateur.** Guide sur l'exécution des analyses de réseaux sociaux avec SPSS Modeler, y compris les analyses de groupe et analyses de diffusion.
- **Text Analytics for SPSS Modeler Guide de l'utilisateur.** Informations sur l'utilisation des analyses de texte avec SPSS Modeler, notamment sur les nœuds de Text Mining, l'espace de travail interactif, les modèles et d'autres ressources.
- Guide de l'utilisateur de **Text Analytics for IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer IBM® SPSS® Modeler Server pour l'utiliser avec Text Analytics for SPSS Modeler. La console est implémentée en tant que plug-in à l'application Deployment Manager.

Exemples d'application

Tandis que les outils de Data mining de SPSS Modeler peuvent vous aider à résoudre une grande variété de problèmes commerciaux et organisationnels, les exemples d'application fournissent des introductions brèves et ciblées aux méthodes et aux techniques de modélisation. Les ensembles de données utilisés ici sont beaucoup plus petits que les énormes entrepôts de données gérés par certains Data miners, mais les concepts et les méthodes impliqués doivent pouvoir être adaptés à des applications réelles.

Vous pouvez accéder aux exemples en cliquant Exemples d'application dans le menu Aide de SPSS Modeler. Les fichiers de données et les flux d'échantillons sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. Pour plus d'informations, reportez-vous à la section Dossier Demos sur p. 6.

Exemples de modélisation de bases de données. Consultez les exemples dans le *IBM SPSS ModelerGuide d'exploration de base de données*.

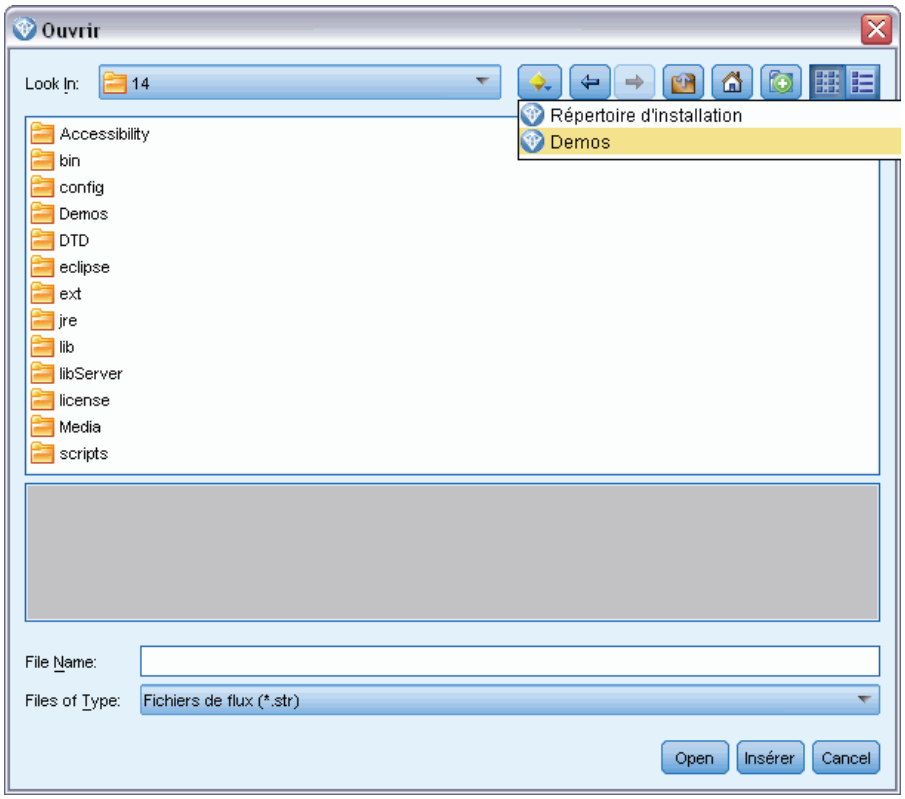
Exemples de génération de scripts. Consultez les exemples dans le *IBM SPSS ModelerGuide de génération de scripts et d'automatisation*.

Dossier Demos

Les fichiers de données et les flux d'échantillons utilisés avec les exemples d'application sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. Ce dossier est également accessible à partir du groupe de programmes sous IBM SPSS Modeler 15 dans le menu

Démarrer de Windows, ou en cliquant sur *Demos* dans la liste des répertoires récents de la boîte de dialogue Ouverture de fichier.

Figure 1-1
Sélection du dossier Demos dans la liste des répertoires récemment consultés



Chapitre
2

Noeuds source

Présentation

Les noeuds source vous permettent d'importer des données stockées dans différents formats : fichiers plats, IBM® SPSS® Statistics (*.sav*), SAS, Microsoft Excel et bases de données relationnelles compatibles ODBC, entre autres. Vous pouvez également générer des données synthétiques à l'aide du nœud Utilisateur.

La palette Sources contient les noeuds suivants :

Le noeud Enterprise View crée une connexion à un IBM® SPSS® Collaboration and Deployment Services Repository, vous permettant de lire des données Enterprise View dans un flux et de regrouper un modèle dans un scénario accessible depuis le référentiel par d'autres utilisateurs. Pour plus d'informations, reportez-vous à la section Noeud Enterprise View sur p. 9.

Le noeud SGBD peut être utilisé pour importer des données provenant de nombreux autres logiciels utilisant la connectivité ODBC (Open Database Connectivity), tels que Microsoft SQL Server, DB2, Oracle, etc. Pour plus d'informations, reportez-vous à la section Nœud Source de base de données sur p. 15.

Le noeud Délimité lit les données de fichiers texte de longueur variable, c'est-à-dire les fichiers dont les enregistrements contiennent un nombre fixe de champs et un nombre variable de caractères. Ce noeud est également utile pour les fichiers contenant des textes d'en-tête de longueur fixe et certains types d'annotation. Pour plus d'informations, reportez-vous à la section Noeud Délimité sur p. 25.

Le noeud Fixe permet d'importer les données de fichiers texte de longueur fixe, c'est-à-dire les fichiers dont les champs ne sont pas délimités, mais commencent au même endroit et sont de longueur fixe. Les données générées automatiquement ou héritées sont souvent stockées au format de longueur fixe. Pour plus d'informations, reportez-vous à la section Noeud Fixe sur p. 29.

Le noeud Statistics lit les données du format de fichier *.sav* utilisé par SPSS Statistics ainsi que des fichiers cache enregistrés dans IBM® SPSS® Modeler, qui utilisent le même format. Pour plus d'informations, reportez-vous à la section Noeud Statistics dans le chapitre 8 sur p. 497.

Le noeud IBM® SPSS® Data Collection importe des données d'enquête dans différents formats utilisés par le logiciel d'étude de marché et conformément au modèle de données de Data Collection. Pour pouvoir utiliser ce noeud, vous devez avoir installé auparavant Data Collection Developer Library. Pour plus d'informations, reportez-vous à la section Noeud Data Collection sur p. 36.

Le noeud source IBM Cognos BI importe des données depuis les bases de données Cognos BI. Pour plus d'informations, reportez-vous à la section Importation des données Cognos sur p. 46.

© Copyright IBM Corporation 1994, 2012.

Le noeud Fichier SAS permet d'importer des données SAS dans SPSS Modeler. Pour plus d'informations, reportez-vous à la section Noeud source SAS sur p. 51.

Le noeud Excel permet d'importer des données issues de n'importe quelle version de Microsoft Excel. Aucune source de données ODBC n'est requise. Pour plus d'informations, reportez-vous à la section Noeud source Excel sur p. 53.

Le noeud source XML importe des données au format XML dans le flux. Vous pouvez importer un fichier ou tous les fichiers dans un répertoire. Vous pouvez aussi spécifier un fichier de schéma à partir duquel lire la structure XML. Pour plus d'informations, reportez-vous à la section Noeud source XML sur p. 54.

Le noeud Utilisateur représente une façon simple de créer des données synthétiques (à partir de zéro ou en modifiant des données existantes). Ceci est utile, par exemple, si vous souhaitez créer un ensemble de données de test pour la modélisation. Pour plus d'informations, reportez-vous à la section Noeud Utilisateur sur p. 58.

Pour commencer un flux, ajoutez un noeud source à l'espace de travail de flux. Ensuite, double-cliquez sur le noeud pour ouvrir la boîte de dialogue correspondante. Dans les différents onglets de la boîte de dialogue, vous pouvez lire les données, afficher les champs et les valeurs, et définir différentes options, comme les filtres, les types de données, le rôle du champ et la détection des valeurs manquantes.

Noeud Enterprise View

Le noeud Enterprise View vous permet de créer et de conserver une connexion entre une session IBM® SPSS® Modeler et une vue Enterprise View dans un IBM® SPSS® Collaboration and Deployment Services Repository partagé. Vous pouvez ainsi lire les données à partir d'une vue Enterprise View dans un flux SPSS Modeler et regrouper un modèle SPSS Modeler dans un scénario accessible par d'autres utilisateurs du référentiel partagé.

Un **scénario** est un fichier contenant un flux SPSS Modeler avec des noeuds spécifiques, des modèles et des propriétés supplémentaires qui permettent de le déployer vers un IBM SPSS Collaboration and Deployment Services Repository à des fins de scoring ou en vue de rafraîchissement de modèle automatique. L'utilisation des noeuds Enterprise View avec des scénarios garantit que, dans une situation multi-utilisateur, tous les utilisateurs utilisent les mêmes données. Une **connexion** est un lien d'une session SPSS Modeler vers une vue Enterprise View dans le IBM SPSS Collaboration and Deployment Services Repository.

La **Enterprise View** est l'ensemble complet des données appartenant à une organisation, quel que soit l'endroit où les données sont physiquement situées. Chaque connexion comprend une sélection spécifique d'une seule **Vue d'application** (sous-ensemble d'Enterprise View adapté à une application particulière), une **Définitions du fournisseur de données** ou DPD (relie les tableaux/colonnes d'une Vue d'application logique à une source de données physique) et d'un **environnement** (identifie les colonnes spécifiques qu'il convient d'associer aux segments de marché définis). La vue Enterprise View, la Vue d'application et les définitions DPD sont mémorisées avec les versions dans le référentiel, bien que les données réelles résident dans une ou plusieurs bases de données ou d'autres sources externes.

Lorsqu'une connexion a été établie, vous indiquez un **tableau** de Vue d'application à utiliser dans SPSS Modeler. Dans une Vue d'application, un tableau est une vue logique comprenant certaines colonnes ou toutes les colonnes d'un ou de plusieurs tableaux dans une ou plusieurs bases de données physiques. Ainsi, le noeud Enterprise View permet de voir des enregistrements de plusieurs tables de bases de données comme un seul tableau dans SPSS Modeler.

Conditions requises

- Pour utiliser le noeud Enterprise View, un IBM SPSS Collaboration and Deployment Services Repository doit d'abord être installé et configuré sur votre site, avec une vue Enterprise View, des vues d'application et des DPD déjà définis.

Remarque : Une licence distincte est requise pour accéder à un référentiel IBM® SPSS® Collaboration and Deployment Services. Pour plus d'informations, reportez-vous à *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*

- En outre, le IBM® SPSS® Collaboration and Deployment Services Enterprise View Driver doit être installé sur chaque ordinateur utilisé pour modifier ou exécuter le flux. Pour Windows, installez simplement le lecteur sur l'ordinateur où IBM® SPSS® Modeler ou IBM® SPSS® Modeler Server est installé et où aucune autre configuration du lecteur n'est nécessaire. Sur UNIX, une référence au script *pev.sh* doit être ajoutée au script de démarrage. Contactez votre administrateur local pour des détails sur l'installation du IBM SPSS Collaboration and Deployment Services Enterprise View Driver.
- Un DPD est défini par rapport à une source de données ODBC particulière. Pour utiliser un DPD à partir de SPSS Modeler, vous devez avoir une source de données ODBC définie sur l'hôte serveur SPSS Modeler qui a le même nom et qui se connecte au même magasin de données que celui référencé dans le DPD.

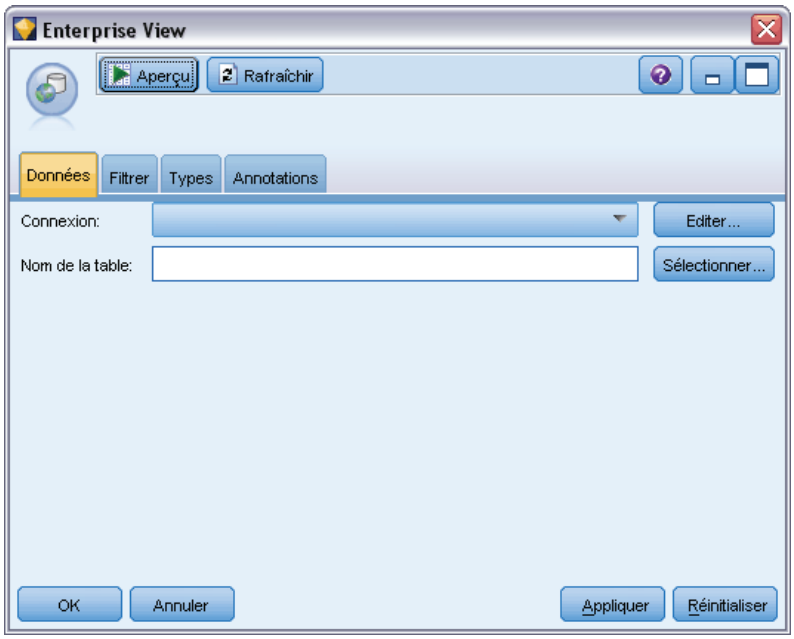
Paramétrage des options du noeud Enterprise View

Vous pouvez utiliser les options de l'onglet Données de la boîte de dialogue Enterprise View pour :

- Sélectionner une connexion existante à un référentiel
- Editer une connexion existante à un référentiel
- Créer une connexion à un référentiel
- Sélectionner une table Vue d'application

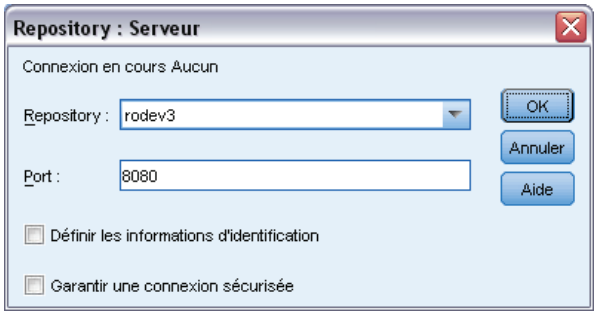
Pour plus de détails sur l'utilisation des référentiels, reportez-vous au manuel *Guide de l'administrateur IBM® SPSS® Collaboration and Deployment Services*.

Figure 2-1
Ajout d'une connexion à un IBM SPSS Collaboration and Deployment Services Repository



Connexion. La liste déroulante fournit des options permettant de sélectionner une connexion existante à un référentiel, d'éditer une connexion existante ou d'ajouter une connexion. Si vous êtes déjà connecté à un référentiel via IBM® SPSS® Modeler, le choix de l'option Ajouter/Editer une connexion affiche la boîte de dialogue Connexions Enterprise View, où vous pouvez définir ou éditer les détails requis pour la connexion active. Si vous n'êtes pas connecté, cette option affiche la boîte de dialogue Connexion du référentiel.

Figure 2-2
Connexion à un référentiel



Pour plus d'informations sur la connexion au référentiel, reportez-vous au *guide de l'utilisateur de SPSS Modeler*.

Lorsqu'une connexion à un référentiel a été établie, cette connexion reste en place jusqu'à ce que vous quittiez SPSS Modeler. Une connexion peut être partagée par d'autres noeuds dans le même flux, mais vous devez créer une nouvelle connexion pour chaque nouveau flux.

Une connexion réussie affiche la boîte de dialogue Connexions Enterprise View.

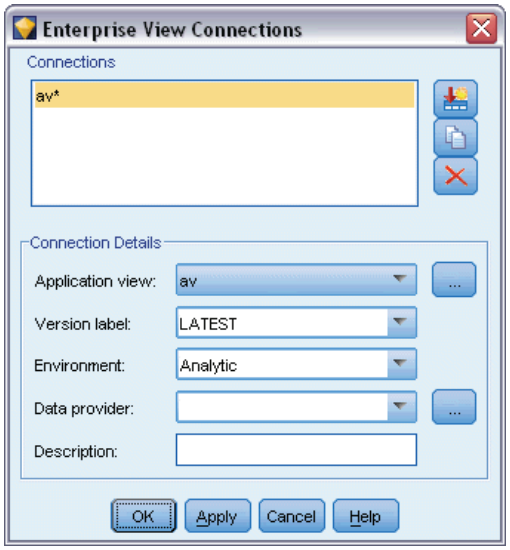
Nom de la table. Ce champ est vide à l'origine et ne peut pas être rempli tant que vous ne créez pas une connexion. Si vous connaissez le nom de la table de Vue d'application à laquelle vous souhaitez accéder, indiquez-le dans le champ Nom de la table. Dans le cas contraire, cliquez sur le bouton Sélectionner pour ouvrir une boîte de dialogue répertoriant les tables de Vue d'application disponibles.

Connexions Enterprise View

Cette boîte de dialogue permet de définir ou d'éditer les détails requis relatifs à la connexion au référentiel. Vous pouvez spécifier les éléments suivants :

- Vue d'application et version
- Environnement
- Définitions du fournisseur de données (DPD)
- Description de la connexion.

Figure 2-3
Choix d'une Vue d'application



Connexions. Répertorie les connexions existantes au référentiel.

- **Ajouter une nouvelle connexion.** Affiche la boîte de dialogue Extraire l'objet où vous pouvez rechercher et sélectionner une Vue d'application dans le référentiel.
- **Copier la connexion sélectionnée.** Copie une connexion sélectionnée, vous épargnant d'avoir de nouveau à rechercher la même Vue d'application.
- **Supprimer la connexion sélectionnée.** Supprime la connexion sélectionnée de la liste.

Détails de connexion. Pour la connexion sélectionnée dans le panneau Connexions, cette option affiche la Vue d'application, l'étiquette de version, l'environnement, la définition DPD et le texte descriptif.

- **Vue d'application.** La liste déroulante affiche la Vue d'application sélectionnée, le cas échéant. S'il existe des connexions à d'autres Vues d'application dans la session actuelle, elles apparaissent aussi dans la liste déroulante. Cliquez sur le bouton Naviguer pour rechercher d'autres Vues d'application dans le référentiel.
- **Etiquette de version.** Ce champ déroulant répertorie toutes les étiquettes de version définies pour la Vue d'application spécifiée. Les étiquettes de version permettent d'identifier des versions d'objets de référentiel spécifiques. Par exemple, il peut exister deux versions d'une Vue d'application spécifique. Vous pouvez, par exemple, spécifier l'étiquette TEST pour la version utilisée dans l'environnement de développement et l'étiquette PRODUCTION pour la version utilisée dans l'environnement de production. Sélectionnez l'étiquette appropriée.

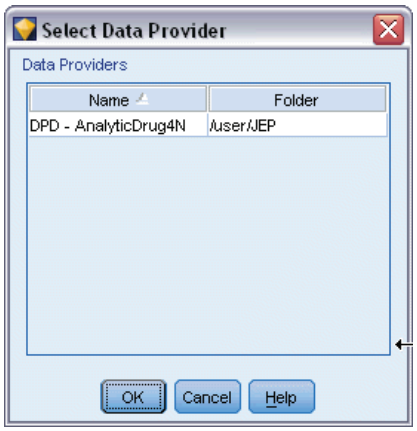
Remarque : Les étiquettes ne doivent pas contenir de caractère "[", car le nom de la table ne s'afficherait pas dans l'onglet Données de la boîte de dialogue Enterprise View.

- **Environnement.** Ce champ déroulant répertorie tous les environnements valides. Le paramètre d'environnement détermine les DPD qui sont disponibles, indiquant ainsi les colonnes particulières à associer aux segments commerciaux définis. Par exemple, lorsque l'option Analytique est sélectionnée, seules les colonnes de Vue d'application définies comme Analytique sont renvoyées. L'environnement par défaut est Analytique ; vous pouvez également choisir l'option Opérationnel.
- **Fournisseur de données.** La liste déroulante affiche les noms d'un maximum de dix Définitions du fournisseur de données pour la Vue d'application sélectionnée. Seuls les DPD qui font référence à la Vue d'application sélectionnée sont représentés. Cliquez sur le bouton adjacent Naviguer pour afficher le nom et le chemin de tous les DPD liés à la Vue d'application actuelle.
- **Description :** Texte descriptif relatif à la connexion au référentiel. Ce texte sera utilisé pour le nom de la connexion. Si vous cliquez sur OK, le texte apparaît dans la liste déroulante Connexion, dans la barre de titre de la boîte de dialogue Enterprise View et en tant qu'étiquette du noeud Enterprise View dans l'espace de travail.

Choix du DPD

La boîte de dialogue Sélection du fournisseur de données affiche le nom et le chemin de tous les DPD qui font référence à la Vue d'application actuelle.

Figure 2-4
Choix d'un DPD



Les Vues d'application peuvent avoir plusieurs DPD pour prendre en charge les différentes étapes d'un projet. Par exemple, les données historiques utilisées pour créer un modèle peuvent provenir d'une seule base de données, tandis que les données opérationnelles proviennent d'une autre.

Un DPD est défini par rapport à une source de données ODBC particulière. Pour utiliser un DPD à partir de IBM® SPSS® Modeler, vous devez avoir une source de données ODBC définie sur l'hôte serveur SPSS Modeler qui a le même nom et qui se connecte au même magasin de données que celui référencé dans le DPD.

- Pour choisir un DPD à utiliser, sélectionnez son nom dans la liste et cliquez sur OK.

Choix de la table

La boîte de dialogue Sélectionner un tableau répertorie tous les tableaux qui sont référencés dans la Vue d'application actuelle. La boîte de dialogue est vide si aucune connexion au IBM SPSS Collaboration and Deployment Services Repository n'a été effectuée.

Figure 2-5
Choix d'une table

- Pour choisir un tableau à utiliser, sélectionnez son nom dans la liste et cliquez sur OK.

Nœud Source de base de données

Le noeud Source de base de données peut être utilisé pour importer des données provenant de nombreux autres logiciels utilisant la connectivité ODBC (Open Database Connectivity), tels que Microsoft SQL Server, DB2, Oracle, etc.

Pour lire ou écrire sur une base de données, vous devez installer et configurer une source de données ODBC pour la base de données appropriée, avec, le cas échéant, des autorisations en lecture et en écriture. Le IBM® SPSS® Data Access Pack contient un ensemble de pilotes ODBC qui peuvent être utilisés dans ce but, et ces pilotes sont disponibles sur le DVD de IBM SPSS Data Access Pack ou depuis le site de téléchargement. Si vous avez des questions sur la création ou la définition d'autorisations pour les sources de données ODBC, contactez l'administrateur de votre base de données.

Dans IBM® SPSS® Modeler, la prise en charge de la base de données est classée en trois niveaux différents de prise en charge pour l'optimisation et le pushback SQL, en fonction du fournisseur de la base de données. Les différents niveaux de prise en charge sont implémentés au moyen d'un certain nombre de paramètres système qui peuvent être personnalisés pour faire partie du contrat de services SPSS.

Les trois niveaux de prise en charge de la base de données sont :

Table 2-1
niveaux de prise en charge de la base de données

Niveau de prise en charge	Description
Niveau 1	Tout pushback SQL possible est disponible, avec l'optimisation SQL spécifique à la base de données.
Niveau 2	La plupart des pushback SQL possibles sont disponibles, sans optimisation SQL spécifique à la base de données.
Niveau 3	Aucune répercussion SQL ou optimisation : uniquement la lecture des données depuis la base de données et l'écriture des données dans la base de données sont disponibles.

Pilotes ODBC pris en charge

Pour obtenir les informations les plus récentes sur les bases de données et pilotes ODBC pris en charge et testés pour une utilisation avec SPSS Modeler 15, consultez les matrices de compatibilité des produits sur le site Web de support technique de l'entreprise (*http://www.ibm.com/support*).

Où installer les pilotes

Vous devez installer et configurer les pilotes ODBC sur chaque ordinateur où le traitement a lieu.

- Si vous exécutez IBM® SPSS® Modeler en mode local (autonome), vous devez installer les pilotes sur l'ordinateur local.
- Si vous exécutez SPSS Modeler en mode distribué sur IBM® SPSS® Modeler Server en mode distant, les pilotes ODBC doivent être installés sur le même ordinateur d'installation que SPSS Modeler Server. Pour SPSS Modeler Server sur les systèmes UNIX, consultez également « Configuration des pilotes ODBC sur les systèmes UNIX » plus avant dans cette section.

- Si vous devez accéder aux mêmes sources de données provenant de SPSS Modeler et de SPSS Modeler Server, les pilotes ODBC doivent être installés sur les deux ordinateurs.
- Si vous exécutez SPSS Modeler sur Terminal Services, vous devez installer les pilotes ODBC sur le serveur Terminal Services sur lequel vous disposez de SPSS Modeler.
- Si vous utilisez IBM® SPSS® Modeler Solution Publisher Runtime pour exécuter des flux publiés sur un ordinateur distinct, vous devez aussi installer et configurer les pilotes ODBC sur cet ordinateur.

Remarque : Si vous utilisez SPSS Modeler Server sous UNIX pour accéder à une base de données Teradata, vous devez utiliser le gestionnaire de pilote ODBC installé avec le pilote ODBC Teradata. Afin de procéder à ces modifications dans SPSS Modeler Server, veuillez spécifier une valeur pour ODBC_DRIVER_MANAGER_PATH en haut du script modelersrv.sh, à l'endroit indiqué par les commentaires. Cette variable d'environnement doit être définie sur l'emplacement du gestionnaire de pilote ODBC fourni avec le pilote ODBC Teradata (/usr/odbc/lib dans une installation du pilote ODBC Teradata par défaut). Vous devez redémarrer SPSS Modeler Server pour que la modification prenne effet. Pour obtenir plus de détails sur les plateformes de SPSS Modeler Server qui prennent en charge l'accès à Teradata, et sur la version du pilote ODBC Teradata prise en charge, consultez le site Web de support technique de l'entreprise à l'adresse *http://www.ibm.com/support*.

Configuration des pilotes ODBC sur les systèmes UNIX

Par défaut, le gestionnaire de pilote DataDirect n'est pas configuré pour SPSS Modeler Server sur les systèmes UNIX. Pour configurer le chargement du gestionnaire de pilote DataDirect sur UNIX, saisissez les commandes suivantes :

```
cd modeler_server_install_directory/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

Le lien par défaut est alors supprimé et un lien vers le gestionnaire de pilote DataDirect est créé.

Pour accéder aux données à partir d'une base de données, utilisez la procédure générale suivante :

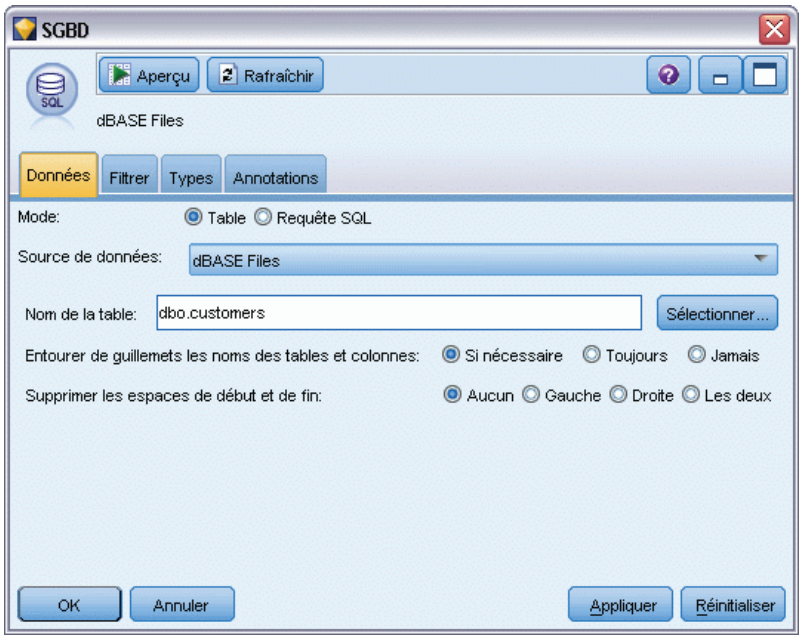
- Installez un pilote ODBC et configurez une source de données pour la base de données à utiliser.
- Dans la boîte de dialogue du nœud SGBD, connectez-vous à une base de données en mode Table ou en mode Requête SQL.
- Sélectionnez une table dans la base de données.
- A l'aide des onglets de la boîte de dialogue du nœud SGBD, vous pouvez modifier les types d'utilisation et filtrer les champs de données.

Cette procédure est détaillée dans les rubriques suivantes.

Définition des options du nœud SGBD

Vous pouvez utiliser les options de l'onglet Données de la boîte de dialogue du noeud Source de base de données pour accéder à une base de données et lire les données d'une table sélectionnée.

Figure 2-6
Chargement de données par sélection d'une table



Mode. Sélectionnez Table pour vous connecter à une table à l'aide des commandes de la boîte de dialogue.

Sélectionnez Requête SQL pour interroger la base de données sélectionnée en utilisant SQL. Pour plus d'informations, reportez-vous à la section Interrogation de la base de données sur p. 23.

Source de données. En mode Table et Requête SQL, vous pouvez entrer un nom dans le champ Source de données ou sélectionner Ajouter une nouvelle connexion à la base de données dans la liste déroulante.

Les options suivantes vous permettent de vous connecter à une base de données et de sélectionner une table à l'aide de la boîte de dialogue :

Nom de la table. Si vous connaissez le nom de la table à laquelle vous souhaitez accéder, indiquez-le dans le champ Nom de la table. Dans le cas contraire, cliquez sur le bouton Sélectionner pour ouvrir une boîte de dialogue répertoriant les tables disponibles.

Entourer de guillemets les noms des tables et colonnes. Indiquez si vous souhaitez que les noms des tables et des colonnes soient placés entre guillemets lors de l'envoi des requêtes à la base de données (par exemple, s'ils contiennent des espaces ou des signes de ponctuation).

- Si vous sélectionnez Si nécessaire, les noms des tables et des champs seront placés entre guillemets *uniquement* s'ils contiennent des caractères non standard. Les caractères non standard sont les caractères non ASCII, l'espace et tous les caractères non alphanumériques autres que le point (.).

- Sélectionnez Jamais si vous ne souhaitez *jamais* mettre les noms des tables et des champs entre guillemets.
- Sélectionnez Toujours si vous souhaitez que *tous* les noms de tables et de champs soient mis entre guillemets.

Supprimer les espaces de début et de fin. Sélectionnez les options permettant la suppression des espaces situés en début et en fin des chaînes.

Remarque. Les comparaisons entre les chaînes qui utilisent ou non les répercussions SQL peuvent produire des résultats différents lorsqu'il existe des espaces en fin de chaîne.

Lecture de chaînes vides issues d'Oracle. Lorsque vous lisez une base de données Oracle ou que vous écrivez dedans, souvenez-vous que, contrairement à IBM® SPSS® Modeler et à la plupart des autres bases de données, Oracle traite et stocke les valeurs de chaîne vides comme des valeurs nulles. Autrement dit, les mêmes données extraites d'une base de données Oracle, ou d'un fichier ou d'une autre base de données peuvent se comporter différemment, et donc renvoyer des résultats différents.

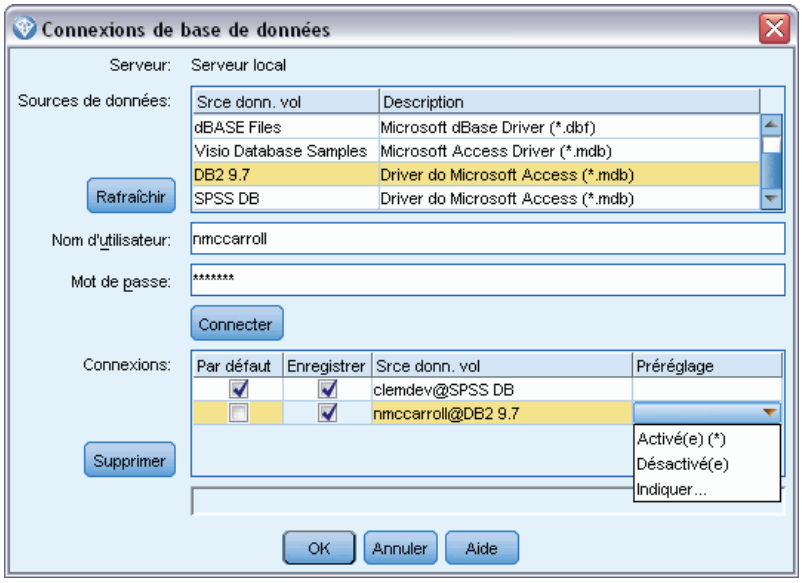
Ajout d'une connexion à la base de données

Pour ouvrir une base de données, vous devez d'abord sélectionner la source de données avec laquelle vous souhaitez établir une connexion. Dans l'onglet Données, sélectionnez Ajouter une nouvelle connexion à la base de données dans la liste déroulante Source de données.

La boîte de dialogue Connexions de base de données apparaît.

Remarque : Pour ouvrir cette boîte de dialogue à partir du menu principal, choisissez : Outils > Bases de données...

Figure 2-7
Boîte de dialogue Connexions de base de données



Sources de données. Répertorie les sources de données disponibles. Si la base de données souhaitée n'apparaît pas, faites défiler la liste. Une fois que vous avez sélectionné la source de données et entré les mots de passe, cliquez sur Connecter. Cliquez sur Rafraîchir pour mettre à jour la liste.

Nom d'utilisateur. Si la source de données est protégée par un mot de passe, entrez votre nom d'utilisateur.

Mot de passe. Si la source de données est protégée par un mot de passe, entrez-le ici.

Connexions. Indique les bases de données actuellement connectées.

- **Défaut :** En option, vous pouvez choisir une connexion par défaut. Cette action prédéfinit cette connexion comme source de données des noeuds source de base de données et d'exportation, mais peut être modifiée au besoin.
- **Enregistrer.** Vous pouvez également sélectionner une ou plusieurs connexions à afficher de nouveau dans les prochaines sessions.
- **Source de données.** Les chaînes de connexion pour les bases de données actuellement connectées.
- **Prédéfinir.** Indique (à l'aide du caractère *) si des valeurs prédéfinies ont été spécifiées pour la connexion de la base de données. Pour spécifier des valeurs prédéfinies, cliquez sur cette colonne dans la ligne correspondant à la connexion de la base de données et choisissez Spécifier dans la liste. Pour plus d'informations, reportez-vous à la section Spécification de valeurs prédéfinies pour une connexion de la base de données sur p. 19.

Pour supprimer des connexions, sélectionnez-les dans la liste, puis cliquez sur Supprimer.

Une fois vos sélections effectuées, cliquez sur OK.

Spécification de valeurs prédéfinies pour une connexion de la base de données

Pour certaines bases de données, il est possible de spécifier des paramètres par défaut pour la connexion de la base de données. Les paramètres concernent tous l'exportation de la base de données.

Les bases de donnés prenant en charge cette fonctionnalité sont les suivantes.

- IBM InfoSphere Warehouse s'exécutant sur DB2 9.1 ou une version ultérieure. Pour plus d'informations, reportez-vous à la section Paramètres pour IBM DB2 InfoSphere Warehouse sur p. 20.
- SQL Server 2008 ou des éditions Enterprise et Developer ultérieures. Pour plus d'informations, reportez-vous à la section Paramètres pour SQL Server sur p. 20.
- Oracle 10g et 11gR1 ou des éditions Enterprise ou Personal ultérieures. Pour plus d'informations, reportez-vous à la section Paramètres pour Oracle sur p. 20.
- IBM Netezza, IBM DB2 pour z/OS et Teradata se connectent à une base de données ou à un schéma de la même façon. Pour plus d'informations, reportez-vous à la section Paramètres de IBM Netezza, IBM DB2 pour z/OS et Teradata sur p. 22.

Si vous êtes connecté à une base de données ou à un schéma qui ne prend pas en charge cette fonctionnalité, le message Aucun préréglage ne peut être configuré pour cette connexion à la base de données apparaît.

Paramètres pour IBM DB2 InfoSphere Warehouse

Ces paramètres s'affichent pour IBM InfoSphere Warehouse s'exécutant sur DB2 9.1 ou une version ultérieure.

Espace Table. L'espace Table utilisé pour l'exportation. Les administrateurs de la base de données peuvent créer ou configurer des espaces tables partitionnés. Nous vous recommandons de sélectionner un de ces espaces tables (plutôt que celui par défaut) pour l'exportation vers la base de données.

Utiliser la compression. Si cette option est sélectionnée, elle crée des tables compressées pour l'exportation (par exemple, l'équivalent de CREATE TABLE MYTABLE(…) COMPRESS YES; en SQL).

Ne pas enregistrer les mises à jour. Si cette option est sélectionnée, aucune consignation n'est effectuée lors de la création de tables, ni lors de l'insertion de données (l'équivalent de CREATE TABLE MYTABLE(…) NOT LOGGED INITIALLY; en SQL).

Paramètres pour SQL Server

Ces paramètres sont affichés pour SQL Server 2008 ou les éditions Enterprise et Developer ultérieures.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Ligne.** Active la compression au niveau des lignes (par exemple, l'équivalent de CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW); en SQL).
- **Page.** Active la compression au niveau des pages (par exemple, CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE); en SQL).

Paramètres pour Oracle

Paramètres d'Oracle 10g

Ces paramètres sont affichés pour Oracle 10g édition Enterprise ou Personal.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées. Pour cette version de la base de données, seule la compression basique est disponible (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL).

Paramètres d'Oracle 11gR1

Ces paramètres sont affichés pour Oracle 11g R1 édition Enterprise ou Personal.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Opérations de chargement direct.
- **Opérations de chargement direct.** Active la compression des opérations d'insertion en masse (directes) uniquement (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR DIRECT_LOAD OPERATIONS; en SQL).
- **Toutes les opérations.** Active la compression de toutes les opérations (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR ALL OPERATIONS; en SQL).

Paramètres d'Oracle 11gR2 - option basique

Ces paramètres sont affichés pour Oracle 11g R2 édition Enterprise ou Personal avec l'option Basique.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Basique.
- **Basique.** Active la compression basique (par exemple, CREATE TABLE MYTABLE(…) COMPRESS BASIC; en SQL).

Paramètres d'Oracle 11gR2 - option avancée

Ces paramètres sont affichés pour Oracle 11g R2 édition Enterprise ou Personal avec l'option Avancée.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Basique.
- **Basique.** Active la compression basique (par exemple, CREATE TABLE MYTABLE(…) COMPRESS BASIC; en SQL).
- **OLTP.** Active la compression OLTP (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP; en SQL).
- **Requête faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les requêtes (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW; or CREATE TABLE MYTABLE(…)COMPRESS

FOR QUERY HIGH; en SQL). La compression des requêtes est utile dans les environnements d'entreposage de données ; HIGH fournit un taux de compression plus grand que LOW.

- **Archive faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les archives (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW; ou CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH; en SQL). La compression des archives est utile pour compresser des données qui seront stockées pendant de longues périodes ; HIGH fournit un taux de compression plus grand que LOW.

Paramètres de IBM Netezza, IBM DB2 pour z/OS et Teradata

Lorsque vous spécifiez des préréglages pour IBM Netezza, IBM DB2 pour z/OS, ou Teradata, il vous est demandé de sélectionner les éléments suivants :

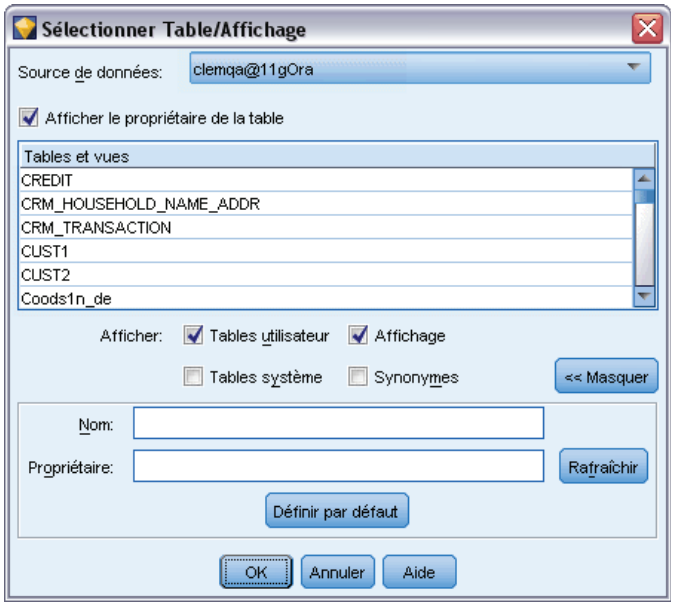
Utiliser la base de données / schéma de l'adaptateur de scoring de serveur. Si cette option est sélectionnée, l'option Base de données / schéma de l'adaptateur de scoring de serveur est activée.

Base de données / schéma de l'adaptateur de scoring de serveur. Dans la liste déroulante, sélectionnez la connexion adaptée.

Sélection d'une table de base de données

Une fois connecté à une source de données, vous pouvez importer des champs à partir d'une table ou d'une vue donnée. Dans l'onglet Données de la boîte de dialogue SGBD, vous pouvez entrer le nom d'une table dans le champ Nom de la table ou cliquer sur Sélectionner pour ouvrir une boîte de dialogue répertoriant les tables et vues disponibles.

Figure 2-8
Sélection d'un tableau à partir de la base de données actuellement connectée



Afficher le propriétaire de la table. Cochez cette case si l'accès à une table d'une source de données requiert que le propriétaire de la table soit identifié. Désélectionnez cette option pour les sources de données qui ne requièrent pas cette identification.

Remarque : pour les bases de données SAS et Oracle, l'identification du propriétaire est généralement obligatoire.

Tables et vues. Sélectionnez la table ou la vue à importer.

Afficher. Répertorie les colonnes de la source de données à laquelle vous êtes actuellement connecté. Cliquez sur l'une des options suivantes pour personnaliser la vue des tables disponibles :

- Cliquez sur Tables utilisateur pour afficher les tables de base de données ordinaires créées par les utilisateurs de la base de données.
- Cliquez sur Tables système pour afficher les tables de base de données appartenant au système (par exemple, les tables qui fournissent des informations sur la base de données, comme les détails des index). Cette option peut être utilisée pour afficher les onglets utilisés dans les bases de données Excel. (Un noeud source Excel distinct est également disponible. Pour plus d'informations, reportez-vous à la section Noeud source Excel sur p. 53.)
- Cliquez sur Affichage pour afficher les tables virtuelles basées sur une requête impliquant des tables ordinaires.
- Cliquez sur Synonymes pour afficher les synonymes créés dans la base de données pour toute table existante.

Filtres Nom/Propriétaire. Ces champs vous permettent de filtrer la liste des tables affichées par nom ou propriétaire. Par exemple, saisissez SYS pour répertorier les tables de ce propriétaire uniquement. Pour les recherches basées sur des caractères génériques, un caractère de soulignement (_) peut être utilisé pour représenter un caractère ; un caractère pourcentage (%) peut correspondre à une séquence d'au moins zéro caractère.

Définir par défaut. Enregistre les paramètres actuels en tant que paramètres par défaut de l'utilisateur actuel. Ces paramètres seront restaurés ultérieurement lorsqu'un utilisateur ouvrira une nouvelle boîte de dialogue du sélecteur de table *pour les mêmes nom de source de données et connexion utilisateur uniquement*.

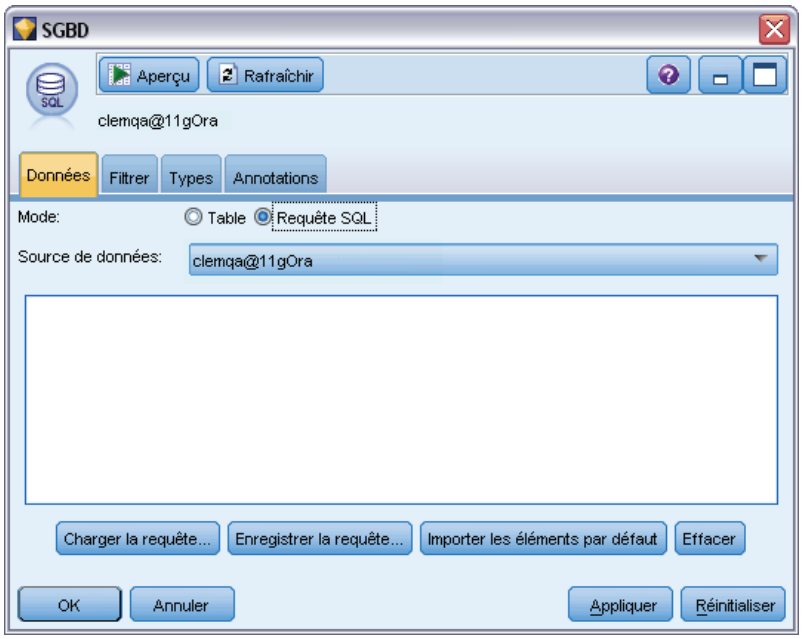
Interrogation de la base de données

Une fois connecté à une source de données, vous pouvez choisir d'importer des champs à l'aide de requêtes SQL. Dans la boîte de dialogue principale, sélectionnez Requête SQL comme mode de connexion. Cette opération ajoute une fenêtre d'éditeur de requêtes dans la boîte de dialogue. A l'aide de l'éditeur de requêtes, vous pouvez créer ou charger une ou plusieurs requêtes SQL dont les résultats seront lus dans le flux de données.

Si vous spécifiez plusieurs requêtes SQL, séparez-les par des points-virgules (;) et assurez-vous qu'il n'y a pas d'instruction SÉLECTIONNER multiple.

Pour annuler et fermer la fenêtre de l'éditeur de requêtes, sélectionnez Table comme mode de connexion.

Figure 2-9
Chargement de données à l'aide de requêtes SQL



Vous pouvez inclure des paramètres de flux SPSS Modeler (un type de variable définie par l'utilisateur) dans les requêtes SQL. Pour plus d'informations, reportez-vous à la section Utilisation des paramètres de flux dans une requête SQL sur p. 24.

Charger la requête. Cliquez sur cette option pour ouvrir le navigateur, que vous pouvez utiliser pour charger une requête précédemment enregistrée.

Enregistrer la requête. Cliquez sur cette option pour ouvrir la boîte de dialogue Enregistrer la requête, que vous pouvez utiliser pour enregistrer la requête en cours.

Importer les éléments par défaut. Cliquez sur cette option pour importer un exemple d'instruction SQL SELECT créé automatiquement à l'aide de la table et des colonnes sélectionnées dans la boîte de dialogue.

Effacer. Efface le contenu de la zone de travail. Utilisez cette option pour tout annuler et recommencer.

Utilisation des paramètres de flux dans une requête SQL

Lors de la rédaction d'une requête SQL pour l'importation de champs, vous pouvez inclure des paramètres de flux SPSS Modeler précédemment définis. Tous les types de paramètre de flux sont pris en charge.

Le tableau suivant présente quelques exemples d'interprétation de paramètres de flux dans les requêtes SQL.

Table 2-2
Exemples de paramètres de flux

Nom du paramètre de flux (exemple)	Stockage	Valeur du paramètre de flux	Interprété comme
PString	Chaîne	ss	'ss'
PInt	Entier	5	5
PReal	Réel	5.5	5.5
PTime	Temps	23:05:01	t{'23:05:01'}
PDate	Date	2011-03-02	d{'2011-03-02'}
PTimeStamp	Horodatage	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Inconnu	IntValue	IntValue

Dans la requête SQL, vous spécifiez un paramètre de flux de la même manière qu'une expression CLEM, à savoir par '$P-<parameter_name>', où <parameter_name> est le nom qui a été défini pour le paramètre de flux.

Lors du référencement d'un champ, le type de stockage doit être défini comme Unknown, et la valeur du paramètre doit être entre guillemets, le cas échéant. Par exemple, à l'aide des exemples du tableau, si vous saisissez la requête SQL suivante :

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

elle sera interprétée comme :

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

Si vous référencez le champ IntValue à l'aide du paramètre PColumn, vous devrez spécifier la requête comme suit pour obtenir le même résultat :

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

Noeud Délimité

Vous pouvez utiliser des noeuds Délimité pour lire les données de fichiers texte de longueur variable (fichiers dont les enregistrements contiennent un nombre fixe de champs et un nombre variable de caractères), aussi appelés fichiers texte délimités. Ce type de nœud est également utile pour les fichiers contenant des textes d'en-tête de longueur fixe et certains types d'annotation. Les enregistrement sont lus un à la fois et transmis via le flux jusqu'à la lecture de la totalité du fichier.

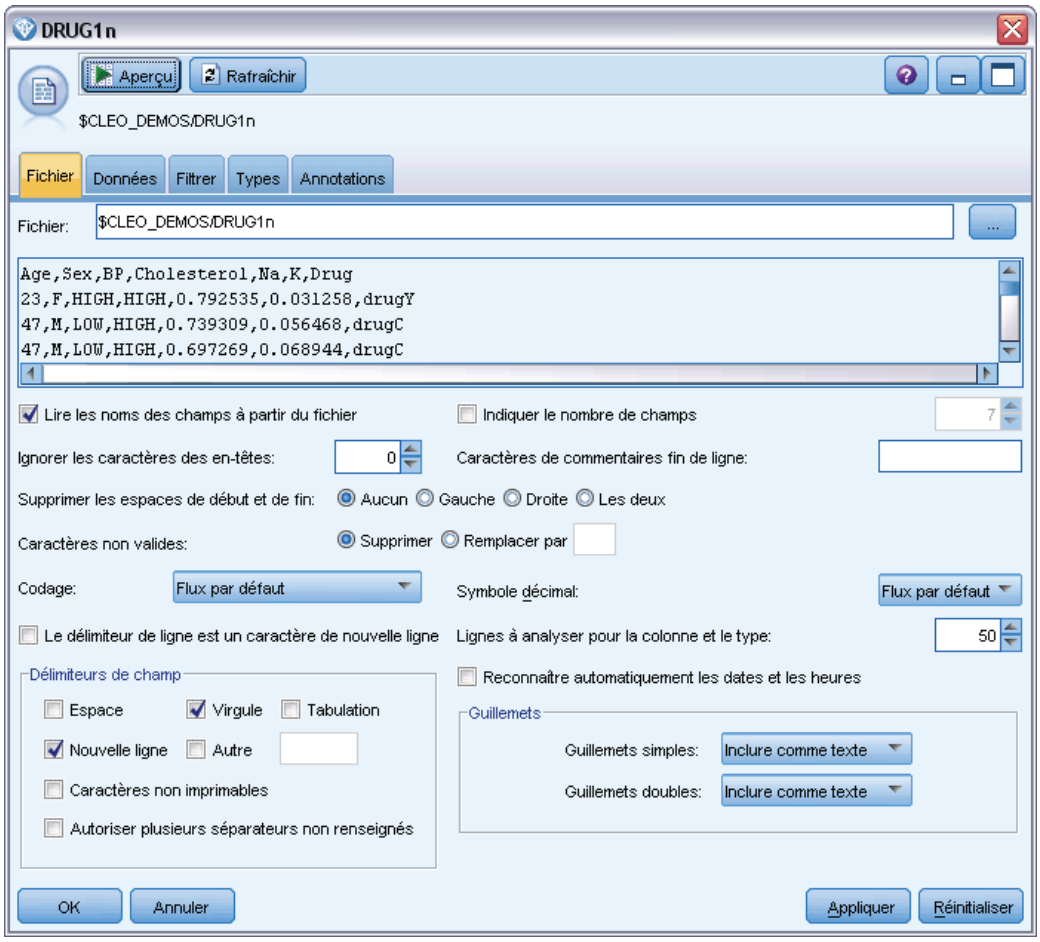
Remarques pour la lecture dans les données de texte délimitées

- Les enregistrements doivent être délimités par un caractère nouvelle ligne à la fin de chaque ligne. Le caractère nouvelle ligne ne doit pas être utilisé à une autre fin (par exemple, dans un nom ou une valeur de champ). Les espaces situés en début et en fin doivent être supprimés pour économiser de l'espace. Ceci n'est cependant pas crucial. En option, le noeud peut effectuer cette opération.

- Les champs doivent être délimités par une virgule ou un autre caractère utilisé uniquement comme séparateur. Ceci indique qu'il ne figure pas dans les noms ou les valeurs de champ. Si ceci n'est pas possible, tous les champs de texte peuvent être mis entre guillemets doubles, à condition qu'aucun des noms et valeurs de champ ne contienne de guillemet double. Si les noms ou les valeurs de champ contiennent des guillemets doubles, les champs de texte peuvent être mis à la place entre guillemets simples, de nouveau à condition que ces derniers ne soient pas utilisés ailleurs dans les valeurs. Si vous ne pouvez utiliser ni les guillemets simples ni les guillemets doubles, les valeurs de texte doivent être modifiées pour supprimer ou remplacer le caractère séparateur, ou les guillemets simples ou doubles.
- Chaque ligne, dont celle d'en-tête, doit contenir le même nombre de champs.
- La première ligne doit contenir les noms de champ. Si ce n'est pas le cas, désélectionnez Lire les noms des champs à partir du fichier pour donner à chaque champ un nom générique tel que *Champ1, Champ2*, etc.
- La deuxième ligne doit contenir le premier enregistrement de données. Il ne doit y avoir ni ligne vierge ni commentaire.
- Les valeurs numériques ne doivent pas inclure le séparateur des milliers ou le symbole de groupement—sans la virgule dans 3,000.00, par exemple. L'indicateur décimal (point aux E-U ou au RU) ne doit être utilisé que dans les cas appropriés.
- Les valeurs date et heure doivent être dans l'un des formats reconnus dans la boîte de dialogue Options de flux, comme DD/MM/YYYY ou HH:MM:SS. Tous les champs date et heure dans le fichier doivent respecter le même format. En outre, tout champ contenant une date doit utiliser le même format pour toutes les valeurs dans ce champ.

Définition des options du nœud Délimité

Figure 2-10
Boîte de dialogue nœud Délimité



Fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (...) pour sélectionner un fichier. Le chemin d'accès au fichier apparaît une fois que vous avez sélectionné un fichier. Son contenu est affiché avec des séparateurs dans le panneau situé en dessous.

Vous pouvez copier et coller le texte d'exemple affiché à partir de votre source de données à l'aide des commandes suivantes : Caractères de commentaires de fin de ligne et séparateurs spécifiés par l'utilisateur. Utilisez les raccourcis clavier Ctrl+C et Ctrl+V pour effectuer le copier-coller.

Lire les noms des champs à partir du fichier. Sélectionnée par défaut, cette option traite la première ligne du fichier de données en tant qu'étiquettes pour la colonne. Si la première ligne n'est pas un en-tête, désélectionnez l'option pour attribuer automatiquement à chaque champ de l'ensemble de données un nom générique, comme *Champ1, Champ2*.

Indiquer le nombre de champs. Indiquez le nombre de champs dans chaque enregistrement. Le nombre de champs peut être détecté automatiquementsi les enregistrements se terminent par des caractères de nouvelle ligne. Vous pouvez également entrer directement un nombre.

Ignorer les caractères des en-têtes. Indiquez le nombre de caractères à ignorer au début du premier enregistrement.

Caractères de commentaires fin de ligne. Spécifiez quels caractères (comme # ou !) indiquent des annotations dans les données. Lorsque l'un de ces caractères apparaît dans le fichier, toutes les données situées entre ce caractère et le caractère de nouvelle ligne suivant (non inclus) sont ignorées.

Supprimer les espaces de début et de fin. Sélectionnez les options permettant la suppression des espaces situés en début et en fin des chaînes lors de l'importation.

Remarque. Les comparaisons entre les chaînes qui utilisent ou non les répercussions SQL peuvent produire des résultats différents lorsqu'il existe des espaces en fin de chaîne.

Caractères non valides. Sélectionnez Supprimer pour supprimer les caractères non valides de la source de données. Sélectionnez Remplacer par pour remplacer les caractères non valides par le symbole indiqué (un caractère uniquement). Les caractères non valides sont des caractères nuls (0) ou des caractères qui n'existent pas dans la méthode de codage spécifiée.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal. Sélectionnez le type de séparateur décimal utilisé dans votre source de données. Le flux par défaut est le caractère sélectionné dans l'onglet Options de la boîte de dialogue des propriétés du flux. Sinon, sélectionnez Point (.) ou Virgule (,) pour lire toutes les données de cette boîte de dialogue à l'aide du caractère choisi comme séparateur décimal.

Le séparateur de ligne est un caractère nouvelle ligne. Pour utiliser le caractère nouvelle ligne comme séparateur de ligne, à la place d'un séparateur de champ, vous devez sélectionner cette option. Par exemple, ceci peut être utile si une ligne est coupée du fait que le nombre de séparateurs sur cette ligne est impair. Remarque : si vous sélectionnez cette option, vous ne pouvez pas sélectionner Nouvelle ligne dans la liste des séparateurs.

Séparateurs. A l'aide des cases à cocher répertoriées pour cette commande, vous pouvez spécifier quels caractères (comme la virgule) marquent les limites des champs dans le fichier. Vous pouvez indiquer plusieurs séparateurs(, | par exemple) pour les enregistrements qui font appel à des séparateurs multiples. Le séparateur par défaut est la virgule.

Remarque : si la virgule est également définie en tant que séparateur décimal, les paramètres par défaut fournis ne fonctionnent pas. Si la virgule sert à la fois de séparateur de champs et de séparateur décimal, sélectionnez Autre dans la liste Séparateurs. Ensuite, ajoutez manuellement une virgule dans le champ de saisie.

Sélectionnez Autoriser plusieurs séparateurs non renseignés pour considérer plusieurs séparateurs non renseignés adjacents comme un séparateur unique. Par exemple, une séquence constituée d'une valeur de données suivie de quatre espaces, puis d'une autre valeur de données, sera considérée comme séquence à deux champs, et non comme une séquence à cinq champs.

Lignes à analyser pour la colonne et le type. Spécifiez le nombre de lignes et de colonnes à analyser pour les types de données spécifiés.

Reconnaître automatiquement les dates et les heures. Pour permettre à IBM® SPSS® Modeler d'essayer de reconnaître automatiquement des entrées de date ou d'heures, cochez cette case. Par exemple, cela signifie qu'une entrée telle que 07-11-1965 sera identifiée comme une date et que 02:35:58 sera identifié comme une heure. Cependant, des entrées telles que 07111965 ou 023558 seront affichées sous la forme d'un entier car il n'y a pas de séparateurs entre les nombres.

Remarque : Pour éviter de rencontrer des problèmes relatifs aux données lors de l'utilisation de fichiers de données provenant de versions précédentes de SPSS Modeler, cette case n'est pas cochée par défaut pour les informations enregistrées dans les versions antérieures à la version 13.

Guillemets. A l'aide des listes déroulantes, vous pouvez indiquer la façon dont les guillemets simples et doubles sont traités lors de l'importation. Vous pouvez choisir l'option Supprimer (supprime tous les guillemets), Inclure comme texte (inclut les guillemets dans la valeur du champ) ou Apparier et supprimer (supprime des paires de guillemets). Si un guillemet n'est pas apparié, un message d'erreur apparaît. Si vous sélectionnez Supprimer ou Apparier et supprimer, la valeur du champ (sans les guillemets) est stockée sous forme de chaîne.

A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur Rafraîchir pour recharger les champs à partir de la source de données. Ceci est utile lorsque vous modifiez des connexions des données au nœud source ou lorsque vous utilisez les différents onglets de la boîte de dialogue.

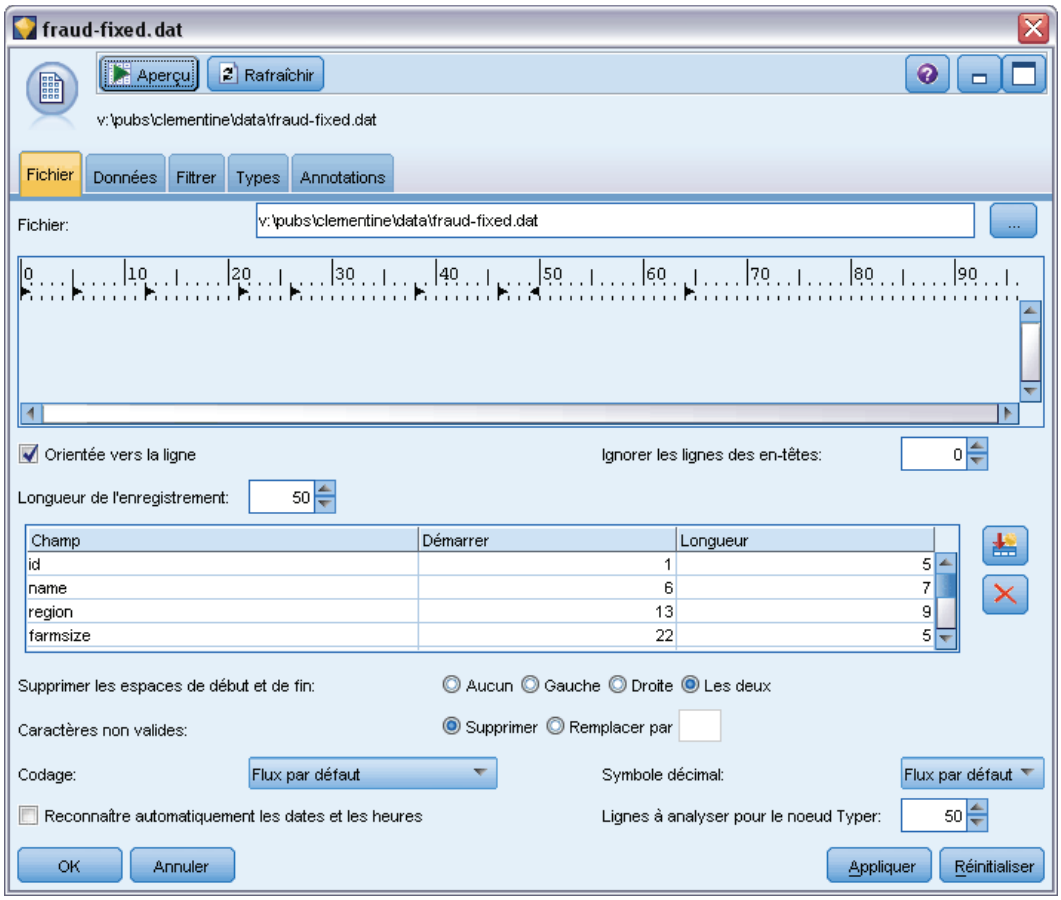
Noeud Fixe

Vous pouvez utiliser des noeuds Fixe pour importer les données de fichiers texte de longueur fixe (fichiers dont les champs ne sont pas délimités, mais qui commencent au même endroit et sont de longueur fixe). Les données générées automatiquement ou héritées sont souvent stockées au format de longueur fixe. Grâce à l'onglet Fichier du nœud Fixe, vous pouvez facilement indiquer la position et la longueur des colonnes de vos données.

Définition des options du nœud Fixe

L'onglet Fichier du noeud Fixe vous permet d'importer des données dans IBM® SPSS® Modeler et de spécifier la position des colonnes et la longueur des enregistrements. Vous pouvez cliquer dans le panneau d'aperçu des données situé au centre de la boîte de dialogue pour ajouter des flèches indiquant les points d'arrêt entre les champs.

Figure 2-11
Spécification de colonnes dans des données de longueur fixe



Fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (...) pour sélectionner un fichier. Une fois que vous avez sélectionné un fichier, son chemin d'accès apparaît. Son contenu est affiché avec des séparateurs dans le panneau situé en dessous.

Vous pouvez utiliser le panneau d'aperçu des données pour spécifier la position et la longueur des colonnes. La règle située en haut de la fenêtre d'aperçu vous permet de mesurer la longueur des variables et de spécifier le point d'arrêt entre elles. Vous pouvez spécifier des lignes de points d'arrêt en cliquant dans la zone de la règle au-dessus des champs. Pour déplacer les points d'arrêt, faites-les glisser. Pour les supprimer, faites-les glisser hors de la zone d'aperçu des données.

- Chaque ligne de points d'arrêt ajoute automatiquement un nouveau champ à ceux du tableau situé en dessous.
- Les positions de départ indiquées par les flèches sont automatiquement ajoutées à la colonne Démarrer du tableau situé en dessous.

Orientée vers la ligne. Indiquez si vous souhaitez ignorer le caractère de nouvelle ligne à la fin de chaque enregistrement.

Ignorer les lignes des en-têtes. Indiquez le nombre de lignes à ignorer au début du premier enregistrement. Cette option est utile si vous souhaitez ignorer les en-têtes de colonne.

Longueur de l'enregistrement. Indiquez le nombre de caractères dans chaque enregistrement.

Champ. Tous les champs que vous avez définis pour ce fichier de données sont répertoriés ici. Vous pouvez définir des champs de deux manières :

- Spécifier les champs de façon interactive à l'aide du panneau d'aperçu des données situé au-dessus.
- Spécifier les champs manuellement en ajoutant des lignes de champs vides dans le tableau en dessous. Cliquer sur le bouton situé à droite du panneau des champs pour ajouter de nouveaux champs. Entrez ensuite dans le champ vide un nom de champ, une position de départ et une longueur. Ces options ajoutent automatiquement des flèches au panneau d'aperçu des données, que vous pouvez ajuster très facilement.

Pour supprimer un champ défini précédemment, sélectionnez-le dans la liste et cliquez sur le bouton de suppression rouge.

Démarrer. Indiquez la position du premier caractère dans le champ. Par exemple, dans le cas d'un enregistrement dont le second champ commence au seizième caractère, vous devez indiquer la valeur 16.

Longueur. Indiquez le nombre de caractères contenus dans la valeur la plus longue de chaque champ. Ceci permet de déterminer le point de césure pour le champ suivant.

Supprimer les espaces de début et de fin. Cochez cette case pour que les espaces situés en début et en fin des chaînes soient supprimés lors de l'importation.

Remarque. Les comparaisons entre les chaînes qui utilisent ou non les répercussions SQL peuvent produire des résultats différents lorsqu'il existe des espaces en fin de chaîne.

Caractères non valides. Sélectionnez Supprimer pour supprimer les caractères non valides de l'entrée de données. Sélectionnez Remplacer par pour remplacer les caractères non valides par le symbole indiqué (un caractère uniquement). Les caractères non valides sont des caractères nuls (0) ou des caractères qui n'existent pas dans le codage en cours.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal. Sélectionnez le type de séparateur décimal utilisé dans votre source de données. La valeur par défaut du flux est le caractère sélectionné dans l'onglet Options de la boîte de dialogue des propriétés du flux. Sinon, sélectionnez Point (.) ou Virgule (,) pour lire toutes les données de cette boîte de dialogue à l'aide du caractère choisi comme séparateur décimal.

Reconnaître automatiquement les dates et les heures. Pour permettre à SPSS Modeler d'essayer de reconnaître automatiquement des entrées de date ou d'heures, cochez cette case. Par exemple, cela signifie qu'une entrée telle que 07-11-1965 sera identifiée comme une date et que 02:35:58 sera identifié comme une heure. Cependant, des entrées telles que 07111965 ou 023558 seront affichées sous la forme d'un entier car il n'y a pas de séparateurs entre les nombres.

Remarque : Pour éviter de rencontrer des problèmes relatifs aux données lors de l'utilisation de fichiers de données provenant de versions précédentes de SPSS Modeler, cette case n'est pas cochée par défaut pour les informations enregistrées dans les versions antérieures à la version 13.

Lignes à analyser pour le nœud Typer. Indiquez le nombre de lignes à traiter pour les types de données indiqués.

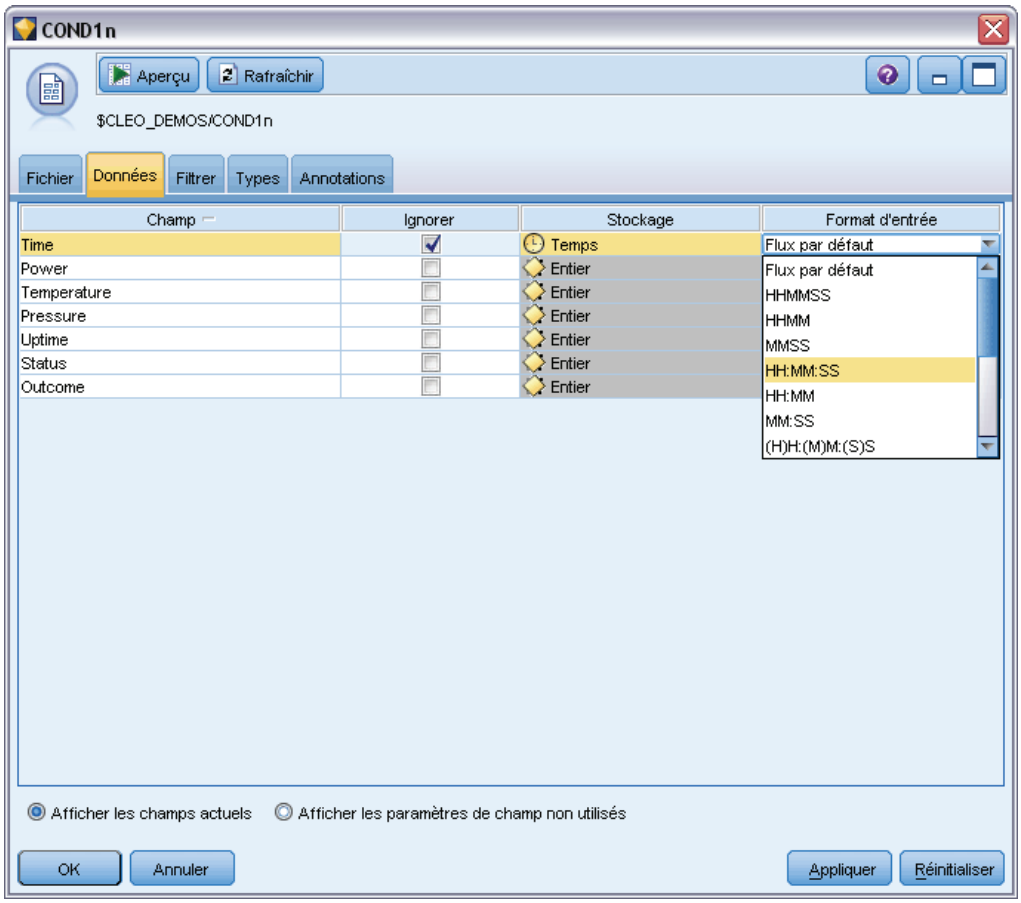
A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur Rafraîchir pour recharger les champs à partir de la source de données. Ceci est utile lors de la modification des connexions des données au nœud source ou lors de l'utilisation des différents onglets de la boîte de dialogue.

Définition du stockage et du formatage des champs

Les options de l'onglet Données des noeuds Fixe, Délimité, Source XML et Utilisateur vous permettent d'indiquer le type de stockage des champs à mesure qu'ils sont importés ou créés dans IBM® SPSS® Modeler. Pour les noeuds Fixe, Délimité et Utilisateur, vous pouvez aussi indiquer le formatage des champs et d'autres métadonnées.

Dans le cas de données lues à partir d'autres sources, le stockage est déterminé automatiquement ; vous pouvez toutefois le modifier par le biais d'une fonction de conversion, telle que to_integer, appliquée dans un noeud Remplacer ou Calculer.

Figure 2-12
Modification du type de stockage et du formatage des champs lors de l'importation



Champ. Utilisez la colonne *Champ* pour afficher et sélectionner des champs dans l'ensemble de données actuel.

Ignorer. Cochez la case de la colonne *Ignorer* pour activer les options des colonnes *Stockage* et *Format d'entrée*.

Stockage des données

Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Par exemple, vous pouvez définir le niveau de mesure d'un champ de nombre entier comportant les valeurs 1 et 0 comme étant un champ *Booléen*. En général, 1 correspond à la valeur *True (vrai)* et 0 à la valeur *False (faux)*. Alors que le stockage doit être déterminé au niveau de la source, le niveau de mesure peut être modifié à l'aide d'un noeud Typer en tout point du flux. Pour plus d'informations, reportez-vous à la section Niveaux de mesure dans le chapitre 4 sur p. 138.

Les types de stockage disponibles sont les suivants :

- **Chaîne.** Utilisé pour les champs contenant des données non numériques, également appelées données alphanumériques. Une chaîne peut inclure n'importe quelle séquence de caractères, telle que *fred*, *Classe 2* ou *1234*. Notez que les nombres utilisés dans les chaînes ne peuvent pas être inclus dans les calculs.
- **Entier.** Champ dont les valeurs sont des entiers.
- **Réel.** Il s'agit de nombres pouvant comporter des décimales (pas uniquement des entiers). Le format d'affichage est indiqué dans la boîte de dialogue Propriétés de flux et peut être ignoré pour des champs individuels dans un noeud Typer (onglet Format).
- **Date.** Valeurs de date indiquées dans un format standard, comme année, mois et jour (par exemple, 26.09.07). Le format exact est indiqué dans la boîte de dialogue Propriétés de flux.
- **Heure :** Valeur indiquant une durée. Par exemple, un appel de service ayant duré 1 heure, 26 minutes et 38 secondes peut être représenté sous la forme 01:26:38, en fonction du format d'heure actuel indiqué dans la boîte de dialogue Propriétés de flux.
- **Horodatage.** Valeurs comportant à la fois un composant de date et d'heure, par exemple 2007–09–26 09:04:00, dépendant une fois de plus des formats de date et d'heure dans la boîte de dialogue Propriétés de flux. Remarque : il se peut que les valeurs d'horodatage doivent être placées entre guillemets doubles pour être interprétées comme une valeur unique et non comme des valeurs date et heure distinctes. (Cela s'applique, par exemple, lors de la saisie de valeurs dans un noeud Utilisateur).

Conversion de stockages. Vous pouvez également convertir le stockage d'un champ à l'aide de diverses fonctions de conversion, comme to_string et to_integer dans un noeud Remplacer. Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer dans le chapitre 4 sur p. 180. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue des propriétés du flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez MOIS AAAA comme format de date par défaut pour le flux. Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne contenant des valeurs catégorielles. Pour plus d'informations, reportez-vous à la section Recodage des valeurs à l'aide du noeud Calculer dans le chapitre 4 sur p. 177.

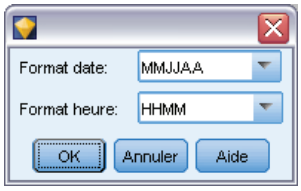
Lecture de données mixtes. Au cours de la lecture des champs de stockage numérique (entier, nombre réel, heure, horodatage ou date), toutes les valeurs non numériques sont définies comme étant nulles ou manquantes dans le système. En effet, contrairement à certaines applications, SPSS Modeler n'autorise pas les types de stockage mixtes au sein d'un champ. Pour éviter ce type de problème, faites en sorte que les champs comportant des données mixtes soient lus en tant que chaînes ; pour cela, modifiez le type de stockage dans le noeud source ou dans l'application externe.

Format d'entrée des champs (noeuds Fixe, Délimité ou Utilisateur uniquement).

Pour tous les types de stockage, à l'exception de Chaîne et Entier, vous pouvez choisir dans la liste déroulante les options de formatage du champ sélectionné. Par exemple, lorsque vous fusionnez les données de plusieurs paramètres régionaux, vous pouvez être amené à utiliser un point (.) comme séparateur décimal pour un champ, alors qu'un autre champ aura une virgule pour séparateur.

Les options d'entrée indiquées dans le noeud source remplacent les options de formatage définies dans la boîte de dialogue des propriétés de flux ; elles n'apparaissent toutefois pas ultérieurement dans le flux. Ces options visent à analyser correctement les entrées fournies en fonction de votre connaissance des données. Les formats spécifiés servent de point de repère à l'analyse de ces données lorsqu'elles sont lues dans SPSS Modeler, et non à déterminer la manière dont elles doivent être formatées après leur lecture dans SPSS Modeler. Pour indiquer le formatage de chaque champ au sein du flux, utilisez l'onglet Format d'un nœud Typer. Pour plus d'informations, reportez-vous à la section Onglet Paramètres du champ dans le chapitre 4 sur p. 153.

Figure 2-13
Spécification des formats de date et d'heure des champs d'horodatage



Les options varient selon le type de stockage utilisé. Par exemple, pour le type de stockage Réel, vous pouvez sélectionner le séparateur décimal Point (.) ou Virgule (,). Pour les champs d'horodatage, une autre boîte de dialogue s'ouvre lorsque vous choisissez Spécifier dans la liste déroulante. Pour plus d'informations, reportez-vous à la section Paramétrage des options de formatage des champs dans le chapitre 4 sur p. 154.

Pour tous les types de stockage, vous pouvez également sélectionner Flux par défaut afin d'utiliser les paramètres par défaut du flux pour l'importation. Les paramètres de flux sont répertoriés dans la boîte de dialogue des propriétés du flux.

Options supplémentaires

Vous pouvez utiliser d'autres options à l'aide de l'onglet Données :

- Pour afficher les paramètres de stockage pour les données qui ne sont plus connectées via le nœud en cours (données d'apprentissage, par exemple), sélectionnez Afficher les paramètres de champ non utilisés. Vous pouvez effacer les champs hérités en cliquant sur Effacer.
- A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur Rafraîchir pour recharger les champs à partir de la source de données. Ceci est utile lorsque vous modifiez des connexions des données au nœud source ou lorsque vous utilisez les différents onglets de la boîte de dialogue.

Noeud Data Collection

Les données d'enquêtes d'importation de noeuds source Data Collection basées sur IBM® SPSS® Data Collection Survey Reporter Developer Kit utilisé par un logiciel d'études de marché de IBM Corp. Ce format distingue les **données d'observation**— (les réponses réelles aux questions rassemblées au cours d'une enquête) des **métadonnées**— qui décrivent la façon dont les données d'observation sont collectées et organisées. Les métadonnées consistent en des informations diverses : texte des questions, nom et description de variables, définitions de variables à réponses multiples, traduction de chaînes de texte et définition de la structure des données d'observation.

Remarque : Ce noeud requiert Data Collection Survey Reporter Developer Kit, fourni avec les logiciels Data Collection de IBM Corp.. Pour de plus amples informations, reportez-vous à la page Web de Data Collection à l'adresse *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*. Mise à part l'installation de Developer Kit, aucune configuration supplémentaire n'est requise.

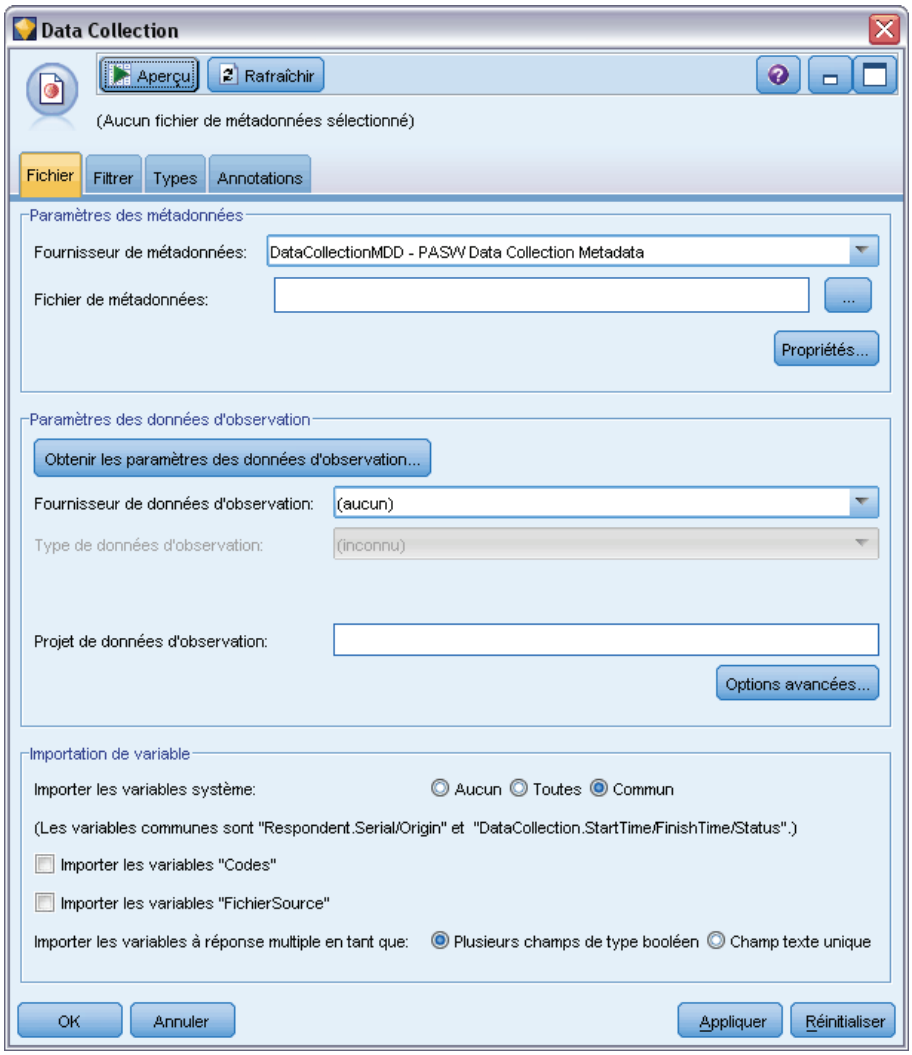
Commentaires

- Les données d'enquête sont lues à partir du format VDATA sous forme de tableau uniforme, ou à partir de sources au format HDATA hiérarchique; si elles incluent une source de métadonnées (requiert Data Collection 4.5 ou version ultérieure).
- Les types sont instanciés automatiquement à partir des informations contenues dans les métadonnées.
- Lorsque des données d'enquête sont importées dans IBM® SPSS® Modeler, les questions sont affichées sous forme de champs et comportent un enregistrement par personne sondée.

Options de fichier d'importation Data Collection

Dans l'onglet Fichier du noeud Data Collection, vous pouvez définir des options pour les métadonnées et les données d'observation à importer.

Figure 2-14
Options de fichier du noeud source Data Collection



Paramètres des métadonnées

Remarque : Afin de visualiser la liste complète des types de fichiers de fournisseurs disponibles, vous devez installer IBM® SPSS® Data Collection Survey Reporter Developer Kit, disponible avec le logiciel Data Collection. Pour plus d'informations, consultez la page Web Data Collection à l'adresse *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*.

Fournisseur de métadonnées. Le logiciel Data Collection Survey Reporter Developer Kit prend en charge un certain nombre de formats d'importation des données d'enquête. Les principaux types de fournisseur disponibles sont les suivants :

- DataCollectionMDD. Lit les métadonnées à partir d'un fichier de définition de questionnaire (*.mdd*). Il s'agit du format de modèle de données Data Collection standard.

- Base de données ADO. Lit les données d'observation et les métadonnées à partir de fichiers ADO. Indiquez le nom et l'emplacement du fichier *.adoinfo* contenant les métadonnées. Le nom interne de ce fichier DSC est *mrADODsc*.
- Base de données In2data. Lit les données d'observation et les métadonnées In2data. Le nom interne de ce fichier DSC est *mrI2dDsc*.
- Fichier journal de collecte des données. Lit les métadonnées issues d'un fichier journal Data Collection standard. Les fichiers journaux sont généralement dotés d'une extension *.tmp*. Toutefois, certains fichiers peuvent comporter une autre extension. Si nécessaire, vous pouvez renommer le fichier et lui attribuer l'extension *.tmp*. Le nom interne de ce fichier DSC est *mrLogDsc*.
- Fichier de définitions Quancept. Convertit les métadonnées en script Quancept. Indiquez le nom du fichier Quancept *.qdi*. Le nom interne de ce fichier DSC est *mrQdiDrsDsc*.
- Base de données Quanvert. Lit les données d'observation et les métadonnées Quanvert. Précisez le nom et l'emplacement du fichier *.qvinfo* ou *.pkd*. Le nom interne de ce fichier DSC est *mrQvDsc*.
- Base de données de participation à la collecte de données. Lit les tables exemple et les tables d'historique d'un projet et crée des variables catégorielles calculées correspondant aux colonnes de ces tables. Le nom interne de ce fichier DSC est *mrSampleReportingMDSC*.
- Fichier Statistics. Lit les données d'observation et les métadonnées issues d'un fichier IBM® SPSS® Statistics*.sav*. Ecrit les données d'observation dans un fichier SPSS Statistics*.sav* en vue de leur analyse dans SPSS Statistics. Ecrit les métadonnées provenant d'un fichier SPSS Statistics*.sav* dans un fichier *.mdd*. Le nom interne de ce fichier DSC est *mrSavDsc*.
- Fichier SurveyCraft. Lit les données d'observation et les métadonnées SurveyCraft. Indiquez le nom du fichier *.vq* SurveyCraft. Le nom interne de ce fichier DSC est *mrSCDsc*.
- Fichier de script de collecte des données. Lit les métadonnées contenues dans un fichier mrScriptMetadata. Ces fichiers sont généralement dotés de l'extension *.mdd* ou *.dms*. Le nom interne de ce fichier DSC est *mrScriptMDSC*.
- Fichier XML Triple-S. Lit les métadonnées à partir d'un fichier Triple-S au format XML. Le nom interne de ce fichier DSC est *mrTripleSDsc*.

Propriétés des métadonnées. Vous pouvez également sélectionner Propriétés pour préciser la version de l'enquête à importer, ainsi que la langue, le contexte et le type d'étiquette à utiliser. Pour plus d'informations, reportez-vous à la section Propriétés relatives aux métadonnées d'importation IBM SPSS Data Collection sur p. 40.

Paramètres des données d'observation

Remarque : Afin de visualiser la liste complète des types de fichiers de fournisseurs disponibles, vous devez installer Data Collection Survey Reporter Developer Kit, disponible avec le logiciel Data Collection. Pour plus d'informations, consultez la page Web Data Collection à l'adresse *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*.

Obtenir les paramètres des données d'observation. Lorsque le système ne lit que des métadonnées issues de fichiers *.mdd*, cliquez sur Obtenir les paramètres des données d'observation pour déterminer les sources de données d'observation associées aux métadonnées sélectionnées, ainsi

que les paramètres nécessaires pour accéder à une source donnée. Cette option est disponible uniquement pour les fichiers *.mdd*.

Fournisseur de données d'observation. Les types de fournisseur suivants sont pris en charge :

- Base de données ADO. Lit les données d'observation par le biais de l'interface Microsoft ADO. Sélectionnez UDL OLE-DB comme type de donnée d'observation et indiquez une chaîne de connexion dans le champ UDL de données d'observation. Pour plus d'informations, reportez-vous à la section Chaîne de connexion de base de données sur p. 41. Le nom interne de ce composant est *mrADODsc*.
- Fichier texte délimité (Excel). Lit des données d'observation à partir d'un fichier délimité par des virgules (.CSV), tel qu'il peut être sorti par Excel. Le nom interne est *mrCsvDsc*.
- Fichier de données de collecte des données. Lit les données d'observation à partir d'un fichier de format de données Data Collection natif (Data Collection 4.5 et versions supérieures). Le nom interne est *mrDataFileDsc*.
- Base de données In2data. Lit les données d'observation et les métadonnées à partir d'un fichier de base de données In2data (*.i2d*). Le nom interne correspondant est *mrI2dDsc*.
- Fichier journal de collecte des données. Lit les données d'observation à partir d'un fichier journal Data Collection standard. Les fichiers journaux sont généralement dotés d'une extension *.tmp*. Toutefois, certains fichiers peuvent comporter une autre extension. Si nécessaire, vous pouvez renommer le fichier et lui attribuer l'extension *.tmp*. Le nom interne correspondant est *mrLogDsc*.
- Fichier de données Quantum. Lit les données d'observation provenant d'un fichier ASCII Quantum (*.dat*). Le nom interne correspondant est *mrPunchDsc*.
- Fichier de données Quancept. Lit les données d'observation issues d'un fichier *.drs*, *.drz* ou *.dru* Quancept. Le nom interne correspondante est *mrQdiDrsDsc*.
- Base de données Quanvert. Lit les données d'observation à partir d'un fichier *qvinfo* ou *.pkd* Quanvert. Le nom interne correspondant est *mrQvDsc*.
- Base de données de collecte des données (MS SQL Server). Lit les données d'observation pour une base de données relationnelle Microsoft SQL Server. Pour plus d'informations, reportez-vous à la section Chaîne de connexion de base de données sur p. 41. Le nom interne correspondante est *mrRdbDsc2*.
- Fichier Statistics. Lit les données d'observation et les métadonnées issues d'un fichier SPSS Statistics*.sav*. Le nom interne correspondant est *mrSavDsc*.
- Fichier SurveyCraft. Lit les données d'observation provenant d'un fichier *.qdt* SurveyCraft. Les fichiers *.vq* et *.qdt* doivent se trouver dans le même répertoire et être accessibles en lecture et en écriture. Cela n'est pas le cas lorsque ces fichiers sont créés par défaut à l'aide de SurveyCraft ; par conséquent, vous devez déplacer l'un des fichiers pour pouvoir importer ensuite des données SurveyCraft. Le nom interne correspondant est *mrScDsc*.
- Fichier de données Triple-S. Lit les données d'observation à partir d'un fichier de données Triple-S, au format délimité par des virgules ou de longueur fixe. Le nom interne est *mr TripleDsc*.
- Collecte de données XML. Lit les données d'observation issues d'un fichier de données XML Data Collection. Vous pouvez généralement utiliser ce format pour transférer des données d'observation d'un emplacement vers un autre. Le nom interne correspondant est *mrXmlDsc*.

Le type des données d'observation. Indique si les données d'observation sont lues à partir d'un fichier, d'un dossier, ou d'un type UDL OLE-DB ou DSN ODBC, et met à jour en conséquence les options de la boîte de dialogue. Les options valides dépendent du type de fournisseur. Pour les fournisseurs de base de données, vous pouvez définir les options de connexion OLE-DB ou ODBC. Pour plus d'informations, reportez-vous à la section Chaîne de connexion de base de données sur p. 41.

Projet de données d'observation. Lorsque vous lisez des données d'observation provenant d'une base de données Data Collection, vous pouvez fournir le nom du projet. Pour tous les autres types de données d'observation, ce paramètre doit rester vide.

Importation de variable

Importation de variables système. Indique si les variables système sont importées, y compris les variables qui indiquent l'état de l'entretien (en cours, terminé, date de fin, etc). Vous pouvez choisir Aucune, Toutes ou Communes.

Importation de variables «Codes». Contrôle l'importation de variables qui représentent des codes utilisés pour des réponses «Autre» ouvertes pour des variables catégorielles.

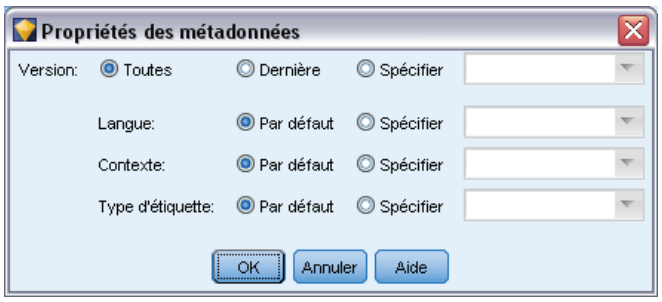
Importation de variables «SourceFile». Contrôle l'importation de variables qui contiennent les noms de fichiers d'images de réponses numérisées.

Importer des variables à réponses multiples sous la forme de. Des variables à réponses multiples peuvent être importées sous la forme de plusieurs champs booléens (un ensemble de dichotomies multiples), qui est la méthode par défaut pour les nouveaux flux. Les flux créés dans les versions de IBM® SPSS® Modeler antérieures à 12.0, des réponses multiples importées dans un champ unique, avec des valeurs séparées par des virgules. L'ancienne méthode reste prise en charge pour permettre l'exécution des flux existants. Toutefois, la mise à jour des anciens flux pour utiliser la nouvelle méthode est recommandée. Pour plus d'informations, reportez-vous à la section Importation des ensembles de réponses multiples sur p. 42.

Propriétés relatives aux métadonnées d'importation IBM SPSS Data Collection

Lorsque vous importez des données d'enquête IBM® SPSS® Data Collection, vous pouvez préciser la version de l'enquête à importer, ainsi que la langue, le contexte et le type d'étiquette à utiliser. Vous ne pouvez importer qu'une langue, un contexte et un type d'étiquette à la fois.

Figure 2-15
Propriétés relatives aux métadonnées d'importation IBM SPSS Data Collection



Version. Chaque version d'enquête peut être considérée comme un instantané des métadonnées utilisées pour collecter un ensemble précis de données d'observation. Au fil des modifications apportées à un questionnaire, plusieurs versions différentes peuvent être créées. Vous pouvez importer la dernière version, toutes les versions ou une version particulière.

- **Toutes.** Sélectionnez cette option si vous souhaitez utiliser une combinaison (ou sur-ensemble) de toutes les versions disponibles. (C'est ce que l'on nomme parfois supervision). En cas de conflit entre versions, les versions les plus récentes sont généralement prioritaires sur les versions plus anciennes. Ainsi, si une étiquette de catégorie diffère dans l'une des versions, c'est le texte de la version la plus récente qui est utilisé.
- **Version la plus récente.** Sélectionnez cette option pour n'utiliser que la version la plus récente.
- **Spécifier une version.** Sélectionnez cette option si vous souhaitez utiliser une version d'enquête particulière.

Sélectionner l'ensemble des versions s'avère utile, par exemple, lorsque vous souhaitez exporter des données d'observation à partir de plusieurs versions et que des modifications ont été apportées aux définitions des variables et des catégories (autrement dit, lorsque les données d'observation collectées dans une version ne sont pas valides dans une autre version). En sélectionnant toutes les versions pour lesquelles effectuer une exportation des données d'observation, vous pouvez généralement exporter simultanément les données souhaitées collectées dans les différentes versions, sans rencontrer d'erreurs de validité dues à des différences entre versions. Toutefois, selon les modifications apportées aux versions, certaines erreurs de validité risquent tout de même de se produire.

Langage : Les questions et le texte associé peuvent être stockés en plusieurs langues dans les métadonnées. Vous pouvez utiliser la langue par défaut de l'enquête ou indiquer une langue particulière. Si un élément n'est pas disponible dans la langue demandée, la langue par défaut est alors utilisée.

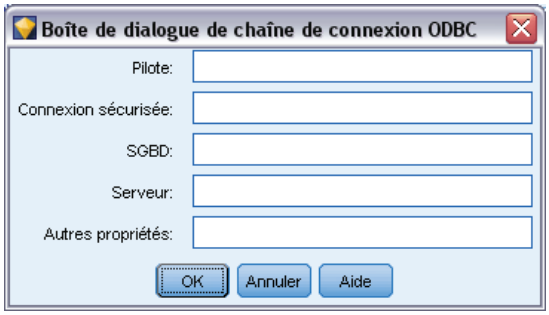
Contexte. Sélectionnez le contexte utilisateur souhaité. Il permet de déterminer les textes à afficher. Par exemple, sélectionnez Question pour afficher le texte des questions ou Analyse pour afficher une version abrégée des textes, mieux adaptée lors de l'analyse des données.

Type d'étiquette. Répertorie les types d'étiquette qui ont été définis. La valeur par défaut, Etiquette, est utilisée pour le texte des questions dans le contexte utilisateur Question et la description des variables dans le contexte utilisateur Analyse. Vous pouvez définir d'autres types d'étiquette pour les instructions, les descriptions, etc.

Chaîne de connexion de base de données

Lorsque vous utilisez le noeud IBM® SPSS® Data Collection pour importer des données d'observation à partir d'une base de données via une connexion OLE-DB ou ODBC, sélectionnez Edition dans l'onglet Fichier pour accéder à la boîte de dialogue de la chaîne de connexion et personnalisez la chaîne de connexion transmise au fournisseur afin d'affiner la connexion.

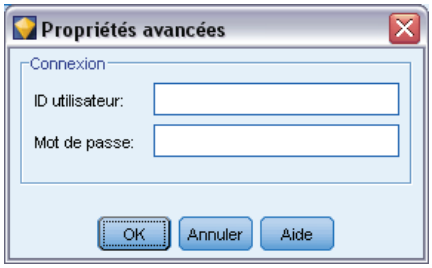
Figure 2-16
Chaîne Connexion d'importation IBM SPSS Data Collection



Propriétés avancées

Lorsque vous utilisez le noeud IBM® SPSS® Data Collection afin d'importer des données d'observation à partir d'une base de données qui requiert une connexion explicite, sélectionnez Options avancées pour fournir un ID utilisateur et un mot de passe permettant d'accéder à la source de données.

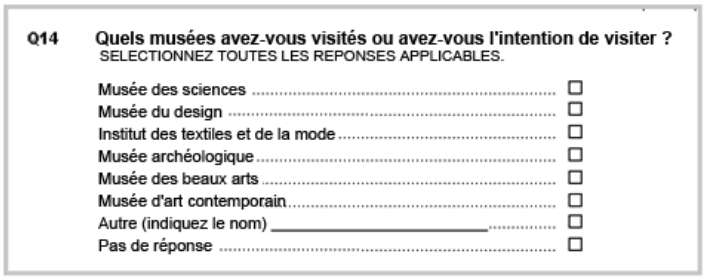
Figure 2-17
Propriétés d'importation avancées IBM SPSS Data Collection



Importation des ensembles de réponses multiples

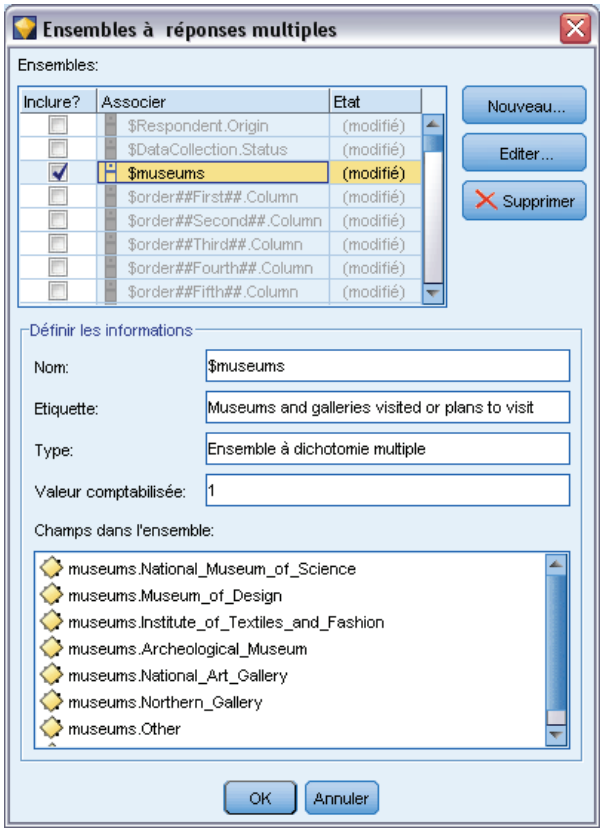
Les variables à réponses multiples peuvent être importées à partir de IBM® SPSS® Data Collection comme ensembles de plusieurs dichotomies, avec un champ booléen distinct pour chaque valeur possible de variable. Par exemple, s'il est demandé aux personnes interrogées de sélectionner les musées qu'ils ont visités dans une liste, l'ensemble inclut un champ booléen distinct pour chaque musée répertorié.

Figure 2-18
Question à réponses multiples



Une fois les données importées, vous pouvez ajouter ou modifier des ensembles de réponses multiples à partir d'un noeud quelconque qui inclut un onglet Filtrer. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples dans le chapitre 4 sur p. 159.

Figure 2-19
Boîte de dialogue Ensembles à réponses multiples



Importation de réponses multiples dans un champ unique (pour les flux créés dans les versions précédentes)

Dans les versions antérieures de IBM® SPSS® Modeler, au lieu d'importer des réponses multiples comme décrit ci-dessus, elles ont été importées dans un champ unique, avec des valeurs séparées par des virgules. Cette méthode reste acceptée afin de prendre en charge les flux existants. Toutefois, il est recommandé de mettre à jour ces flux pour utiliser la nouvelle méthode.

Remarques sur l'importation de colonnes IBM SPSS Data Collection

Les colonnes issues des données IBM® SPSS® Data Collection sont lues dans IBM® SPSS® Modeler de la manière indiquée dans le tableau suivant.

Type de colonne Data Collection	Stockage de SPSS Modeler	Niveau de mesure
Commutateur booléen (yes/no (oui/non))	Chaîne	Commutateur (valeurs 0 et 1)
Catégoriel	Chaîne	Nominal
Date ou horodatage	Horodatage	Continu
Double (valeur à virgule flottante comprise dans un intervalle défini)	Réel	Continu
Long (valeur entière comprise dans un intervalle défini)	Entier	Continu
Texte (description libre)	Chaîne	Sans type
Niveau (indique des grilles ou des boucles dans une question)	Ne s'applique pas aux données VDATA et n'est pas importé dans SPSS Modeler	
Objet (données binaires, comme la télécopie d'un texte griffonné ou un enregistrement sonore)	Pas importé dans SPSS Modeler	
Aucun (type inconnu)	Pas importé dans SPSS Modeler	
Colonne Respondent.Serial (associe un ID unique à chaque personne sondée)	Entier	Sans type

Pour éviter toute incohérence éventuelle entre les étiquettes de valeur lues dans les métadonnées et les valeurs réelles, toutes les valeurs des métadonnées sont converties en minuscules. Par exemple, l'étiquette de valeur *E1720_ans* est convertie en *e1720_ans*.

Noeud source IBM Cognos BI

Le noeud source IBM Cognos BI permet d'inclure des données de base de données Cognos BI ou des rapports de liste unique dans votre session de Data mining. Ainsi, vous pouvez combiner les fonctionnalités de veille économique de Cognos aux capacités d'analyses prédictives de IBM® SPSS® Modeler. Vous pouvez importer des données relationnelles, DMR (dimensionally-modeled relational) et OLAP.

À partir d'une connexion au serveur Cognos, commencez par sélectionner un emplacement à partir duquel importer les données ou les rapports. Un emplacement contient un modèle Cognos et tous les dossiers, requêtes, rapports, vues, raccourcis, URL et définitions de tâches associés à ce modèle. Un modèle Cognos définit les règles commerciales, les descriptions de données, les relations entre les données, les dimensions et les hiérarchies commerciales et d'autres tâches administratives.

Si vous importez des données, vous sélectionnez ensuite les objets que vous voulez importer depuis le package sélectionné. Parmi les objets que vous pouvez importer : les objets de requête (qui représentent les tables de la base de données) ou les éléments de requête individuels (qui représentent les colonnes de la table). Pour plus d'informations, reportez-vous à la section Icônes d'objet Cognos sur p. 45.

Si des filtres sont définis dans le package, vous pouvez importer un ou plusieurs d'entre eux. Si un des filtres que vous importez est associée à des données importées, ce filtre est appliqué avant l'importation des données. *Remarque* : Les données à importer doivent être au format UTF-8.

Si vous importez un rapport, vous sélectionnez un package, ou un dossier dans un package, contenant un ou plusieurs rapports. Vous sélectionnez ensuite le rapport individuel que vous voulez importer. *Remarque* : Seuls les rapports de liste unique peuvent être importés, les listes multiples ne sont pas prises en charge.

Si les paramètres ont été définis, soit pour un objet de données soit pour un rapport, vous pouvez spécifier les valeurs de ces paramètres avant d'importer l'objet ou le rapport.

Icônes d'objet Cognos

Les divers types d'objets pouvant être importés depuis une base de données Cognos BI sont représentés par différentes icônes, comme l'illustre le tableau suivant.

Table 2-3
Icônes d'objet Cognos

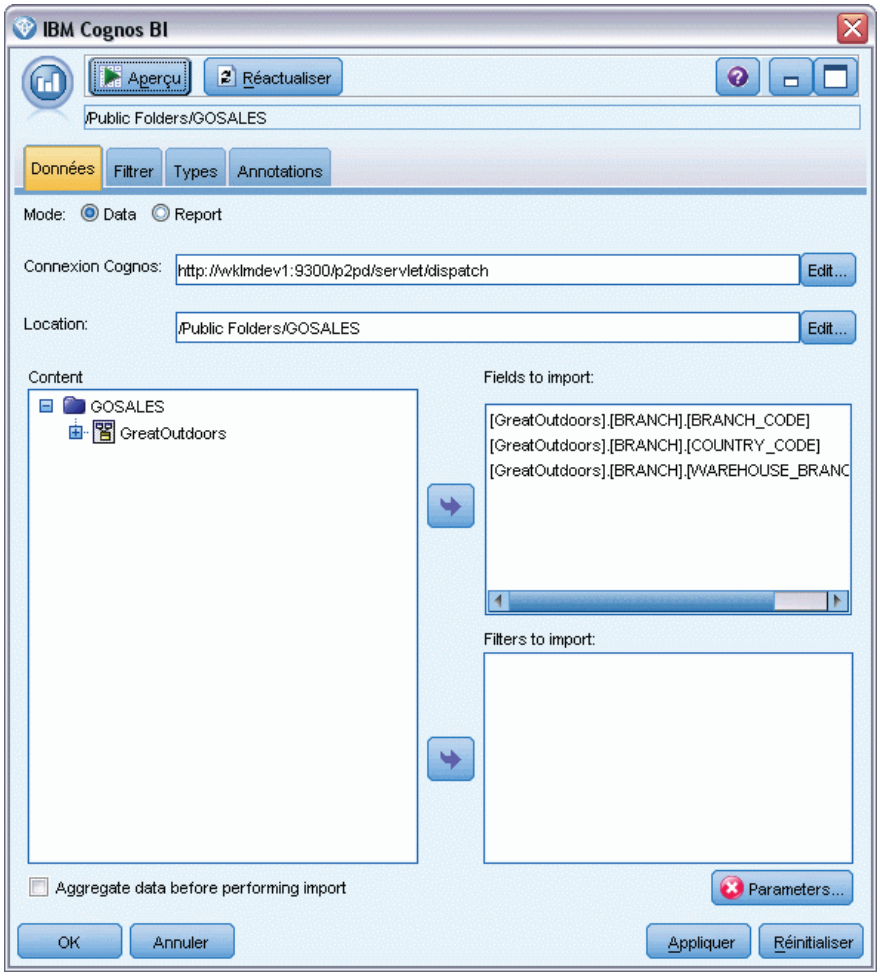
Icône	Objet
	Package
	Espace de nommage
	Objet de requête
	Élément de requête
	Dimension de mesure
	Mesure
	Dimension
	Hiérarchie de niveau
	Niveau

Icône	Objet
	Filtrer
	Rapport
	Calcul autonome

Importation des données Cognos

Pour importer des données d'une base de données Cognos BI IBM, assurez-vous que le Mode est réglé sur Données et remplissez la boîte de dialogue comme suit.

Figure 2-20
Importation des données Cognos



Connexion. Cliquez sur le bouton Modifier pour afficher une boîte de dialogue dans laquelle vous pourrez définir les détails d'une connexion Cognos à partir de laquelle importer les données ou les rapports. Si vous êtes déjà connecté à un serveur Cognos via IBM® SPSS® Modeler,

vous pouvez également modifier les détails de la connexion actuelle. Pour plus d'informations, reportez-vous à la section Connexions Cognos sur p. 49.

Emplacement. Lorsque la connexion au serveur Cognos est établie, cliquez sur le bouton Modifier à côté de ce champ pour afficher une liste des packages disponibles depuis lesquels importer le contenu. Pour plus d'informations, reportez-vous à la section Sélection de l'emplacement de Cognos sur p. 50.

Contenu. Affiche le nom du package sélectionné, avec les espaces de nommage associés au package. Double-cliquez sur un espace de nommage pour afficher les objets que vous pouvez importer. Les divers types d'objets sont représentés par différentes icônes. Pour plus d'informations, reportez-vous à la section Icônes d'objet Cognos sur p. 45.

Pour choisir un objet à importer, sélectionnez l'objet et cliquez sur la flèche supérieure des deux flèches droites pour déplacer l'objet dans le volet Champs à importer. La sélection d'un objet de requête importe tous ses éléments de requête. Double-cliquer sur un objet de requête le développe et vous pouvez ainsi choisir un ou plusieurs éléments de requête individuels. Vous pouvez effectuer plusieurs sélections avec Ctrl-clic (sélectionne des éléments individuels), Maj-clic (sélectionne un block d'éléments) et Ctrl-A (sélectionne tous les éléments).

Pour choisir un filtre à appliquer (si des filtres sont définis pour le package), accédez au filtre dans le volet Contenu, sélectionnez le filtre et cliquez sur la flèche inférieure des deux flèches de droite pour déplacer le filtre dans le volet Filtres à appliquer. Vous pouvez faire plusieurs choix avec Ctrl-clic (sélectionner des filtres individuels) et Shift-clic (sélectionner un bloc de filtres).

Champs à importer. Répertorie les objets de la base de données que vous avez choisi d'importer dans SPSS Modeler pour être traités. Si un des objets spécifiques n'est plus requis, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet Contenu. Vous pouvez effectuer plusieurs sélections de la même façon que pour le Contenu.

Filtres à appliquer. Répertorie les filtres que vous avez choisi d'appliquer aux données avant de les importer. Si l'un des filtres spécifiques n'est plus requis, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet Contenu. Vous pouvez effectuer plusieurs sélections de la même façon que pour le Contenu.

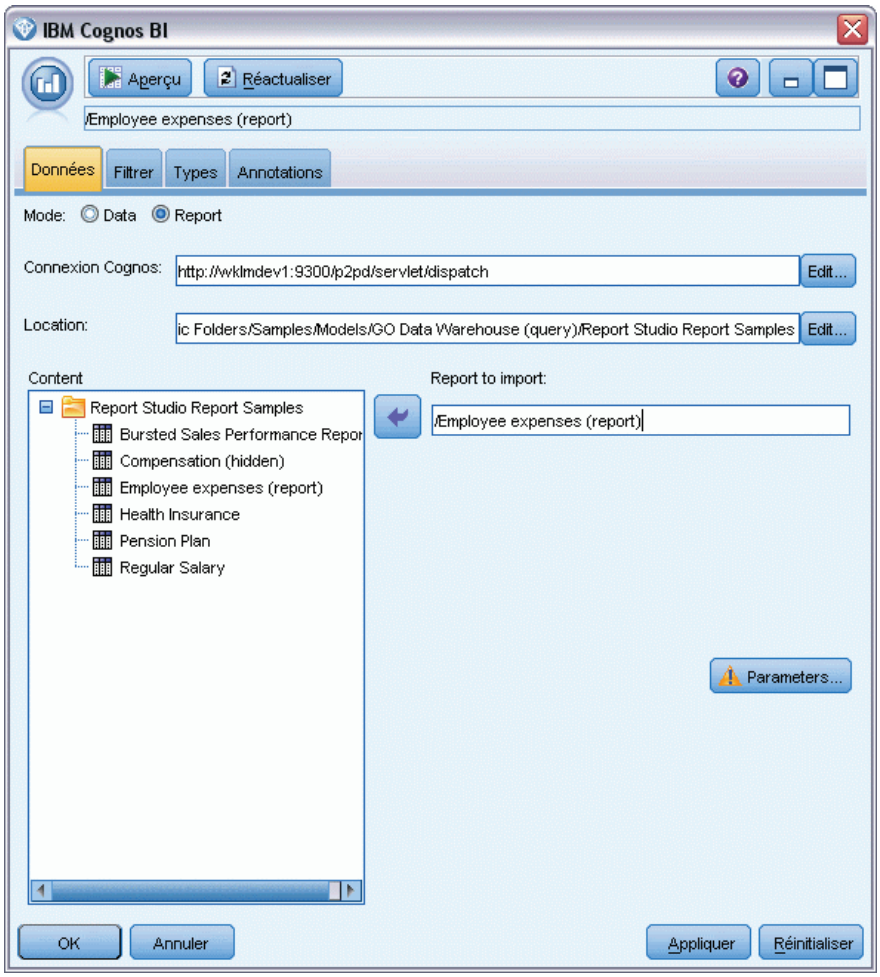
Paramètres. Si ce bouton est activé, les paramètres de l'objet sélectionné sont définis. Vous pouvez utiliser les paramètres pour effectuer des réglages (par exemple effectuer un calcul paramétré) avant d'importer les données. Si les paramètres sont définis mais qu'aucune valeur par défaut n'est fournie, le bouton affiche un triangle d'avertissement. Cliquez sur le bouton pour afficher les paramètres et les modifier le cas échéant. Si le bouton est désactivé, aucun paramètre n'est défini pour le rapport.

Agrégez les données avant l'importation. Cochez cette case si vous souhaitez importer des données agrégées plutôt que des données brutes.

Importer des rapports Cognos

Pour importer un rapport prédéfini d'une base de données Cognos BI IBM, assurez-vous que le Mode est réglé sur Rapport et remplissez la boîte de dialogue comme suit. *Remarque* : Seuls les rapports de liste unique peuvent être importés, les listes multiples ne sont pas prises en charge.

Figure 2-21
Importer des rapports Cognos



Connexion. Cliquez sur le bouton Modifier pour afficher une boîte de dialogue dans laquelle vous pourrez définir les détails d'une connexion Cognos à partir de laquelle importer les données ou les rapports. Si vous êtes déjà connecté à un serveur Cognos via IBM® SPSS® Modeler, vous pouvez également modifier les détails de la connexion actuelle. Pour plus d'informations, reportez-vous à la section Connexions Cognos sur p. 49.

Emplacement. Lorsque la connexion au serveur Cognos est établie, cliquez sur le bouton Modifier à côté de ce champ pour afficher une liste des packages disponibles depuis lesquels importer le contenu. Pour plus d'informations, reportez-vous à la section Sélection de l'emplacement de Cognos sur p. 50.

Contenu. Affiche le nom du package ou dossier sélectionné contenant les rapports. Accédez à un rapport spécifique, sélectionnez-le et cliquez sur la flèche droite pour amener le rapport dans le champ Rapport à importer.

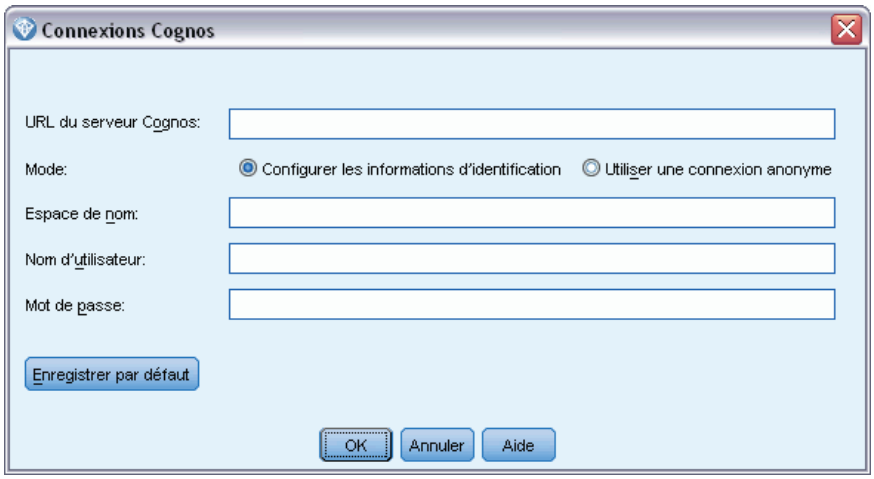
Rapport à importer. Indique le rapport que vous avez choisi d'importer dans SPSS Modeler. Si vous n'avez plus besoin de ce rapport, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet Contenu ou insérer un autre rapport dans ce champ.

Paramètres. Si ce bouton est activé, les paramètres du rapport sélectionné sont définis. Vous pouvez utiliser les paramètres pour effectuer des réglages avant d'importer le rapport (par exemple, spécifier une date de début et de fin pour les données de rapport). Si les paramètres sont définis mais qu'aucune valeur par défaut n'est fournie, le bouton affiche un triangle d'avertissement. Cliquez sur le bouton pour afficher les paramètres et les modifier le cas échéant. Si le bouton est désactivé, aucun paramètre n'est défini pour le rapport.

Connexions Cognos

La boîte de dialogue Connexions Cognos vous permet de sélectionner le serveur Cognos BI duquel vous souhaitez importer ou exporter les objets de base de données.

Figure 2-22
Sélection du serveur Cognos



URL du serveur Cognos. Saisissez l'URL du serveur Cognos BI depuis lequel vous souhaitez importer ou exporter. Ceci est la valeur de la propriété d'environnement « URL de répartiteur externe » de la configuration Cognos IBM sur le serveur Cognos BI. Contactez votre administrateur système Cognos si vous ne savez pas quel URL choisir.

Mode. Sélectionnez Définir les informations d'identification si vous souhaitez vous connecter avec un espace de nommage, nom d'utilisateur et mot de passe Cognos spécifiques (par exemple, en tant qu'administrateur). Sélectionnez Utiliser une connexion anonyme pour vous connecter sans informations de connexion utilisateur, auquel cas vous n'avez pas à remplir les autres champs.

Espace de nommage. Spécifiez le fournisseur de sécurité pour l'authentification Cognos utilisé pour se connecter au serveur. Le fournisseur pour l'authentification permet de définir et de gérer les utilisateurs, les groupes et les rôles et de contrôler le processus d'authentification.

Nom d'utilisateur. Entrez le nom d'utilisateur Cognos avec lequel effectuer la connexion au serveur.

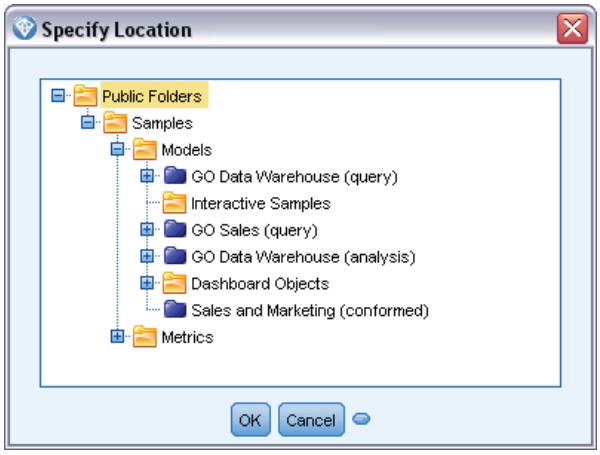
Mot de passe. Entrez le mot de passe associé au nom d'utilisateur défini.

Enregistrer comme paramètres par défaut. Cliquez sur ce bouton pour stocker ces paramètres comme paramètres par défaut et ainsi éviter d'avoir à les saisir à chaque fois que vous ouvrez le noeud.

Sélection de l'emplacement de Cognos

La boîte de dialogue Spécifier l'emplacement vous permet de sélectionner un package Cognos à partir duquel importer des données, ou un package ou dossier à partir duquel importer des rapports.

Figure 2-23
Sélection de l'emplacement de Cognos



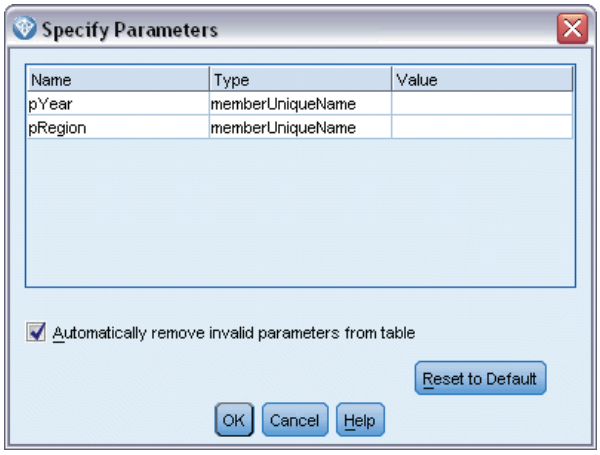
Dossiers publics. Si vous importez des données, ceci répertorie les packages et dossiers disponibles sur le serveur choisi. Sélectionnez le package désiré et cliquez sur OK. Vous ne pouvez sélectionner qu'un seul package par noeud source Cognos BI.

Si vous importez des rapports, ceci répertorie les packages et dossiers contenant des rapports disponibles sur le serveur choisi. Sélectionnez un dossier de package ou de rapport et cliquez sur OK. Vous ne pouvez sélectionner qu'un seul dossier de package ou de rapport par noeud source Cognos BI, bien que les dossiers de rapport puissent contenir d'autres dossiers de rapport ainsi que des rapports individuels.

Spécification de paramètres pour les données ou les rapports

Si les paramètres ont été définis dans Cognos BI, soit pour un objet de données soit pour un rapport, vous pouvez spécifier les valeurs de ces paramètres avant d'importer l'objet ou le rapport. Par exemple, les paramètres d'un rapport peuvent être les dates de début et de fin associées au contenu du rapport.

Figure 2-24
Paramètres Cognos



Nom. Le nom du paramètre tel qu'il est spécifié dans la base de données Cognos BI.

Type. Une description du paramètre.

Valeur. La valeur à attribuer au paramètre. Pour saisir ou modifier une valeur, double-cliquez sur sa cellule dans le tableau. Les valeurs ne sont pas validées ici, par conséquent les valeurs non valides sont détectées au moment de l'exécution.

Supprimer automatiquement les paramètres non valides de la table. Cette option est sélectionnée par défaut et supprimera tout paramètre non valide trouvé dans l'objet de données ou le rapport.

Noeud source SAS

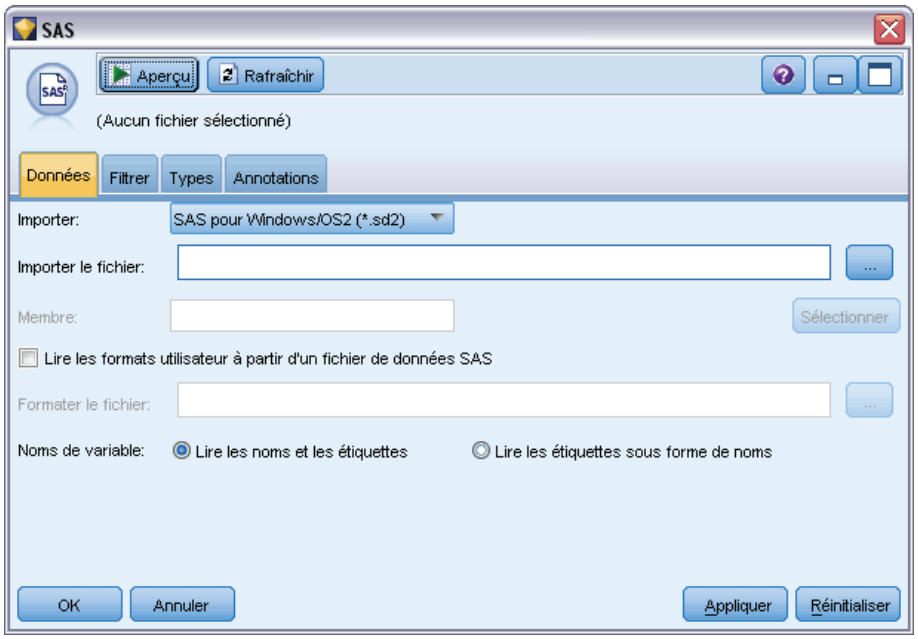
Remarque : cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud source SAS permet d'inclure des données SAS dans votre session de Data mining. Vous pouvez importer quatre types de fichier :

- SAS pour Windows/OS2 (*.sd2*)
- SAS pour UNIX (*.ssd*)
- Fichier Transport SAS (*.tpt*)
- SAS Version 7/8/9 (*.sas7bdat*)

Une fois les données importées, toutes les variables sont conservées et aucun type de variable n'est modifié. Toutes les observations sont sélectionnées.

Figure 2-25
Importation d'un fichier SAS



Définition des options du noeud source SAS

Importer. Sélectionnez le type de fichier SAS à importer. Vous pouvez choisir SAS pour Windows/OS2 (.sd2), SAS pour UNIX (.SSD), Fichier Transport SAUVEGARDES (.tpt) ou SAS Version 7/8/9 (.sas7bdat).

Importer le fichier. Indiquez le nom du fichier. Entrez directement le nom ou cliquez sur le bouton ... pour parcourir l'arborescence à la recherche du fichier.

Membre. Choisissez un membre à importer depuis le fichier Transport SAS sélectionné au-dessus. Vous pouvez entrer un nom de membre ou cliquer sur Sélectionner pour parcourir tous les membres du fichier.

Lire les formats utilisateur à partir d'un fichier de données SAS. Cochez cette case pour que les formats utilisateur soient lus. Les données et les formats de données (comme les étiquettes de variables) sont stockés dans différents fichiers. La plupart du temps, il est conseillé d'importer également les formats. Cependant, dans le cas d'ensembles de données volumineux, il peut être préférable de désélectionner cette option afin d'économiser la mémoire.

Formater le fichier. Si un fichier format est requis, cette zone de texte est active. Entrez directement le nom ou cliquez sur le bouton ... pour parcourir l'arborescence à la recherche du fichier.

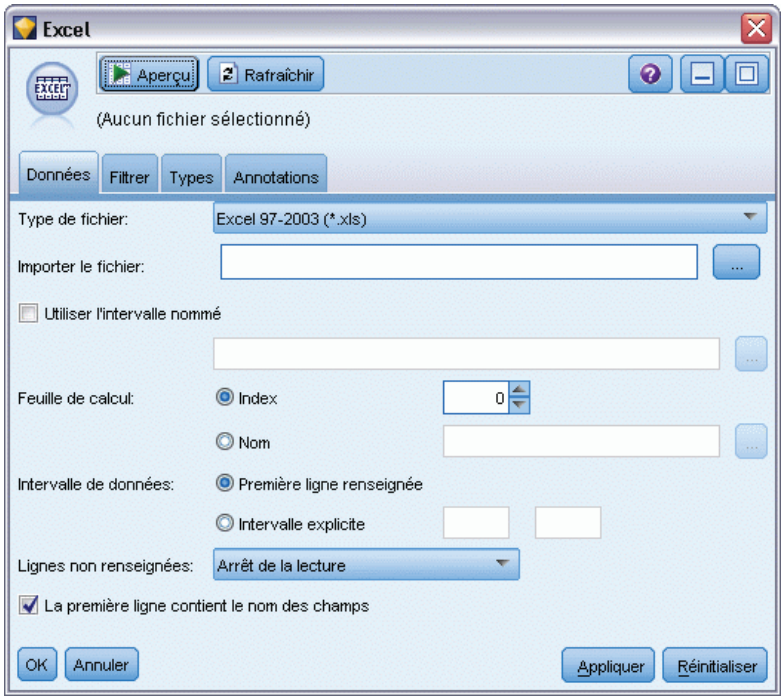
Noms des variables. Sélectionnez une méthode de gestion des noms et étiquettes de variable lors de l'importation des données d'un fichier SAS. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM® SPSS® Modeler. Vous pouvez également les réexporter pour les utiliser dans SAS.

- **Lire les noms et les étiquettes.** Sélectionnez cette option afin de lire les noms et les étiquettes de variable dans SPSS Modeler. Par défaut, cette option est sélectionnée et les noms de variable affichés dans le nœud Typer. Les étiquettes peuvent être affichées dans le Générateur de formules, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux.
- **Lire les étiquettes sous forme de nom.** Sélectionnez cette option pour lire les étiquettes de variable descriptives du fichier SAS au lieu des noms de champ abrégés, puis utilisez ces étiquettes en tant que noms de variable dans SPSS Modeler.

Noeud source Excel

Le noeud source Excel vous permet d'importer des données issues de n'importe quelle version de Microsoft Excel.

Figure 2-26
Noeud source Excel



Type de fichier. Sélectionnez le fichier de type Excel que vous souhaitez importer.

Importer le fichier. Indique le nom et l'emplacement de la feuille de calcul à importer.

Utiliser l'intervalle nommé. Permet d'indiquer un intervalle de cellules nommé, tel qu'il est défini dans la feuille de calcul Excel. Cliquez sur le bouton représentant des points de suspension (...) pour sélectionner la valeur souhaitée dans la liste des intervalles disponibles. Si vous utilisez un intervalle nommé, les autres paramètres de feuille de calcul et d'intervalle de données ne s'appliquent plus et sont, par conséquent, désactivés.

Choisissez une feuille de calcul. Indique la feuille de calcul à importer, via un index ou un nom.

- **Par index :** Définit la valeur d'index de la feuille de calcul à importer, 0 désignant la première feuille de calcul, 1, la deuxième et ainsi de suite.
- **Par nom.** Indiquez le nom de la feuille de calcul à importer. Cliquez sur le bouton représentant des points de suspension (...) pour sélectionner la valeur souhaitée dans la liste des feuilles de calcul disponibles.

Intervalle sur la feuille de calcul. Vous pouvez importer des données en partant de la première ligne renseignée ou en indiquant un intervalle de cellules explicite.

- **L'intervalle commence à la première ligne non vide.** Repère la première cellule renseignée et l'utilise comme angle supérieur gauche de l'intervalle de données.
- **Intervalle de cellules explicite.** Vous permet de spécifier un intervalle explicite par ligne et par colonne. Par exemple, pour spécifier l'intervalle Excel A1:D5 , vous pouvez entrer A1 dans le premier champ et D5 dans le second (ou bien, R1C1 et R5C4). Toutes les lignes de l'intervalle indiqué sont renvoyées, y compris les lignes vides.

Sur les lignes vides. Si plusieurs lignes vides sont rencontrées, vous pouvez Arrêter la lecture ou cliquer sur Renvoyer des lignes non renseignées pour poursuivre la lecture des données jusqu'à la fin de la feuille de calcul (lignes vides comprises).

La première ligne contient les noms de colonnes. Indique que la première ligne de l'intervalle spécifié doit être utilisée pour les noms de champ (de colonne). Si vous ne sélectionnez pas cette option, les noms de champ sont générés automatiquement.

Stockage de champ et niveau de mesure

Lors de la lecture de valeurs issues d'Excel, les champs de stockage numérique sont lus avec un niveau de mesure *continu* par défaut et les champs de chaîne sont lus avec un niveau *nominal*. Vous pouvez modifier manuellement le niveau de mesure (continu ou nominal) dans l'onglet Type, mais le stockage est, lui, déterminé automatiquement (il est toutefois possible, si nécessaire, de le modifier à l'aide d'une fonction de conversion, telle que to_integer, appliquée dans un noeud Remplacer ou Calculer). Pour plus d'informations, reportez-vous à la section Définition du stockage et du formatage des champs sur p. 32.

Par défaut, les champs comportant à la fois des valeurs numériques et des valeurs de type chaîne sont considérés comme numériques ; autrement dit, toute valeur de chaîne prendra la valeur nulle (manquante dans le système) dans IBM® SPSS® Modeler. Cela s'explique par le fait que—contrairement à Excel—SPSS Modeler n'autorise pas les types de stockage mixtes dans un même champ. Pour éviter ce type de problème, vous pouvez définir manuellement le format de cellule sur Texte dans la feuille de calcul Excel ; toutes les valeurs (y compris les nombres) sont ainsi lues en tant que chaînes.

Noeud source XML

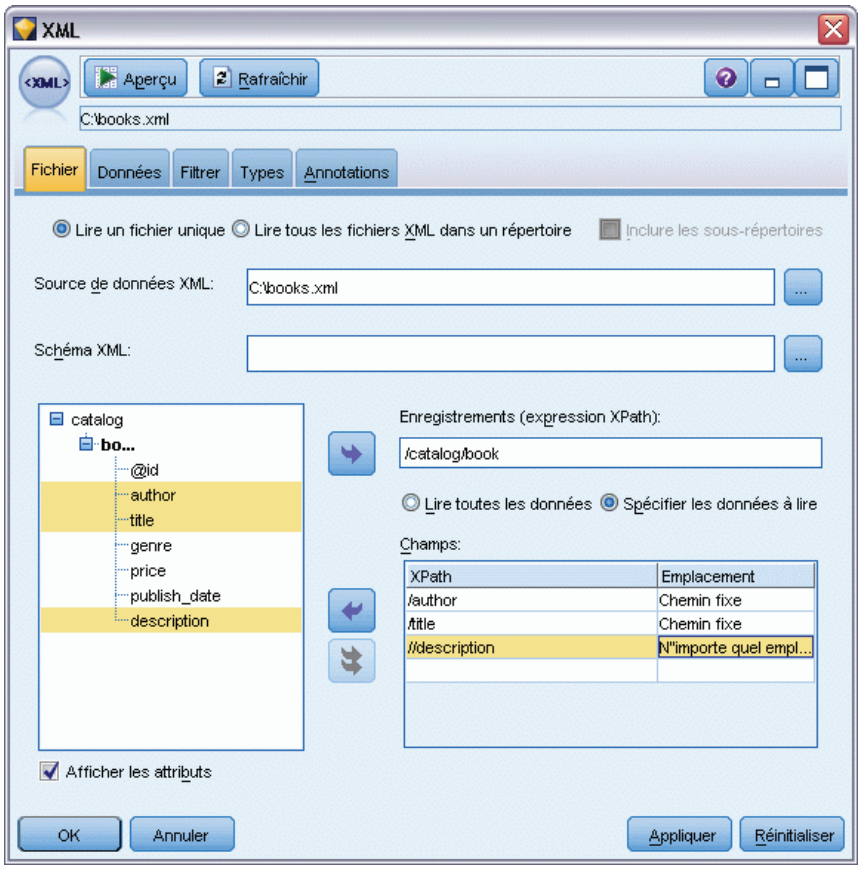
Remarque : cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud source XML vous permet d'importer des données depuis un fichier au format XML dans un flux IBM® SPSS® Modeler. XML est un langage standard d'échange de données, et représente pour beaucoup d'entreprises un format de choix approprié. Par exemple, un organisme

gouvernemental d'imposition souhaite analyser des données provenant de déclarations de revenus soumises en ligne et dont les données sont au format XML.

L'importation de données XML dans un flux SPSS Modeler vous permet d'exécuter une large gamme de fonctions d'analyse prédictive sur la source. Les données XML sont analysées dans un format tabulaire dans lequel les colonnes correspondent à différents niveaux d'imbrication des attributs et des éléments XML. Les éléments XML sont affichés au format XPath (voir *http://www.w3.org/TR/xpath20/*).

Figure 2-27
Importation des données XML



Lire un seul fichier. Par défaut, SPSS Modeler lit un seul fichier, que vous spécifiez dans le champ Source de données XML.

Lire tous les fichiers XML d'un répertoire. Sélectionnez cette option si vous souhaitez lire tous les fichiers XML d'un répertoire particulier. Spécifiez l'emplacement dans le champ Répertoire qui s'affiche. Cochez la case Inclure les sous-répertoires pour lire des fichiers XML supplémentaires dans tous les sous-répertoires du répertoire spécifié.

Sources de données XML. Saisissez le chemin complet et le nom du fichier de la source XML que vous souhaitez importer, ou utilisez le bouton Parcourir pour rechercher le fichier.

Schéma XML. (Facultatif) Spécifiez le chemin complet et le nom du fichier d'un fichier XSD ou DTD à partir duquel la structure XML est lue, ou utilisez le bouton Parcourir pour rechercher ce fichier. Si vous laissez ce champ vierge, la structure est lue à partir du fichier source XML. Un fichier XSD ou DTD peut avoir plus d'un élément racine. Dans ce cas, lorsque vous déplacez l'activation vers un autre champ, une boîte de dialogue s'affiche dans laquelle vous pouvez choisir l'élément racine à utiliser. Pour plus d'informations, reportez-vous à la section Sélection de plusieurs éléments racine sur p. 56.

Structure XML. Un arbre hiérarchique affichant la structure du fichier source XML (ou le schéma, si vous en avez spécifié un dans le champ Schéma XML). Pour définir une limite d'enregistrement, sélectionnez un élément et cliquez sur le bouton de la flèche droite pour copier l'élément dans le champ Enregistrements.

Afficher les attributs. Affiche ou masque les attributs, des éléments XML dans le champ Structure XML.

Enregistrements (expression XPath). Affiche la syntaxe XPath d'un élément copié à partir du champ de structure XML. Cet élément est alors mis en évidence dans la structure XML et définit la limite de l'enregistrement. À chaque fois que cet élément est rencontré dans le fichier source, un nouvel enregistrement est créé. Si ce champ est vide, le premier élément enfant sous la racine est utilisé comme limite d'enregistrement.

Lire toutes les données. Par défaut, toutes les données du fichier source sont lues dans le flux.

Spécifier les données à lire. Sélectionnez cette option pour importer des attributs, des éléments individuels ou les deux. Sélectionner cette option active le tableau Champs dans lequel vous pouvez spécifier les données que vous souhaitez importer.

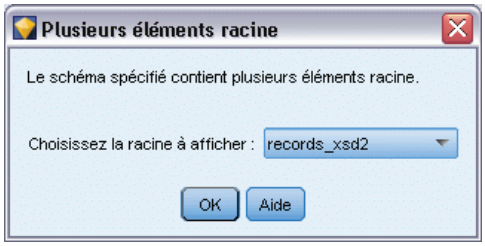
Champs. Ce tableau répertorie les éléments et les attributs sélectionnés pour l'importation, si vous avez sélectionné l'option Spécifier les données à lire. Vous pouvez soit saisir la syntaxe XPath d'un élément ou d'un attribut directement dans la colonne XPath, soit sélectionner un élément ou un attribut dans la structure XML et cliquer sur le bouton de la flèche droite pour copier l'élément dans le tableau. Pour copier tous les éléments enfants et les attributs d'un élément, sélectionnez l'élément dans la structure XML et cliquez sur le bouton en forme de double-flèche.

- **XPath.** La syntaxe Xpath des éléments à importer.
- **Emplacement.** L'emplacement dans la structure XML des éléments à importer. Chemin fixe affiche le chemin de l'élément en relation avec l'élément mis en évidence dans la structure XML (ou le premier élément enfant sous la racine, si aucun élément n'est mis en évidence). N'importe quel emplacement indique un élément du nom donné à n'importe quel emplacement de la structure XML. Personnalisé s'affiche si vous saisissez l'emplacement directement dans la colonne XPath.

Sélection de plusieurs éléments racine

Alors qu'un fichier XML correctement formé ne peut contenir qu'un seul élément racine, un fichier XSD ou DTD peut en contenir plusieurs. Si l'un des éléments racines correspond à celui du fichier source XML, cet élément racine est utilisé, sinon vous devez en sélectionnez un.

Figure 2-28
Sélection de plusieurs éléments racine



Choisissez la racine à afficher. Sélectionnez l'élément racine à utiliser. L'élément par défaut est le premier élément racine dans la structure XSD ou DTD.

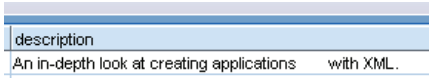
Suppression des espaces superflus des données source XML

Des sauts de ligne peuvent être implémentés dans les données source XML par une combinaison des caractères [CR][LF]. Dans certain cas, ces sauts de ligne peuvent apparaître au milieu d'une chaîne de texte, par exemple :

```
<description>Un examen en profondeur de la création d'applications[CR][LF]
avec XML.</description>
```

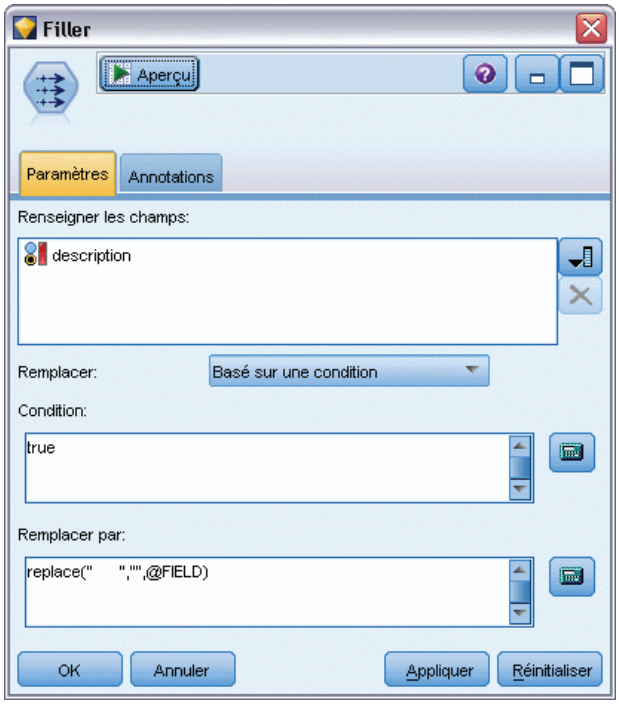
Ces sauts de ligne peuvent ne pas être visibles lorsque le fichier est ouvert dans certaines applications, par exemple dans un navigateur Web. Toutefois, lorsque les données sont lues dans le flux à travers le noeud source XML, les sauts de ligne sont convertis en une série de caractères d'espacement, par exemple :

Figure 2-29
enregistrement XML avec saut de ligne affiché en tant qu'espace



Vous pouvez corriger ceci en utilisant un noeud Remplacer et supprimer ces espaces superflus :

Figure 2-30
Noeud Remplacer avec des paramètres de suppression des espaces



Voici un exemple de la manière de traiter ce problème :

- Reliez un noeud Remplacer au noeud source XML.
- Ouvrez le noeud Remplacer et utilisez le sélecteur de champs pour sélectionner le champ contenant les espace superflus.
- Définissez Remplacer sur Basé sur une condition et Condition sur true (vrai).
- Dans le champ Remplacer par, entrez replace(" ","",@FIELD) et cliquez sur OK.
- Reliez un noeud Table au noeud Remplacer et exécutez le flux.

Dans la sortie du noeud Table, le texte apparaît désormais comme suit :

Figure 2-31
Enregistrement XML dont les espaces superflus ont été supprimés

description
An in-depth look at creating applications with XML.

Noeud Utilisateur

Le nœud Utilisateur représente une façon simple de créer des données synthétiques (à partir de zéro ou en modifiant des données existantes). Ceci est utile, par exemple, si vous souhaitez créer un ensemble de données de test pour la modélisation.

Création intégrale de données

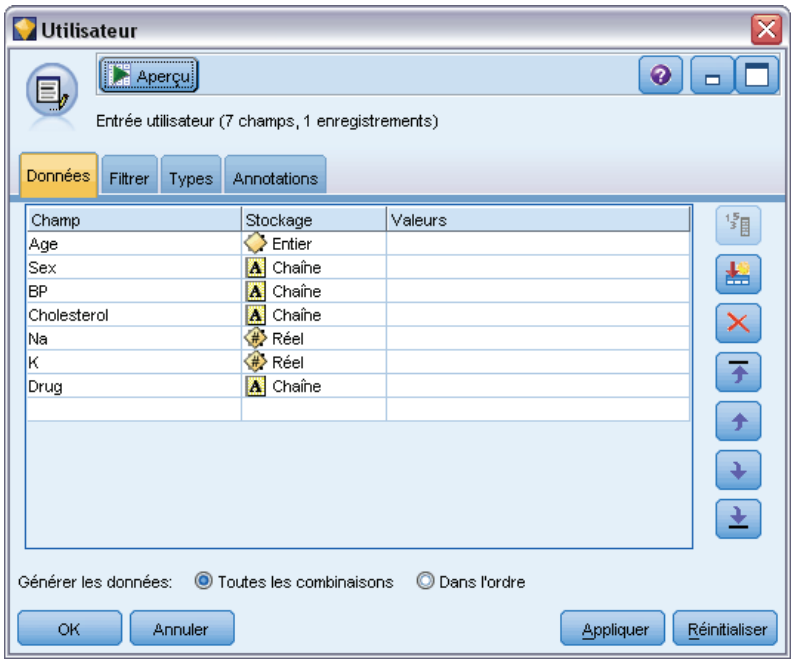
Le noeud Utilisateur est disponible dans la palette Sources. Vous pouvez l'ajouter directement à l'espace de travail de flux.

- Cliquez sur l'onglet Sources de la palette de noeuds.
- Faites glisser le nœud Utilisateur ou double-cliquez dessus pour l'ajouter à l'espace de travail de flux.
- Double-cliquez pour ouvrir la boîte de dialogue correspondante, et spécifiez les champs et les valeurs.

Remarque : les noeuds Utilisateur sélectionnés depuis la palette Sources sont entièrement vides (ils ne contiennent aucun champ et aucune information sur les données). Ceci permet de créer intégralement des données synthétiques.

Génération de données à partir d'une source de données existante

Figure 2-32
Noeud Utilisateur généré à partir d'un noeud de flux



Vous pouvez également générer un noeud Utilisateur à partir de tout noeud non terminal dans le flux :

- Déterminez l'emplacement dans le flux du nœud à remplacer.
- Cliquez avec le bouton droit de la souris sur le noeud qui alimentera le noeud Utilisateur en données, puis sélectionnez Générer le noeud Utilisateur dans le menu.
- Le nœud Utilisateur sera associé à tous les processus en aval qui lui sont connectés, remplaçant le nœud existant à cet emplacement dans votre flux de données. Une fois généré, le nœud

hérite de toute la structure des données et des informations de type de champ (le cas échéant) des métadonnées.

Remarque : si les données ne sont pas transmises par tous les noeuds du flux, ces derniers ne sont pas entièrement instanciés. Cela signifie que les valeurs de stockage et de données ne seront peut-être pas disponibles lors du remplacement par un nœud Utilisateur.

Définition des options du noeud Utilisateur

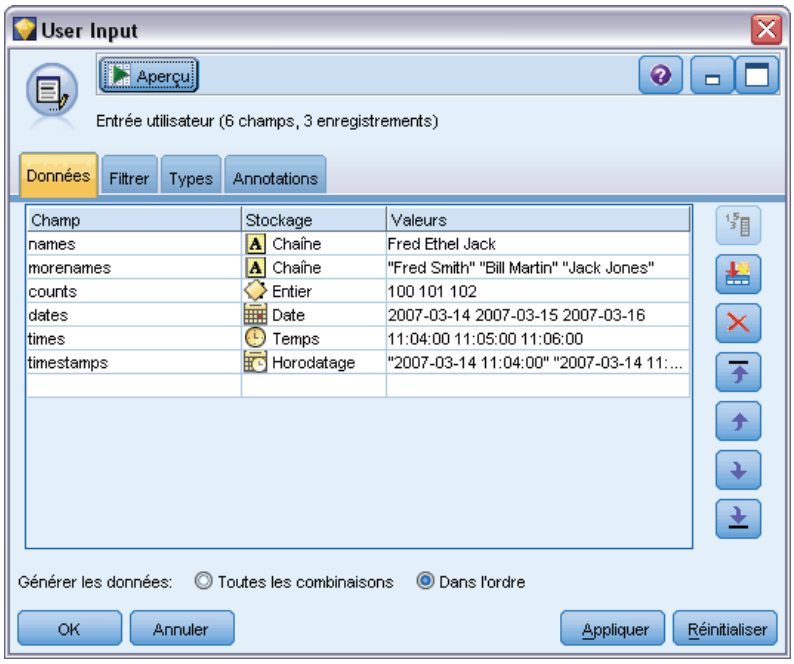
La boîte de dialogue d'un noeud Utilisateur contient différents outils que vous pouvez utiliser pour entrer des valeurs et définir la structure des données synthétiques. Pour un nœud généré, le tableau de l'onglet Données contient les noms de champ de la source de données d'origine. Pour un nœud ajouté à partir de la palette Sources, le tableau est vide. A l'aide des options du tableau, vous pouvez effectuer les opérations suivantes :

- Ajouter de nouveaux champs à l'aide du bouton Ajouter un nouveau champ situé à droite du tableau.
- Renommer des champs existants.
- Spécifier le stockage des données pour chaque champ.
- Indiquer des valeurs.
- Modifier l'ordre des champs affichés.

Saisie de données

Pour chaque champ, vous pouvez spécifier des valeurs ou en insérer à partir de l'ensemble de données d'origine via le bouton de sélection des valeurs situé à droite du tableau. Pour plus d'informations sur la spécification des valeurs, reportez-vous aux règles décrites ci-dessous. Vous pouvez également choisir de laisser le champ vide. Les champs vides sont renseignés avec la valeur système nulle (**$null$**).

Figure 2-33
Spécification du type de stockage pour les champs dans un nœud Utilisateur généré



Pour indiquer des valeurs de chaîne, saisissez-les dans la colonne de valeurs, séparées par des espaces :

```
Fred Ethel Martin
```

Les chaînes qui incluent des espaces peuvent être mises entre doubles guillemets :

```
«Bill Smith» «Fred Martin» «Jack Jones»
```

Pour les champs numériques, vous pouvez entrer des valeurs multiples de la même manière (utiliser des espaces pour délimiter les valeurs) :

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

Vous pouvez également spécifier les mêmes valeurs en indiquant le premier et le dernier nombre (10, 20), et l'incrément qui les sépare (2). Dans ce cas, vous entrerez :

```
10,20,2
```

Ces deux méthodes peuvent être combinées par imbrication, comme suit :

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

Cette syntaxe produira les valeurs suivantes :

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

Les valeurs date et heure peuvent être saisies à l'aide du format par défaut actuel sélectionné dans la boîte de dialogue Propriétés du flux, par exemple :

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

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

Pour les valeurs d'horodatage, qui comportent à la fois un composant date et heure, des guillemets doubles doivent être utilisés :

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

Pour plus d'informations, reportez-vous aux commentaires sur le stockage des données ci-dessous.

Générer les données. Permet d'indiquer comment les enregistrements sont générés lorsque vous exécutez le flux.

- **Toutes les combinaisons.** Génère des enregistrements contenant toutes les combinaisons possibles des valeurs de champ de façon à ce que chaque valeur de champ apparaisse dans plusieurs enregistrements. Cette procédure peut parfois générer plus de données que ce que vous souhaitez. Par conséquent, vous pouvez faire suivre ce noeud d'un noeud Echantillonner.
- **Dans l'ordre**. Génère des enregistrements dans l'ordre dans lequel les valeurs du champ de données sont indiquées. Chaque valeur de champ n'apparaît que dans un enregistrement. Le nombre total d'enregistrements est égal au plus grand nombre de valeurs pour un seul champ. Lorsque le nombre de valeurs des champs est inférieur au plus grand nombre de valeurs, des valeurs non définies ($null$) sont insérées.

Par exemple, les entrées suivantes généreront les enregistrements répertoriés dans les tableaux ci-dessous.

- **Age.** 30,60,10
- **TA.** FAIBLE
- **Cholestérol.** NORMAL ELEVE
- **Médicament.** (vide)

Option Générer les données définie sur Toutes les combinaisons :

Age	TA	Cholestérol	Médicament
30	FAIBLE	NORMAL	$null$
30	FAIBLE	ELEVEE	$null$
40	FAIBLE	NORMAL	$null$
40	FAIBLE	ELEVEE	$null$
50	FAIBLE	NORMAL	$null$
50	FAIBLE	ELEVEE	$null$
60	FAIBLE	NORMAL	$null$
60	FAIBLE	ELEVEE	$null$

Option Générer les données définie sur Dans l'ordre :

Age	TA	Cholestérol	Médicament
30	FAIBLE	NORMAL	$null$
40	$null$	ELEVEE	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Stockage des données

Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Par exemple, vous pouvez définir le niveau de mesure d'un champ de nombre entier comportant les valeurs 1 et 0 comme étant un champ *Booléen*. En général, 1 correspond à la valeur *True (vrai)* et 0 à la valeur *False (faux)*. Alors que le stockage doit être déterminé au niveau de la source, le niveau de mesure peut être modifié à l'aide d'un noeud Typer en tout point du flux. Pour plus d'informations, reportez-vous à la section Niveaux de mesure dans le chapitre 4 sur p. 138.

Les types de stockage disponibles sont les suivants :

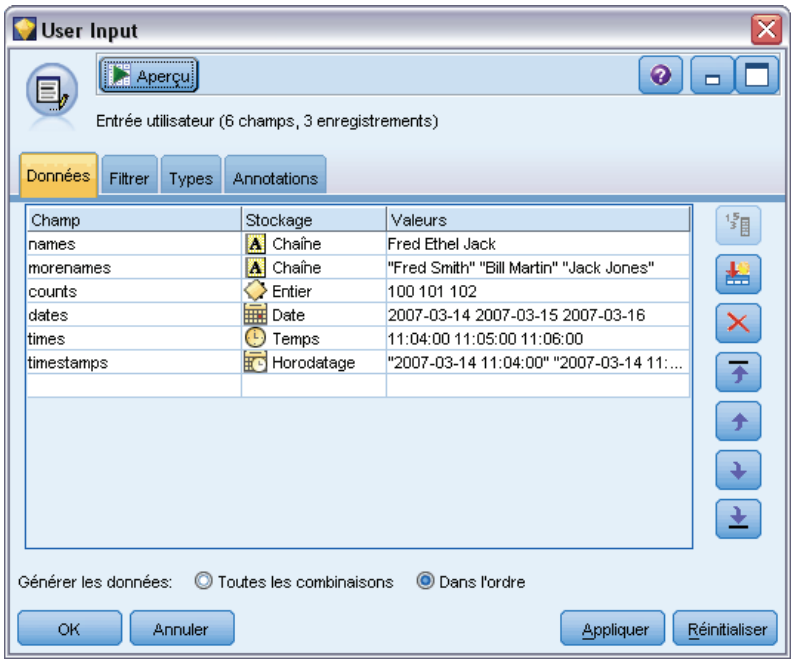
- **Chaîne.** Utilisé pour les champs contenant des données non numériques, également appelées données alphanumériques. Une chaîne peut inclure n'importe quelle séquence de caractères, telle que *fred*, *Classe 2* ou *1234*. Notez que les nombres utilisés dans les chaînes ne peuvent pas être inclus dans les calculs.
- **Entier.** Champ dont les valeurs sont des entiers.
- **Réel.** Il s'agit de nombres pouvant comporter des décimales (pas uniquement des entiers). Le format d'affichage est indiqué dans la boîte de dialogue Propriétés de flux et peut être ignoré pour des champs individuels dans un noeud Typer (onglet Format).
- **Date.** Valeurs de date indiquées dans un format standard, comme année, mois et jour (par exemple, 26.09.07). Le format exact est indiqué dans la boîte de dialogue Propriétés de flux.
- **Heure :** Valeur indiquant une durée. Par exemple, un appel de service ayant duré 1 heure, 26 minutes et 38 secondes peut être représenté sous la forme 01:26:38, en fonction du format d'heure actuel indiqué dans la boîte de dialogue Propriétés de flux.
- **Horodatage.** Valeurs comportant à la fois un composant de date et d'heure, par exemple 2007–09–26 09:04:00, dépendant une fois de plus des formats de date et d'heure dans la boîte de dialogue Propriétés de flux. Remarque : il se peut que les valeurs d'horodatage doivent être placées entre guillemets doubles pour être interprétées comme une valeur unique et non comme des valeurs date et heure distinctes. (Cela s'applique, par exemple, lors de la saisie de valeurs dans un noeud Utilisateur).

Conversion de stockages. Vous pouvez également convertir le stockage d'un champ à l'aide de diverses fonctions de conversion, comme to_string et to_integer dans un noeud Remplacer. Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer dans le chapitre 4 sur p. 180. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue des propriétés du flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez MOIS AAAA comme format de date par défaut pour le flux. Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne contenant des valeurs catégorielles. Pour plus d'informations, reportez-vous à la section Recodage des valeurs à l'aide du noeud Calculer dans le chapitre 4 sur p. 177.

Lecture de données mixtes. Au cours de la lecture des champs de stockage numérique (entier, nombre réel, heure, horodatage ou date), toutes les valeurs non numériques sont définies comme étant nulles ou manquantes dans le système. En effet, contrairement à certaines applications, IBM® SPSS® Modeler n'autorise pas les types de stockage mixtes au sein d'un champ. Pour éviter ce type de problème, faites en sorte que les champs comportant des données mixtes soient lus en tant que chaînes ; pour cela, modifiez le type de stockage dans le noeud source ou dans l'application externe.

Remarque : les noeuds Utilisateur générés peuvent déjà contenir ces informations de stockage, recueillies à partir du noeud source si ce dernier a été instancié. Un nœud non instancié ne contient pas d'informations de stockage ou de type d'utilisation.

Figure 2-34
Spécification du type de stockage pour les champs dans un nœud Utilisateur généré



Règles pour la spécification des valeurs

Dans le cas des champs symboliques, vous devez séparer les valeurs multiples par des espaces, par exemple :

ELEVE MOYEN FAIBLE

Pour les champs numériques, vous pouvez entrer des valeurs multiples de la même manière (utiliser des espaces pour délimiter les valeurs) :

10 12 14 16 18 20

Vous pouvez également spécifier les mêmes valeurs en indiquant le premier et le dernier nombre (10, 20), et l'incrément qui les sépare (2). Dans ce cas, vous entrerez :

10,20,2

Ces deux méthodes peuvent être combinées par imbrication, comme suit :

1 5 7 10,20,2 21 23

Cette syntaxe produira les valeurs suivantes :

1 5 7 10 12 14 16 18 20 21 23

Onglets communs des noeuds source

Les options suivantes peuvent être spécifiées pour tous les noeuds source en cliquant sur l'onglet correspondant :

- **Onglet Données.** Permet de modifier le type de stockage par défaut.
- **Onglet Filtrer.** Permet d'éliminer ou de renommer des champs de données. Cet onglet offre les mêmes fonctions que le nœud Filtrer. Pour plus d'informations, reportez-vous à la section Paramétrage des options de filtrage dans le chapitre 4 sur p. 156.
- **Onglet Types.** Permet de définir les niveaux de mesure. Cet onglet offre les mêmes fonctions que le nœud Typer.
- **Onglet Annotations.** Utilisé pour tous les noeuds, cet onglet propose des options permettant de renommer les noeuds, de créer des info-bulles personnalisées et de stocker de longues annotations.

Définition des niveaux de mesure dans le nœud source

Les propriétés de champ peuvent être indiquées dans un noeud source ou dans un noeud Typer distinct. Les fonctionnalités sont similaires dans les deux noeuds. Les propriétés suivantes sont disponibles :

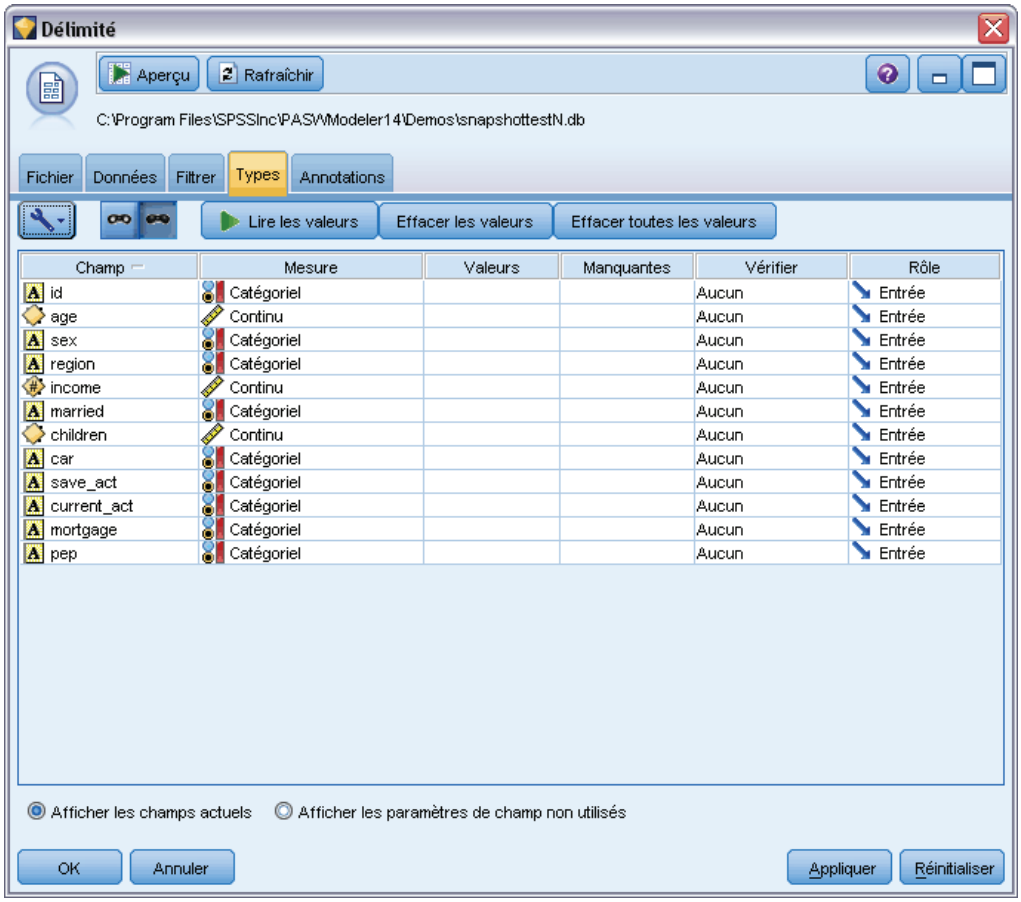
- **Champ.** Double-cliquez sur un nom de champ pour indiquer les étiquettes de valeur et de champ des données de IBM® SPSS® Modeler. Par exemple, vous pouvez consulter ou modifier ici les métadonnées de champ importées à partir de IBM® SPSS® Statistics. De même, vous pouvez créer des étiquettes pour les champs et leurs valeurs. La présence des étiquettes indiquées ici dans SPSS Modeler dépend des sélections effectuées dans la boîte de dialogue Propriétés du flux.
- **Mesure.** Il s'agit d'un niveau de mesures qui permet de décrire les caractéristiques des données d'un champ précis. Si tous les détails d'un champ sont connus, il est dit **complètement instancié**. Pour plus d'informations, reportez-vous à la section Niveaux de mesure dans le chapitre 4 sur p. 138.

  Remarque : Le niveau de mesure d'un champ est différent de son type de stockage, qui indique si les données sont stockées sous forme de chaînes, d'entiers, de nombres réels, de dates, d'heures ou d'horodatages.
- **Valeurs :** Cette colonne vous permet de spécifier des options pour la lecture de valeurs de données à partir de l'ensemble de données ou d'utiliser l'option Spécifier afin de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte. Vous pouvez également choisir de transférer les champs sans lire leurs valeurs. Pour plus d'informations, reportez-vous à la section Valeurs de données dans le chapitre 4 sur p. 143.

- **Manquant :** Permet de spécifier le traitement des valeurs manquantes du champ. Pour plus d'informations, reportez-vous à la section Définition de valeurs manquantes dans le chapitre 4 sur p. 149.
- **Vérifier.** Dans cette colonne, vous pouvez définir des options pour garantir que les valeurs de champ sont conformes aux intervalles ou aux valeurs spécifiés. Pour plus d'informations, reportez-vous à la section Vérification des valeurs de type dans le chapitre 4 sur p. 149.
- **Rôle.** Permet d'indiquer aux noeuds de modélisation si les champs sont des champs d'entrée (champs prédicteurs) ou de cible (champs prédits) pour un processus d'apprentissage automatique. Sont également disponibles les rôles Les deux et Aucun, et l'option Partition. Cette dernière signale les champs utilisés pour partitionner les enregistrements en échantillons distincts à des fins d'apprentissage, de test et de validation. La valeur Diviser spécifie que des modèles séparés seront construits pour chaque valeur possible du champ. Pour plus d'informations, reportez-vous à la section Définition du rôle du champ dans le chapitre 4 sur p. 150.

Pour plus d'informations, reportez-vous à la section Noeud Typer dans le chapitre 4 sur p. 136.

Figure 2-35
Options de l'onglet Types



A quel moment procéder à l'instanciation au niveau du nœud source ?

Vous pouvez obtenir des informations sur le stockage et les valeurs de données de vos champs de deux façons différentes. Cette **instanciation** peut se produire au niveau du noeud source lorsque vous introduisez des données pour la première fois dans IBM® SPSS® Modeler, ou lorsque vous ajoutez un noeud Typer dans le flux de données.

L'instanciation au niveau du noeud source est utile dans les cas suivants :

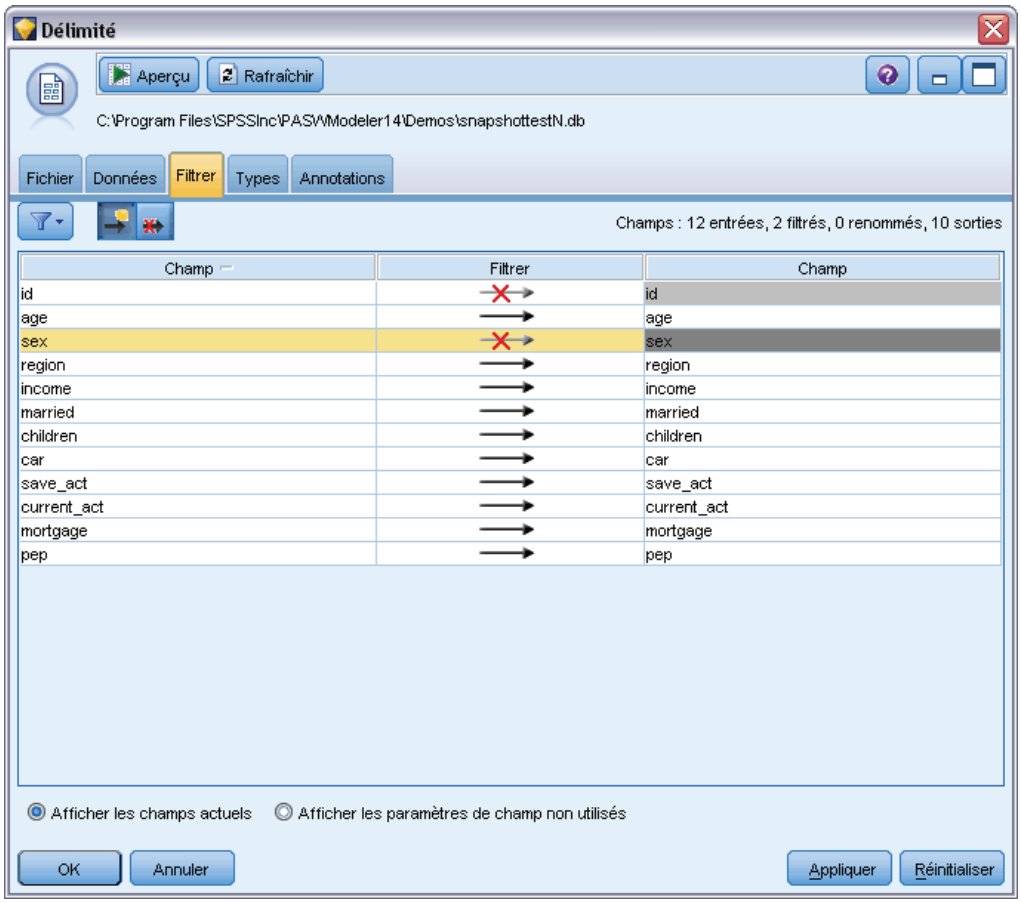
- L'ensemble de données n'est pas volumineux.
- Vous prévoyez de calculer de nouveaux champs à l'aide du Générateur de formules (l'instanciation rend les valeurs des champs disponibles à partir du Générateur de formules).

En général, si votre ensemble de données n'est pas trop volumineux et si vous ne prévoyez pas d'ajouter des champs au flux par la suite, l'instanciation au niveau du noeud source est la méthode la plus pratique.

Filtrage des champs à partir du noeud source

L'onglet Filtrer de la boîte de dialogue d'un noeud source vous permet d'exclure des champs des opérations en aval en fonction de votre examen initial des données. Ceci est utile, par exemple, s'il existe des champs en double dans les données ou si vous êtes déjà suffisamment familiarisé avec les données pour exclure les champs qui ne sont pas pertinents. Vous pouvez également ajouter ultérieurement au flux un noeud Filtrer distinct. Les fonctionnalités sont similaires dans les deux cas. Pour plus d'informations, reportez-vous à la section Paramétrage des options de filtrage dans le chapitre 4 sur p. 156.

Figure 2-36
Filtrage des champs à partir du noeud source



Chapitre
3

Noeuds d'opérations sur les lignes

Présentation des noeuds d'opérations sur les lignes

Les noeuds d'opérations sur les lignes permettent d'apporter des modifications aux enregistrements de données. Ces opérations sont importantes durant les phases de **compréhension des données** et de **préparation des données** du Data mining parce qu'elles vous permettent d'adapter les données à vos besoins commerciaux.

Par exemple, selon les résultats de l'audit que vous avez mené à l'aide du noeud Audit données (palette Sortie), vous pouvez décider de fusionner les enregistrements achat client des trois derniers mois. Le noeud Fusionner permet de fusionner les enregistrements en fonction des valeurs d'un champ-clé, par exemple *ID client*. Vous pouvez également constater qu'une base de données d'informations relatives à la fréquentation d'un site Web devient impossible à gérer lorsqu'elle comporte plus d'un million d'enregistrements. Dans ce cas, utilisez un noeud Echantillonner pour sélectionner un sous-ensemble de données à utiliser lors de la modélisation.

La palette Opérations sur les lignes contient les noeuds suivants :

Le noeud Sélectionner permet de sélectionner ou d'exclure des sous-ensembles d'enregistrements d'un flux de données sur la base d'une condition spécifique. Par exemple, vous pouvez sélectionner les enregistrements qui appartiennent à un secteur de ventes particulier. Pour plus d'informations, reportez-vous à la section Noeud Sélectionner sur p. 70.

Le noeud Echantillonner sélectionne un sous-ensemble d'enregistrements. Divers types d'échantillons sont pris en charge, notamment les échantillons stratifiés, en classe et non aléatoires (structurés). L'échantillonnage peut être utile pour améliorer les performances et pour sélectionner des groupes d'enregistrements associés ou des transactions pour analyse. Pour plus d'informations, reportez-vous à la section Noeud Echantillon sur p. 72.

Le noeud Equilibrer corrige les déséquilibres survenant dans un ensemble de données, de manière à respecter une condition précise. La règle d'équilibrage ajuste la proportion d'enregistrements présentant une condition True (vrai) par rapport au facteur indiqué. Pour plus d'informations, reportez-vous à la section Noeud Equilibrer sur p. 81.

Le noeud Agréger remplace une séquence d'enregistrements d'entrée par des enregistrements de sortie abrégés et agrégés. Pour plus d'informations, reportez-vous à la section Noeud Agréger sur p. 83.

Le noeud agrégé Recency, Frequency, Monetary (RFM) vous permet de prendre les données de l'historique des transactions d'un client, d'en éliminer les éventuelles données inutilisées et de combiner le reste des données de transaction sur une seule ligne qui indique la date de la dernière consultation, le nombre de transactions réalisées et la valeur monétaire totale de ces transactions. Pour plus d'informations, reportez-vous à la section Noeud Agréger RFM sur p. 86.

© Copyright IBM Corporation 1994, 2012.

Le noeud Trier trie les enregistrements par ordre croissant ou décroissant, en fonction de la valeur d'un ou de plusieurs champs. Pour plus d'informations, reportez-vous à la section Noeud Trier sur p. 88.

Le noeud Fusionner permet de créer, à partir de plusieurs enregistrements d'entrée, un seul enregistrement de sortie contenant tout ou partie des champs d'entrée. Il sert notamment à fusionner des données provenant de différentes sources, telles que les données client internes et les données démographiques acquises. Pour plus d'informations, reportez-vous à la section Noeud Fusionner sur p. 90.

Le noeud Ajouter réalise la concaténation d'ensembles d'enregistrements. Il permet de combiner des ensembles de données dont les structures sont similaires, mais les données différentes. Pour plus d'informations, reportez-vous à la section Noeud Ajouter sur p. 101.

Le noeud Distinguer supprime les enregistrements en double, soit en incluant le premier enregistrement dans le flux de données, soit en le supprimant et en incluant ses doublons dans le flux de données. Pour plus d'informations, reportez-vous à la section Noeud Distinguer sur p. 103.

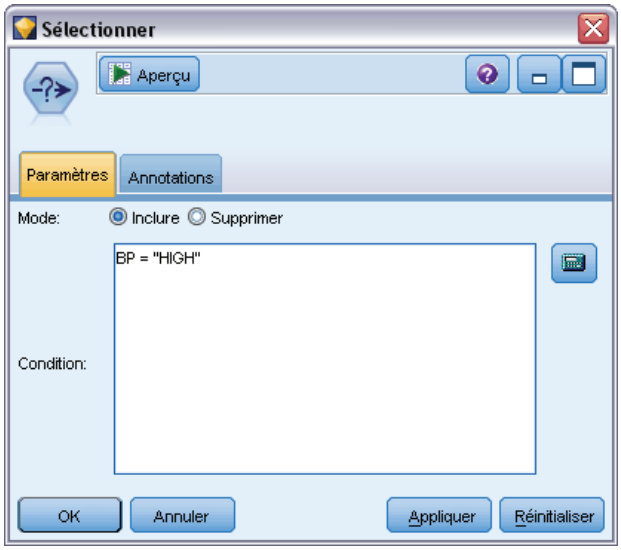
La plupart des noeuds de la palette Opérations sur les lignes nécessitent l'utilisation d'expressions CLEM. Si vous connaissez CLEM, vous pouvez saisir une expression dans le champ. Chaque champ d'expression comporte toutefois un bouton permettant d'ouvrir le Générateur de formules CLEM ; ce dernier vous aide à créer automatiquement de telles expressions.

Figure 3-1
Bouton du Générateur de formules

Noeud Sélectionner

Ce noeud permet de sélectionner ou de supprimer un sous-ensemble d'enregistrements du flux de données en fonction d'une condition spécifique, du type TA (tension artérielle) == ELEVEE.

Figure 3-2
Boîte de dialogue du noeud Sélectionner



Mode. Indique si les enregistrements répondant à la condition seront inclus dans le flux de données ou s'ils en seront exclus.

- **Enlever.** Permet d'inclure les enregistrements qui répondent à la condition de sélection.
- **Supprimer.** Permet d'exclure les enregistrements qui répondent à la condition de sélection.

Condition. Affiche la condition de sélection, spécifiée à l'aide d'une expression CLEM, qui sera utilisée pour tester les enregistrements. Entrez une expression dans la fenêtre ou utilisez le Générateur de formules en cliquant sur le bouton en forme de calculatrice situé à droite de la fenêtre.

Si vous choisissez d'ignorer des enregistrements en fonction d'une condition, comme dans l'exemple suivant :

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

le noeud Sélectionner par défaut ignore également les enregistrements ayant des valeurs nulles pour tous les champs de sélection. Pour éviter cela, ajoutez la condition suivante à la condition d'origine :

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

Les noeuds Sélectionner sont également utilisés pour choisir une proportion d'enregistrements. Normalement, cette opération est effectuée à l'aide d'un noeud Echantillonner. Cependant, si les paramètres disponibles ne sont pas adaptés à la complexité de la condition que vous souhaitez spécifier, vous pouvez créer cette dernière à l'aide d'un noeud Sélectionner. Une condition semblable à la suivante peut être créée :

TA = "ELEVEE" et aléatoire(10) <= 4

Avec cette condition, environ 40 % des enregistrements présentant une tension artérielle élevée seront sélectionnés et transmis aux noeuds en aval pour être analysés plus en détail.

Noeud Echantillon

Vous pouvez utiliser des noeuds Echantillonner pour sélectionner un sous-groupe d'enregistrements à analyser ou définir une proportion d'enregistrements à supprimer. Divers types d'échantillons sont pris en charge, notamment les échantillons stratifiés, en classe et non aléatoires (structurés). Vous pouvez utiliser l'échantillonnage à diverses fins :

- Pour améliorer les performances en évaluant les modèles d'un sous-groupe de données. Les modèles évalués à partir d'un échantillon sont souvent aussi précis que ceux issus de l'ensemble de données complet et plus encore si l'amélioration des performances permet de tester différentes méthodes que vous ne testeriez normalement pas.
- Pour sélectionner des groupes d'enregistrements ou de transactions associés à analyser, tels que tous les articles d'un panier en ligne ou toutes les propriétés d'un voisinage donné.
- Pour identifier des unités ou des observations pour une vérification aléatoire pour le contrôle de qualité, la prévention des fraudes ou la sécurité.

Remarque : Si vous souhaitez simplement diviser les données dans des échantillons d'apprentissage et de test à des fins de validation, vous pouvez utiliser un noeud Partitionner à la place. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 sur p. 207.

Types d'échantillons

Echantillons en classe. Echantillonnent des groupes ou des classes et non des unités individuelles. Supposons que vous disposiez d'un fichier de données comportant un enregistrement pour chaque élève. Si vous classez en fonction de l'école et que la taille de l'échantillon est 50 %, 50 % des écoles seront choisies et tous les élèves de chacune des écoles sélectionnées seront sélectionnés. Les élèves des écoles non sélectionnées seront alors rejetés. En moyenne, 50 % des élèves devraient être sélectionnés, mais étant donné que les écoles ont des tailles différentes, le pourcentage peut ne pas être exact. De même, vous pouvez classer les articles d'un panier en fonction de l'ID de la transaction pour pouvoir conserver tous les articles des transactions sélectionnées. Pour un exemple de classement des propriétés en fonction de la ville, voir le flux d'échantillon *complexsample_property.str*.

Echantillons stratifiés. Sélectionnent les échantillons indépendamment dans des sous-groupes sans chevauchement de population, ou strates. Vous pouvez, par exemple, faire en sorte que tous les hommes et femmes soient échantillonnés dans des proportions égales ou que chaque région ou groupe socio-économique d'une population soient représentés. Vous pouvez également définir une taille d'échantillon différente pour chaque strate (par exemple, si vous pensez qu'un groupe est sous-représenté dans les données d'origine). Pour un exemple de stratification des propriétés en fonction du pays, voir le flux d'échantillon *complexsample_property.str*.

Echantillonnage systématique ou 1 en n Lorsqu'il est difficile d'obtenir une sélection aléatoirement, les unités peuvent être échantillonnées systématiquement (à intervalles fixes) ou séquentiellement.

Pondérations d'échantillonnage : Des pondérations d'échantillonnage sont calculées automatiquement lors de la création du graphique d'un échantillon complexe et elles correspondent approximativement à la "fréquence" que chaque unité échantillonnée représente dans les données

d'origine. Par conséquent, la somme des pondérations sur l'échantillon doit évaluer la taille des données d'origine.

Cadre d'échantillonnage

Un cadre d'échantillonnage définit la source des observations potentielles à inclure dans un échantillon ou une étude. Dans certains cas, il peut être possible d'identifier chaque membre d'une population et d'inclure n'importe quel membre dans un échantillon, par exemple, lors de l'échantillonnage des éléments qui proviennent d'une chaîne de production. Dans la plupart des cas, vous ne pourrez pas accéder à chacune des observations possibles. Par exemple, vous ne pouvez pas savoir qui va voter dans une élection tant que l'élection n'a pas eu lieu. Dans ce cas, vous pouvez utiliser le registre des inscrits comme cadre d'échantillonnage, même si certaines personnes inscrites ne voteront pas, sachant que des personnes peuvent voter bien qu'elles ne figurent pas dans le registre au moment où vous vérifiez le registre. Toute personne ne figurant pas dans le cadre d'échantillonnage ne peut pas être échantillonnée. La représentation de la population à évaluer par le cadre d'échantillonnage doit être traitée pour chaque observation réelle.

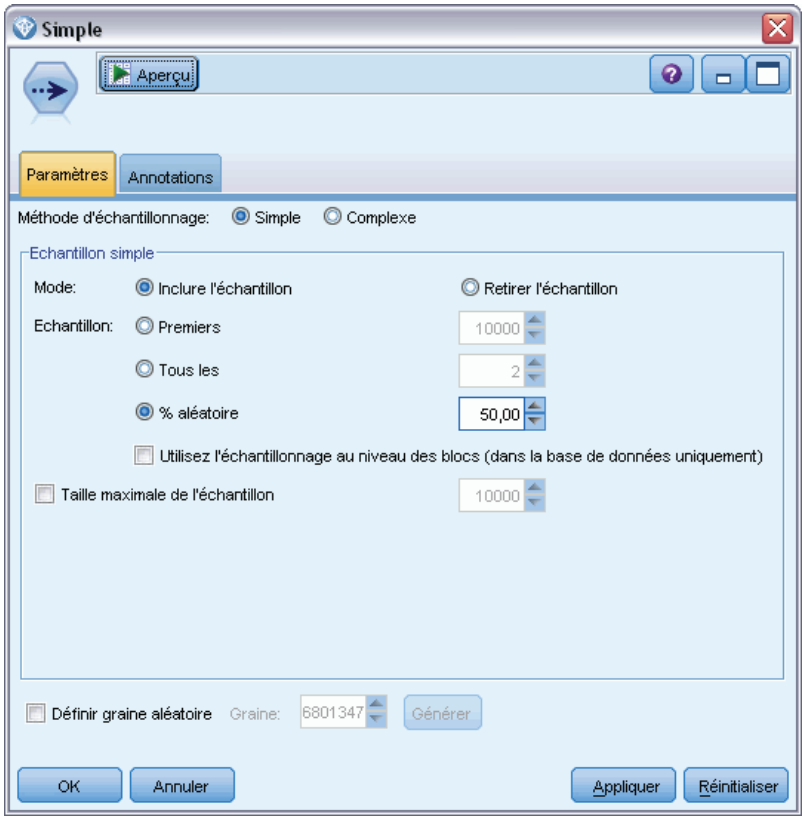
Options de noeud Echantillonner

Vous pouvez choisir la méthode simple ou complexe en fonction de vos besoins.

Options d'échantillonnage simple

La méthode simple permet de sélectionner un pourcentage aléatoire d'enregistrements, de sélectionner des enregistrements contigus ou de sélectionner chaque *nième* enregistrement.

Figure 3-3
Options d'échantillonnage simple



Mode. Choisissez de transmettre (inclure) ou de supprimer (exclure) les enregistrements pour les modes suivants :

- **Inclure l'échantillon.** Inclut les enregistrements sélectionnés dans le flux de données et supprime tous les autres. Par exemple, si vous définissez le mode sur Inclure l'échantillon et l'option 1 en n sur la valeur 5, un enregistrement sur cinq sera inclus pour donner un ensemble de données dont la taille est égale à environ un cinquième de la taille d'origine. Il s'agit du mode par défaut de l'échantillonnage des données et du seul mode disponible avec la méthode complexe.
- **Retirer l'échantillon.** Exclut les enregistrements sélectionnés et inclut tous les autres. Par exemple, si vous définissez le mode sur Retirer l'échantillon et l'option 1 en n sur la valeur 5, un enregistrement sur cinq sera exclu. Ce mode est disponible uniquement avec la méthode simple.

Exemple : Sélectionnez la méthode d'échantillonnage à partir des options suivantes :

- **Premiers.** Permet d'utiliser l'échantillonnage de données adjacentes. Si, par exemple, la taille d'échantillon maximale est 10 000, les 10 000 premiers enregistrements seront sélectionnés.

- **Tous les.** Sélectionnez cette option pour échantillonner les données en incluant ou en excluant un enregistrement sur *n*. Si, par exemple, *n* a la valeur 5, un enregistrement sur 5 sera sélectionné.
- **% aléatoire.** Permet d'échantillonner un pourcentage aléatoire de données. Par exemple, si vous indiquez la valeur 20, 20 % des données seront incluses dans le flux de données ou en seront exclues, selon le mode sélectionné. Dans le champ, indiquez le pourcentage d'échantillonnage. Vous pouvez également spécifier une valeur de graine à l'aide de l'option Définir graine aléatoire.

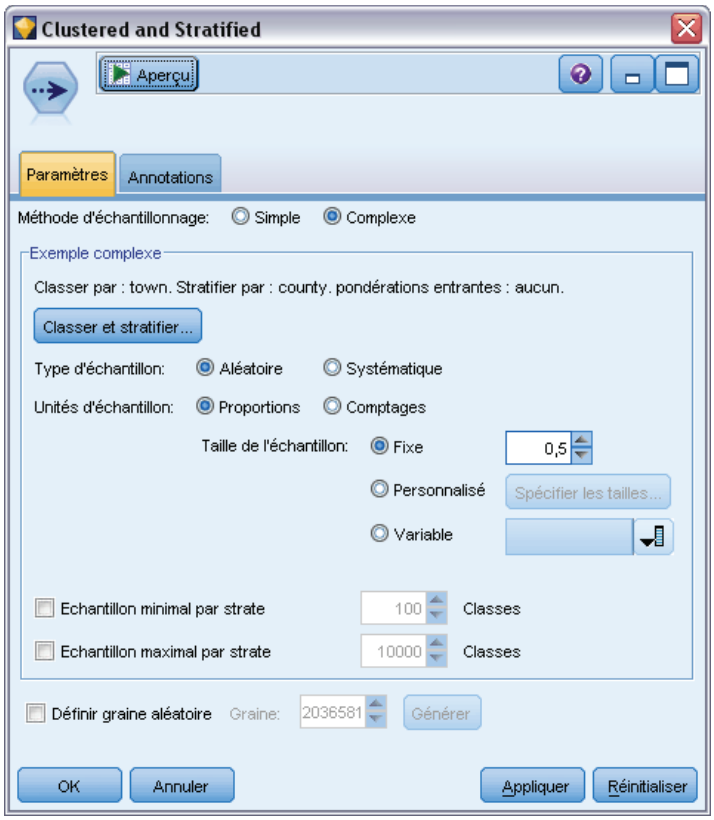
  Utiliser l'échantillonnage des niveaux de bloc (dans la base de données uniquement). Cette option n'est activée que si vous choisissez un échantillonnage de pourcentage aléatoire lors de l'exploration d'une base de données Oracle ou IBM DB2. Dans ce cas, l'échantillonnage des niveaux de bloc peut être plus efficace.

Taille maximale de l'échantillon. Indique le nombre maximum d'enregistrements à inclure dans l'échantillon. Cette option est redondante et par conséquent désactivée si Premiers et Inclure sont sélectionnés. Notez également que lorsqu'il est utilisé avec l'option % aléatoire, ce paramètre peut vous empêcher de sélectionner certains enregistrements. Par exemple, si vous disposez de 10 millions d'enregistrements dans votre ensemble de données et que vous sélectionnez 50 % d'enregistrements avec une taille maximale d'échantillon de 3 millions d'enregistrements, 50 % des 6 premiers millions d'enregistrements seront sélectionnés et les quatre millions d'enregistrements restants ne le seront pas. Pour éviter cette limitation, sélectionnez la méthode d'échantillonnage complexe et demandez un échantillon aléatoire de trois millions d'enregistrements sans définir une variable de classe ou de strate.

Options d'échantillonnage complexe

Les options d'échantillonnage complexe permettent d'affiner le contrôle de l'échantillon, notamment des échantillons en classe, stratifiés et pondérés avec d'autres options.

Figure 3-4
Options d'échantillonnage complexe



Classer et stratifier. Permet de définir des champs de classification, de stratification et de pondération d'entrée, si nécessaire. Pour plus d'informations, reportez-vous à la section Paramètres de classification et de stratification sur p. 77.

Type d'échantillon.

- **Aléatoire.** Sélectionne des classes ou des enregistrements dans chaque strate de manière aléatoire.
- **Systématique.** Sélectionne des enregistrements à une fréquence fixe. Cette option fonctionne comme la méthode *1 en n*, sauf que la position du premier enregistrement change en fonction d'une valeur de départ aléatoire. La valeur *n* est définie automatiquement en fonction de la taille ou de la proportion d'échantillonnage.

Unité d'échantillonnage. Vous pouvez sélectionner des proportions ou des nombres comme unité d'échantillonnage de base.

Taille de l'échantillon : Vous pouvez définir la taille d'échantillonnage de différentes manières :

- **Fixe :** Permet de définir la taille globale de l'échantillon sous la forme d'un nombre ou d'une proportion.

- **Personnalisée.** Permet de définir la taille d'échantillonnage de chaque sous-groupe ou strate. Cette option est disponible uniquement si un champ de stratification a été défini dans la sous-boîte de dialogue Classer et stratifier.
- **Variable :** Permet à l'utilisateur de sélectionner un champ qui définit la taille d'échantillonnage de chaque sous-groupe ou strate. Ce champ doit avoir la même valeur pour chaque enregistrement d'une strate. Si, par exemple, l'échantillon est stratifié en fonction du pays, tous les enregistrements ayant *county = Surrey* doivent avoir la même valeur. Le champ doit être numérique et sa valeur doit correspondre à l'unité d'échantillonnage sélectionnée. Pour les proportions, les valeurs doivent être supérieures à 0 et inférieures à 1. Pour les nombres, la valeur minimale est 1.

Echantillon minimum par strate. Définit le nombre minimum d'enregistrements (ou le nombre minimum de classes si un champ de classe est défini).

Echantillon maximum par strate. Définit le nombre maximum d'enregistrements ou de classes. Si vous sélectionnez cette option sans définir un champ de classe ou de strate, un échantillon aléatoire ou systématique de la taille définie est sélectionné.

Définir graine aléatoire. Lors de l'échantillonnage ou du partitionnement d'enregistrements en fonction d'un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée ou cliquez sur le bouton Générer pour générer automatiquement une valeur aléatoire. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Lorsque vous utilisez l'option Définir graine aléatoire avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la graine aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. Pour plus d'informations, reportez-vous à la section Noeud Trier sur p. 88.

Paramètres de classification et de stratification

La boîte de dialogue Classer et stratifier permet de sélectionner des champs de classe, de stratification et de pondération lors de la création d'n graphique d'échantillon complexe.

Figure 3-5
Paramètres de champ de classification et de stratification

Classes. Définit un champ catégoriel utilisé pour classer les enregistrements. Les enregistrements sont échantillonnés en fonction de leur appartenance aux classes, certaines classes étant incluses et d'autres exclues. Toutefois, si un enregistrement d'une classe est inclus, tous les enregistrements sont inclus. Lors de l'analyse des associations de produits d'un panier, par exemple, vous pouvez classer les articles en fonction de l'ID de transaction pour que tous les articles des transactions sélectionnées soient préservés. Au lieu d'échantillonner les enregistrements—ce qui détruirait les informations sur les articles vendus ensemble—vous pouvez échantillonner les transactions pour que tous les enregistrements des transactions sélectionnées soient préservés.

Stratifier par. Définit un champ catégoriel utilisé pour stratifier les enregistrements pour que les échantillons soient sélectionnés de manière indépendante dans les sous-groupes sans chevauchement de population, ou strates. Si vous sélectionnez un échantillon de 50 % stratifié en fonction du sexe, par exemple, deux échantillons de 50 % sont utilisés, un pour les hommes et un autre pour les femmes. Les strates, par exemple, peuvent être des groupes socio-économiques, des catégories d'emplois ou des groupes ethniques permettant de disposer de tailles d'échantillons adéquates pour les sous-groupes d'intérêt. S'il existe trois fois plus de femmes que d'hommes dans l'ensemble de données d'origine, ce rapport est conservé en échantillonnant séparément depuis chaque groupe. Vous pouvez également définir plusieurs champs de stratification (par exemple, échantillonnage de lignes de produits dans les régions ou vice versa).

Remarque : Si vous stratifiez les données en fonction d'un champ ayant des valeurs manquantes (valeurs nulles ou système manquantes, chaînes vides, espaces et blancs ou valeurs définies par l'utilisateur manquantes), vous ne pouvez pas définir des tailles d'échantillons personnalisées pour les strates. Si vous voulez utiliser des tailles d'échantillon personnalisées lors de la stratification en fonction d'un champ ayant des valeurs manquantes ou vides, vous devez les définir en amont.

Utiliser la pondération d'entrée. Définit un champ utilisé pour pondérer les enregistrements avant l'échantillonnage. Si, par exemple, le champ de pondération a des valeurs comprises entre 1 et 5, les enregistrements pondérés 5 ont cinq fois plus de chance d'être sélectionnés. Les valeurs de ce champ sont remplacées par les pondérations d'entrée finales générées par le noeud (voir le paragraphe suivant).

Nouvelle pondération de sortie. Définit le nom du champ où les pondérations finales sont écrites si aucun champ de pondération d'entrée n'est défini. (Si un champ de pondération d'entrée est défini, ses valeurs sont remplacées par les pondérations finales comme indiqué ci-dessus, et aucun champ de pondération de sortie distinct n'est créé.) Les valeurs de pondération de sortie indiquent le nombre d'enregistrements représentés par chaque enregistrement échantillonné dans les données d'origine. La somme des valeurs de pondération donne l'estimation de la taille d'échantillon. Si, par exemple, un échantillon aléatoire de 10 % est utilisé, la pondération de sortie est 10 pour tous les enregistrements, indiquant que chaque enregistrement échantillonné représente environ dix enregistrements dans les données d'origine. Dans un échantillon stratifié ou pondéré, les valeurs de pondération de sortie peuvent varier en fonction de la proportion d'échantillonnage de chaque strate.

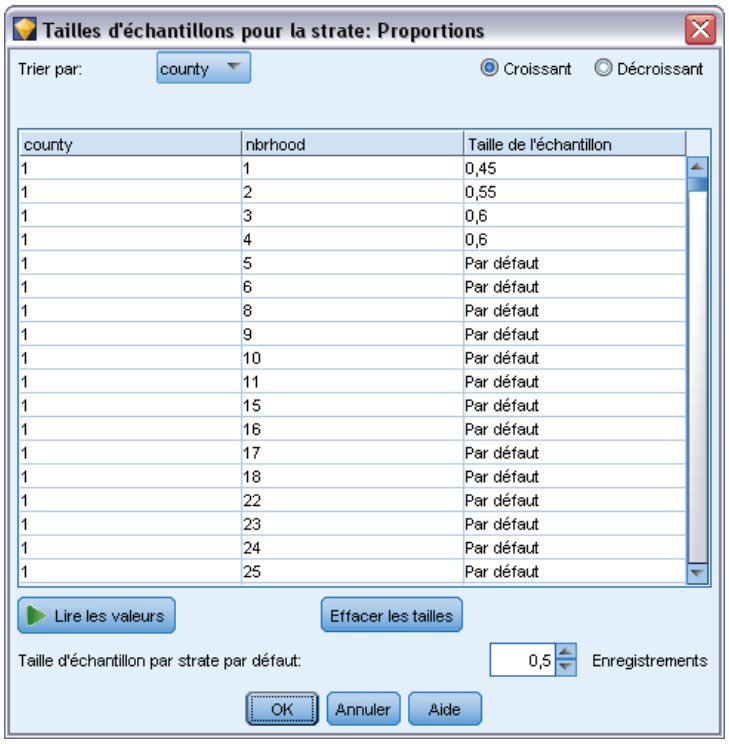
Commentaires

- L'échantillonnage en classe est utile si vous ne pouvez pas obtenir la liste complète de la population à échantillonner, mais pouvez obtenir les listes complètes de certains groupes ou classes. Il est également utilisé lorsqu'un échantillonnage aléatoire produit une liste de sujets de test difficile à contacter. Par exemple, il est plus simple de rendre visite à tous les fermiers d'un pays qu'à des fermiers dispersés dans le pays.
- Vous pouvez définir des champs de classification et de stratification pour échantillonner des classes de manière indépendante dans chaque strate. Par exemple, vous pouvez échantillonner les valeurs de propriétés stratifiées en fonction du pays et effectuer une classification en fonction de la ville dans chaque pays. Ainsi, vous pouvez créer un échantillon indépendant des villes dans chaque pays. Certaines villes seront incluses et d'autres pas, mais pour chaque ville incluse, toutes les propriétés de la ville seront incluses.
- Pour sélectionner un échantillon aléatoire d'unités dans chaque classe, vous pouvez enchaîner deux noeuds Echantillonner. Par exemple, vous pouvez échantillonner en premier les villes stratifiées en fonction du pays, comme indiqué ci-dessus, puis lier un second noeud Echantillonner et sélectionner *ville* comme champ de stratification, ce qui permet d'échantillonner une proportion d'enregistrements dans chaque ville.
- Si une combinaison de champs est nécessaire pour identifier de manière unique les classes, vous pouvez générer un nouveau champ en utilisant un noeud Calculer. Si, par exemple, plusieurs boutiques utilisent le même système de numérotation des transactions, vous pouvez calculer un nouveau champ qui concatène les ID de boutique et de transaction.

Tailles d'échantillons pour la strate

Lorsque vous créez un échantillon stratifié, l'option par défaut consiste à échantillonner la même proportion d'enregistrements ou de classes pour chaque strate. Si un groupe contient plus de membres qu'un autre par un facteur de 3, par exemple, vous voulez généralement préserver le même rapport dans l'échantillon. Si tel n'est pas le cas, vous pouvez définir la taille d'échantillon séparément pour chaque strate.

Figure 3-6
Définition des tailles d'échantillons des strates



La boîte de dialogue Tailles d'échantillons des strates contient les valeurs du champ de stratification permettant de remplacer la valeur par défaut de la strate. S vous sélectionnez plusieurs champs de stratification, toutes les combinaisons de valeurs possibles sont affichées pour vous permettre de définir la taille de chaque groupe ethnique dans chaque ville, par exemple, ou chaque ville dans chaque pays. Les tailles sont définies sous forme de proportions ou de nombres, tel que défini par le paramètre en cours dans le noeud Echantillonner.

Pour définir les tailles d'échantillons des strates

- Dans le noeud Echantillonner, sélectionnez Complexe et un ou plusieurs champs de stratification. Pour plus d'informations, reportez-vous à la section Paramètres de classification et de stratification sur p. 77.

- Sélectionnez Personnaliser et Définir des tailles.

- Dans la boîte de dialogue Tailles d'échantillons des strates, cliquez sur le bouton Lire les valeurs dans la partie inférieure gauche de l'écran. Si nécessaire, vous pouvez être amené à instancier des valeurs dans une source en amont ou un noeud Typer. Pour plus d'informations, reportez-vous à la section Qu'est-ce que l'instanciation ? dans le chapitre 4 sur p. 142.

- Cliquez sur une ligne pour remplacer la taille par défaut de la strate.

Remarques sur les tailles d'échantillons

Des tailles d'échantillons personnalisées peuvent s'avérer utiles si des strates différentes ont des variances différentes, par exemple, pour que les tailles d'échantillons soient proportionnelles à l'écart-type. (Si les observations dans la strate varie davantage, vous devez les échantillonner davantage pour obtenir un échantillon représentatif.) Ou bien, si une strate est petite, vous pouvez utiliser une proportion d'échantillonnage supérieure pour inclure un nombre minimum d'observations.

Remarque : Si vous stratifiez les données en fonction d'un champ ayant des valeurs manquantes (valeurs nulles ou système manquantes, chaînes vides, espaces et blancs ou valeurs définies par l'utilisateur manquantes), vous ne pouvez pas définir des tailles d'échantillons personnalisées pour les strates. Si vous voulez utiliser des tailles d'échantillon personnalisées lors de la stratification en fonction d'un champ ayant des valeurs manquantes ou vides, vous devez les définir en amont.

Noeud Equilibrer

Le noeud Equilibrer permet de corriger les déséquilibres dans les ensembles de données, de sorte que ceux-ci soient conformes aux critères de test spécifiés. Supposons, par exemple, qu'un ensemble de données présente uniquement deux valeurs, *faible* ou *élevée*, et que 90 % des occurrences sont *faibles* tandis que seulement 10 % des occurrences sont *élevées*. De nombreuses techniques de modélisation ne parviennent pas à gérer ce type de données biaisées car elles ont tendance à ne retenir que la valeur *faible* et à ignorer la valeur *élevée*, qui est plus rare. Si les données sont équilibrées, avec des nombres approximativement égaux de valeurs *faibles* et *élevées*, les modèles pourront plus facilement trouver des tendances qui distinguent les deux groupes. Dans ce cas, vous pouvez utiliser un noeud Equilibrer pour créer une directive qui diminue le nombre d'observations de la valeur *faible*.

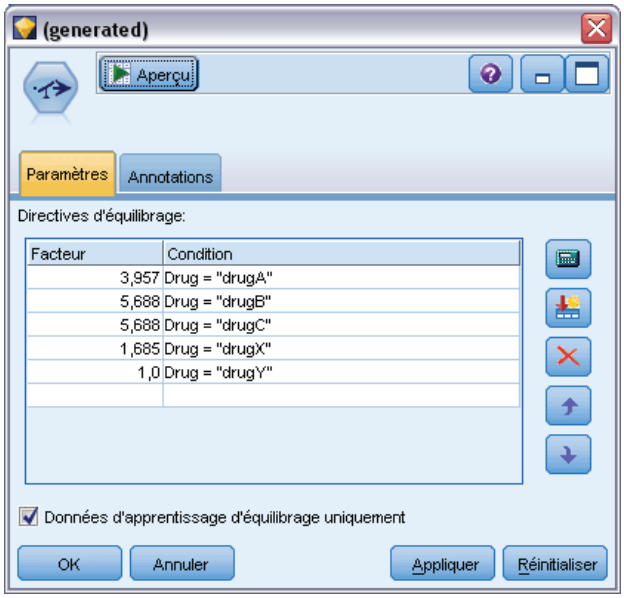
L'équilibrage est obtenu en dupliquant et en supprimant des enregistrements en fonction de conditions spécifiées. Les enregistrements pour lesquels aucune condition n'est vérifiée sont toujours ignorés. Dans la mesure où ce processus implique la duplication et/ou l'exclusion d'enregistrements, la séquence d'origine de vos données est perdue au cours d'opérations effectuées en aval. Veillez à calculer toutes les valeurs dépendant directement de la séquence de vos données avant d'ajouter un noeud Equilibrer au flux de données.

Remarque : les noeuds Equilibrer peuvent être automatiquement générés à partir de graphiques Proportion et Histogramme. Vous pouvez, par exemple, équilibrer les données pour indiquer les proportions égales dans toutes les catégories d'un champ catégoriel, comme indiqué dans une courbe de distribution.

Exemple : Lors de la création d'un flux RFM pour identifier les clients récents qui ont répondu positivement à des campagnes de publicité antérieures, le service Marketing d'une société utilise un noeud Equilibrer pour équilibrer les différences entre les réponses vraies et les réponses fausses dans les données.

Définition des options du noeud Equilibrer

Figure 3-7
Paramètres de noeud Equilibrer



Règles d'équilibrage d'enregistrements. Affiche les règles d'équilibrage en cours. Chaque directive inclut à la fois un facteur et une condition qui indiquent au logiciel « d'augmenter la proportion d'enregistrements d'un facteur spécifié lorsque la condition est vraie ». Un facteur inférieur à 1,0 signifie que la proportion d'enregistrements va être réduite. Par exemple, si vous souhaitez réduire le nombre d'enregistrements pour lesquels le médicament Y est utilisé, vous pouvez créer une règle d'équilibrage avec un facteur de 0,7 et la condition Médicament = "médY". Cette directive indique que le nombre d'enregistrements pour lesquels le médicament Y est utilisé sera réduit de 70 % pour toutes les opérations en aval.

Remarque : les facteurs d'équilibrage de la réduction peuvent avoir quatre décimales. Les facteurs définis en dessous de 0,0001 donnent des résultats erronés car ils ne sont pas calculés correctement.

- **Créez des conditions** en cliquant sur le bouton situé à droite du champ de texte. Une ligne vide est insérée ; elle vous permet de saisir les nouvelles conditions. Pour créer une expression CLEM pour la condition, cliquez sur le bouton Générateur de formules.
- **Supprimez des directives** à l'aide du bouton de suppression rouge.
- **Triez les directives** à l'aide des flèches vers le haut ou vers le bas.

Equilibrer uniquement les données d'apprentissage. Si un champ de partition figure dans le flux, cette option équilibre les données dans la partition d'apprentissage uniquement. Cela peut s'avérer utile, notamment, si vous générez des scores de propension ajustée qui nécessitent une partition de test ou de validation non équilibrée. Si aucun champ de partition ne figure dans le flux (ou si plusieurs champs de partition sont définis), cette option est ignorée et toutes les données sont équilibrées.

Noeud Agréger

L'agrégation est une tâche de préparation des données fréquemment utilisée pour réduire la taille d'un ensemble de données. Avant d'effectuer l'agrégation, vous devez prendre le temps de nettoyer les données, en vous concentrant notamment sur les valeurs manquantes. Une fois l'agrégation terminée, les informations éventuellement utiles concernant les valeurs manquantes risquent d'être perdues.

Les noeuds Agréger permettent de remplacer une séquence d'enregistrements d'entrée par des enregistrements de sortie récapitulatifs et agrégés. Prenons par exemple les enregistrements de ventes d'entrée suivants :

Age	Sexe	Région	Filiale	Ventes
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

Vous pouvez utiliser les champs-clés *Sexe* et *Région* pour agréger ces enregistrements. Agrégez ensuite *Age* avec le mode Moyenne et *Ventes* avec le mode Somme. Sélectionnez Inclure le comptage des enregistrements dans le champ dans la boîte de dialogue du noeud Agréger. Vous obtenez alors la sortie agrégée suivante :

Âge (moyenne)	Sexe	Région	Ventes (somme)	Nombre d'enregistrements
35.5	F	N	25	4
29	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

Ceci vous apprend par exemple que l'âge moyen des quatre membres féminins de l'équipe de vente dans la région nord est de 35,5 ans et que le montant total de leurs ventes était de 25 unités.

Remarque : Les champs comme *Filiale* sont automatiquement exclus lorsqu'aucun mode d'agrégation n'est spécifié.

Définition des options du noeud Agréger

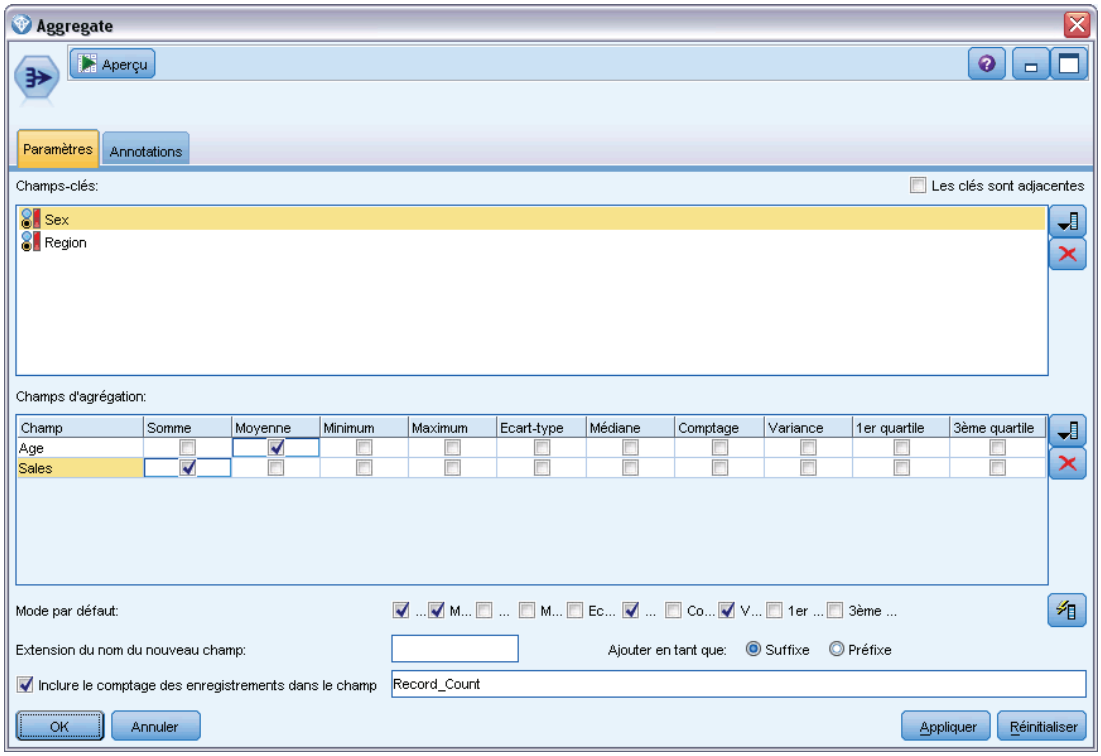
Dans le noeud Agréger, vous spécifiez ce qui suit.

- Un ou plusieurs champs clés à utiliser comme catégories d'agrégation

- Un ou plusieurs champs agrégés pour lesquels calculer les valeurs agrégées
- Un ou plusieurs modes d'agrégation (types d'agrégation) de sortie pour chaque champ agrégé

Vous pouvez également spécifier les modes d'agrégation par défaut à utiliser pour les nouveaux champs ajoutés.

Figure 3-8
Boîte de dialogue du noeud Agréger



Champs-clés. Affiche les champs qui peuvent être utilisés comme catégories pour l'agrégation. Les champs continus (numériques) et catégoriels sont autorisés en tant que clés. Si vous sélectionnez plusieurs champs-clés, les valeurs seront combinées de façon à produire une valeur-clé qui sera utilisée pour l'agrégation des enregistrements. Pour chaque champ-clé unique, un enregistrement agrégé est généré. Par exemple, si vous avez choisi les champs-clés *Sexe* et *Région*, chaque combinaison unique de *M* et *F* avec les régions *N* et *S* (quatre combinaisons uniques) est associée à un enregistrement agrégé. Pour ajouter un champ-clé, utilisez le sélecteur de champs situé à droite dans la fenêtre.

Les clés sont adjacentes. Sélectionnez cette option si vous savez que tous les enregistrements ayant les mêmes valeurs de clés sont regroupés dans l'entrée (si, par exemple, l'entrée est triée sur les champs de clé). Ainsi, vous améliorez les performances.

Champs d'agrégation. Affiche les champs dont les valeurs seront agrégées, ainsi que les modes d'agrégation sélectionnés. Pour ajouter des champs à cette liste, utilisez le sélecteur de champs situé à droite. Les modes d'agrégation suivants sont disponibles.

Remarque : Certains modes ne s'appliquent pas aux champs non numériques (par exemple, Somme pour un champ de date/heure). Les modes qui ne peuvent pas être utilisés avec un champ agrégé sélectionné sont désactivés.

- **Somme :** Cochez cette case pour obtenir les valeurs additionnées de chaque combinaison de champs-clés. La somme ou le total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- **Moyenne.** Cochez cette case pour obtenir les valeurs moyennes de chaque combinaison de champs-clés. La moyenne est une mesure de tendance centrale et est la moyenne arithmétique (la somme divisée par le nombre de cas).
- **Minimum** Cochez cette case pour obtenir les valeurs minimales de chaque combinaison de champs-clés.
- **Maximum** Cochez cette case pour obtenir les valeurs maximales de chaque combinaison de champs-clés.
- **Ecart-type.** Cochez cette case pour obtenir l'écart-type de chaque combinaison de champs-clés. L'écart-type est la mesure de la dispersion des valeurs autour de la moyenne, égale à la racine carrée de la variance.
- **Médiane.** Cochez cette case pour obtenir les valeurs médianes de chaque combinaison de champs-clés. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées. Elle est également nommée 50ème centile ou 2ème quartile.
- **Effectifs.** Cochez cette case pour obtenir le nombre de valeurs non nulles pour chaque combinaison de champs-clés.
- **Variance.** Cochez cette case pour obtenir les valeurs de variance de chaque combinaison de champs-clés. La variance est une mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un.
- **1er quartile.** Cochez cette case pour obtenir les valeurs du 1er quartile (25ème centile) pour chaque combinaison de champs-clés.
- **3ème quartile.** Cochez cette case pour obtenir les valeurs du 3ème quartile (75ème centile) pour chaque combinaison de champs-clés.

Mode par défaut. Indiquez le mode d'agrégation par défaut à utiliser pour les nouveaux champs ajoutés. Si vous utilisez souvent le même type d'agrégation, sélectionnez un ou plusieurs modes, et utilisez le bouton Appliquer partout situé à droite pour appliquer les modes sélectionnés à tous les champs répertoriés.

Extension du nom du nouveau champ. Permet d'ajouter un suffixe ou un préfixe, tel que « 1 » ou « nouveau » pour dupliquer les champs agrégés. Par exemple, l'agrégation des valeurs minimales du champ *Age* produit un champ appelé *Age_Min_1* si vous avez sélectionné l'option du suffixe et spécifié « 1 » comme extension. *Remarque* : les extensions d'agrégation telles que *_Min* ou *Max_* sont automatiquement ajoutées au nouveau champ, indiquant ainsi le type d'agrégation exécuté. Sélectionnez Suffixe ou Préfixe pour indiquer le style d'extension voulu.

Inclure le comptage des enregistrements dans le champ. Permet d'inclure un champ supplémentaire dans chaque enregistrement intitulé *Record_Count*, par défaut. Pour chaque enregistrement de sortie, ce champ indique le nombre d'enregistrements d'entrée qui ont été agrégés. Indiquez le nom de votre choix pour ce champ dans le champ d'édition.

Remarque : les valeurs système nulles sont exclues lors du calcul de l'agrégation, mais sont incluses dans le nombre d'enregistrements. En revanche, les valeurs non renseignées sont incluses dans l'agrégation et dans le nombre d'enregistrements. Pour exclure les valeurs non renseignées, vous pouvez utiliser un noeud Remplacer pour remplacer les valeurs non renseignées par des valeurs nulles. Vous pouvez également supprimer les blancs à l'aide d'un noeud Sélectionner.

Performances

L'activation du traitement parallèle peut profiter aux opérations d'agrégation.

Noeud Agréger RFM

Le noeud Agréger RFM (Recency, Frequency, Monetary) permet d'utiliser les données historiques des transactions des clients, de supprimer les données inutiles et de combiner toutes leurs données de transaction restantes dans une seule ligne, en utilisant leur ID de client unique comme clé, qui indique leur dernier contact avec vous (Recency), le nombre de transactions qu'ils ont effectuées (Frequency) et la valeur totale des transactions (Monetary).

Avant d'effectuer une agrégation, vous devez nettoyer les données en vous concentrant notamment sur les valeurs manquantes.

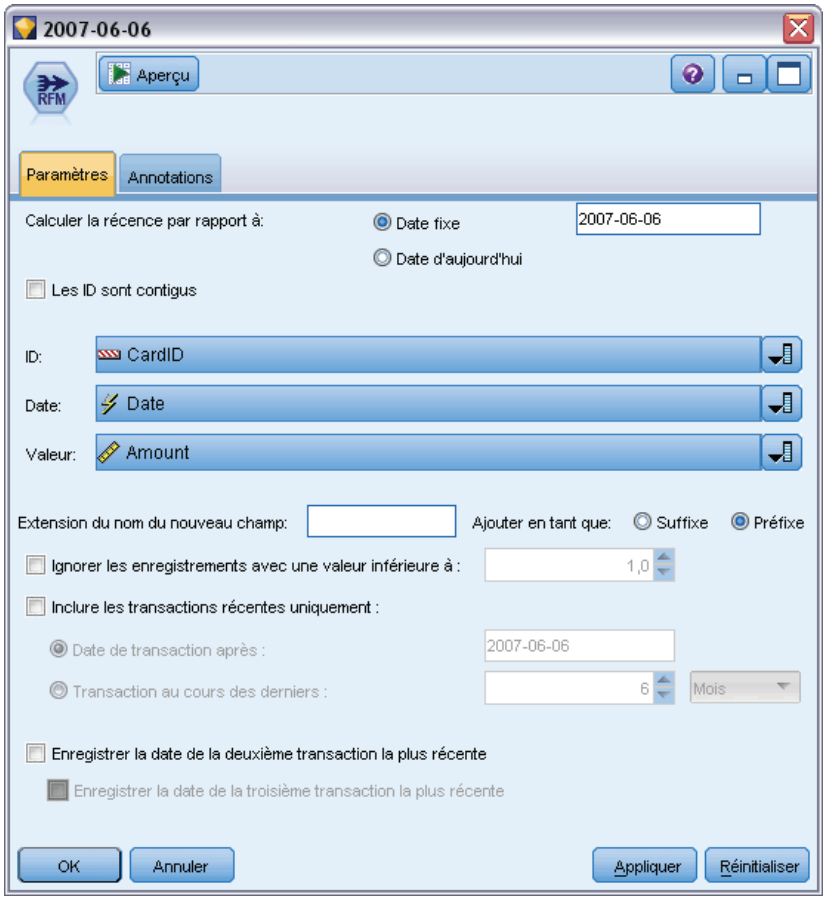
Après avoir identifié et transformé les données en utilisant le noeud Agréger RFM, vous pouvez utiliser un noeud Analyse RFM pour effectuer d'autres analyses. Pour plus d'informations, reportez-vous à la section Noeud Analyse RFM dans le chapitre 4 sur p. 203.

Notez qu'une fois que le fichier de données a été exécuté via le noeud Agréger RFM, il ne dispose plus de valeurs cible. Par conséquent, pour pouvoir l'utiliser comme entrée pour effectuer d'autres analyses prédictives avec des noeuds de modélisation, tels que C5.0 ou CHAID, vous devez le fusionner avec d'autres données client (par exemple, en faisant correspondre les ID client). Pour plus d'informations, reportez-vous à la section Noeud Fusionner sur p. 90.

Les noeuds Agréger RFM et Analyse RFM de IBM® SPSS® Modeler sont configurés pour utiliser la création d'intervalles indépendants ; en d'autres termes, ils classent et espacent les données sur chaque mesure de valeur de proximité dans le temps, d'effectif et de valeur monétaire, sans tenir compte de leur valeur ni des deux autres mesures.

Définition des options du noeud Agréger RFM

Figure 3-9
Paramètres d'agrégation RFM



Calculer la récence par rapport à. Définissez la date à partir de laquelle la récence des transactions sera calculée. Il peut s'agir d'une date fixe que vous entrez ou de la date du jour, telle que définie par votre système. La date du jour est entrée par défaut et elle est mise à jour automatiquement lors de l'exécution du noeud.

Les ID sont contigus. Si vos données sont prétriées de façon à ce que tous les enregistrements avec le même ID apparaissent ensemble dans le flux de données, sélectionnez cette option pour accélérer le traitement. Si vos données ne sont pas pré-triées (ou si vous n'en êtes pas certain), ne sélectionnez pas cette option ; le noeud triera automatiquement les données.

ID. Sélectionnez le champ à utiliser pour identifier le client et ses transactions. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite.

Date. Sélectionnez le champ de date à utiliser pour calculer la récence. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite.

Notez que cela nécessite d'utiliser un champ avec l'enregistrement de date, ou horodatage, dans le format approprié comme entrée. Par exemple, si vous disposez d'un champ de chaîne contenant des valeurs telles que *Jan 2007*, *Fév 2007*, etc., vous pouvez le convertir en champ de date en

utilisant un noeud Remplacer et la fonction to_date(). Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer dans le chapitre 4 sur p. 180.

Valeur. Sélectionnez le champ à utiliser pour calculer la valeur monétaire totale des transactions du client. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite. *Remarque* : Il doit s'agir d'une valeur numérique.

Extension du nom du nouveau champ. Ajoutez un suffixe ou un préfixe, tel que "12_month", dans les nouveaux champs générés de récence, de fréquence et monétaire. Sélectionnez Suffixe ou Préfixe pour indiquer le style d'extension voulu. Par exemple, cela peut s'avérer utile pour examiner différentes périodes.

Supprimer les enregistrements ayant des valeurs inférieures. Si nécessaire, vous pouvez définir une valeur minimale en dessous de laquelle les informations de transaction ne sont pas utilisées pour calculer les totaux RFM. L'unité de valeur fait référence au champ Valeur sélectionné.

Inclure les dernières transactions uniquement. Si vous analysez une base de données volumineuse, vous pouvez indiquer que seuls les derniers enregistrements doivent être utilisés. Vous pouvez utiliser les données enregistrées après une date donnée ou au cours d'une période récente :

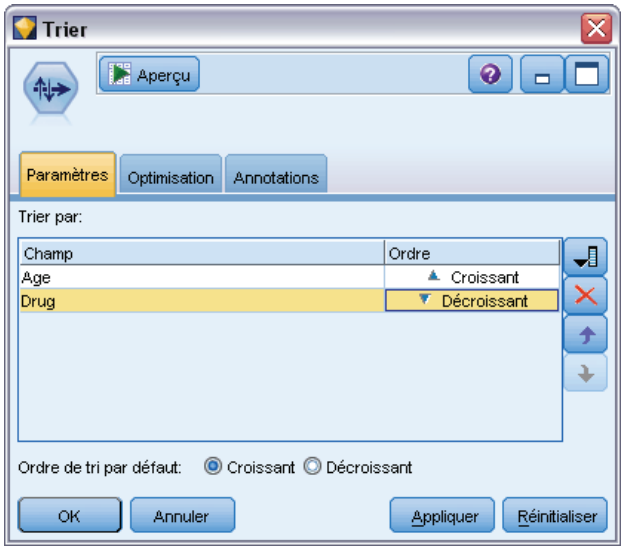
- **Date de transaction après.** Définissez la date de transaction après laquelle les enregistrements seront inclus dans l'analyse.
- **Transaction dans la dernière.** Définissez le nombre et le type des périodes (jours, semaines, mois ou années) par rapport à la date Calculer la récence par rapport à après laquelle les enregistrements seront inclus dans l'analyse.

Enregistrer la date de la deuxième transaction la plus récente. Si vous voulez connaître la date de la deuxième transaction la plus récente de chaque client, cochez cette case. En outre, vous pouvez cochez la case Enregistrer la date de la troisème transaction la plus récente. Elle permet, par exemple, d'identifier les clients qui peuvent avoir exécuté un grand nombre de transactions il y a longtemps, mais une seule transaction récente.

Noeud Trier

Le noeud Trier permet de classer les enregistrements par ordre croissant ou décroissant, en fonction de la valeur d'un ou de plusieurs champs. Les noeuds Trier sont souvent utilisés pour visualiser et sélectionner les enregistrements contenant les valeurs de données les plus répandues. La première opération consiste à agréger les données à l'aide d'un noeud Agréger, puis à utiliser un noeud Trier pour les trier par ordre décroissant de nombre d'enregistrements. Les résultats apparaissent dans un tableau, dans lequel vous pouvez examiner les données et les manipuler (pour sélectionner les enregistrements relatifs aux dix clients les plus fidèles, par exemple).

Figure 3-10
Boîte de dialogue du noeud Trier



Trier par. Tous les champs sélectionnés en tant que clé de tri apparaissent dans un tableau. Le tri est plus efficace lorsque le champ-clé est numérique.

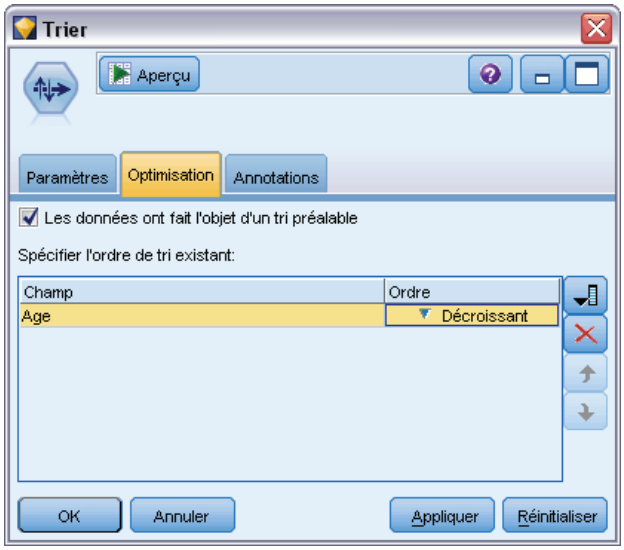
- **Ajoutez des champs** à cette liste en utilisant le sélecteur de champs situé à droite.
- **Sélectionnez un ordre** en cliquant sur la flèche Croissant ou Décroissant de la colonne *Ordre* du tableau.
- **Supprimez des champs** à l'aide du bouton de suppression rouge.
- **Triez les directives** à l'aide des flèches vers le haut ou vers le bas.

Ordre de tri par défaut.Sélectionnez Croissant ou Décroissant pour définir l'ordre de tri par défaut lorsque de nouveaux champs sont ajoutés.

Paramètres d'optimisation du tri

Si vous savez que les données avec lesquelles vous travaillez ont déjà été triées en fonction de certains champs-clés, vous pouvez indiquer les champs ayant déjà fait l'objet d'un tri ; le système peut ainsi trier le reste des données de manière plus efficace. Il se peut, par exemple, que vous souhaitiez effectuer un tri sur les champs *Age* (décroissant) et *Médicament* (croissant) mais que vous sachiez que les données ont déjà été triées en fonction du champ *Age* (décroissant).

Figure 3-11
Paramètres d'optimisation



Les données ont fait l'objet d'un tri préalable. Indique si les données sont déjà triées en fonction d'un ou de plusieurs champs.

Spécifier l'ordre de tri existant. Indiquez les champs déjà triés. Dans la boîte de dialogue Sélectionner les champs, ajoutez des champs à la liste. Dans la colonne *Ordre*, précisez si chaque champ est trié dans l'ordre croissant ou décroissant. Si vous entrez ici plusieurs champs, veillez à les répertorier dans le bon ordre. Cliquez sur les flèches situées à droite de la liste pour trier les champs dans l'ordre souhaité. Si vous définissez un ordre de tri existant incorrect, un message d'erreur apparaît lors de l'exécution du flux, indiquant le numéro d'enregistrement au niveau duquel le tri n'est pas conforme à ce que vous avez indiqué.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de tri.

Noeud Fusionner

Le noeud Fusionner permet de créer à partir de plusieurs enregistrements d'entrée un seul enregistrement de sortie contenant la totalité ou une partie des champs d'entrée. Cette opération permet notamment de fusionner des données provenant de différentes sources, telles que les données client internes et les données démographiques acquises. Vous pouvez fusionner des données des manières suivantes :

- La **fusion par ordre** concatène les enregistrements correspondants issus de toutes les sources dans l'ordre d'entrée, jusqu'à ce que la plus petite source de données soit épuisée. Si vous utilisez cette option, il est important d'avoir trié vos données à l'aide d'un noeud Trier.
- La **fusion à l'aide d'un champ-clé**, tel que *ID client*, vous permet de spécifier le mode de mise en correspondance des enregistrements d'une source de données avec ceux d'une ou d'autres sources de données. Plusieurs types de jointures sont disponibles, notamment les

jointures interne, externe complète, externe partielle et anti-jointure. Pour plus d'informations, reportez-vous à la section Types de jointure sur p. 91.

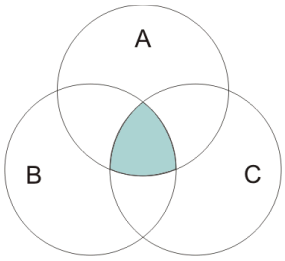
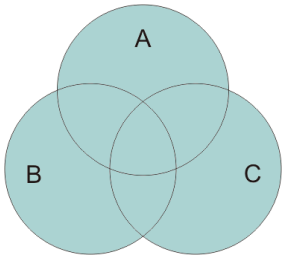
- **Fusion conditionnelle** signifie que vous pouvez spécifier une condition à satisfaire qui détermine si la fusion a lieu ou pas. La condition peut être spécifiée directement dans le noeud, ou il est possible de la générer à l'aide du Générateur de formules.

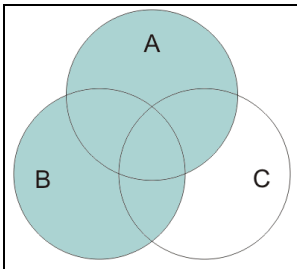
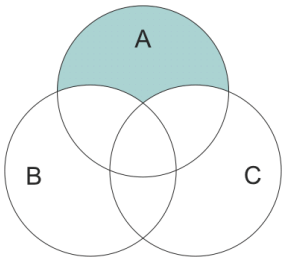
Types de jointure

Lorsque vous utilisez un champ-clé pour la fusion des données, prenez le temps nécessaire pour choisir les enregistrements à inclure et à exclure. Il existe diverses jointures qui sont présentées en détail ci-dessous.

Les deux principaux types de jointure sont appelés jointures internes et jointures externes. Ces méthodes sont souvent utilisées pour fusionner des tables à partir d'ensembles de données associés, sur la base des valeurs communes d'un champ-clé, tel que *ID client*. Les jointures internes permettent d'obtenir des fusions « propres », ainsi qu'un ensemble de données de sortie n'incluant que les enregistrements complets. Les jointures externes comprennent également des enregistrements complets issus des données fusionnées, mais elles vous permettent également d'inclure des données uniques provenant d'une ou de plusieurs tables d'entrée.

Les types de jointures autorisés sont décrits en détail ci-dessous.

	Une **jointure interne** inclut uniquement les enregistrements dans lesquels une valeur pour le champ-clé est commune à toutes les tables d'entrée. Cela signifie que les enregistrements sans correspondance ne sont pas inclus dans l'ensemble de données de sortie.
	Une **jointure externe complète** comprend tous les enregistrements, ceux qui ont une correspondance comme ceux qui n'en ont pas, des tables d'entrée. Les jointures externes gauche et droite sont appelées jointures externes partielles et sont décrites ci-dessous.

	Une **jointure externe partielle** inclut tous les enregistrements mis en correspondance à l'aide du champ-clé, ainsi que les enregistrements sans correspondance issus des tables spécifiées. (Autrement dit, elle inclut tous les enregistrements de certaines tables et uniquement les enregistrements correspondants d'autres tables.) Les tables (par exemple, A et B) peuvent être sélectionnées pour être incluses dans la jointure externe à l'aide du bouton Sélectionner de l'onglet Fusionner. Les jointures partielles sont également appelées jointures externes gauche ou droite lorsque deux tables seulement sont fusionnées. IBM® SPSS® Modeler autorisant la fusion de plusieurs tables, cette jointure est appelée une jointure externe partielle.
	Une **anti-jointure** ne comprend que les enregistrements sans correspondance de la première table d'entrée (table A dans cet exemple). Ce type de jointure est le contraire de la jointure interne ; elle n'inclut pas les enregistrements complets dans l'ensemble de données de sortie.

Par exemple, si un ensemble de données contient des informations sur des fermes et qu'un autre comporte des déclarations de sinistre relatives aux fermes, vous pouvez mettre en correspondance les enregistrements de la première source et ceux de la seconde à l'aide des options de fusion.

Pour déterminer si un client inclus dans cet exemple de fermes a émis une déclaration de sinistre, utilisez l'option de jointure interne pour renvoyer la liste des correspondances de tous les ID de ces deux ensembles de données.

Figure 3-12
Exemple de sortie pour une fusion réalisée par jointure interne

	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...

L'option de jointure externe complète permet de renvoyer à partir des tables d'entrée les enregistrements avec et sans correspondance. La valeur manquante ($null$) du système est utilisée pour les valeurs incomplètes.

Figure 3-13
Exemple de sortie pour une fusion réalisée par jointure externe complète

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.10
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

Une jointure externe partielle inclut tous les enregistrements mis en correspondance à l'aide du champ-clé, ainsi que les enregistrements sans correspondance issus des tables spécifiées. Le tableau affiche tous les enregistrements mis en correspondance à partir du champ d'ID, ainsi que ceux mis en correspondance à partir du premier ensemble de données.

Figure 3-14
Exemple de sortie pour une fusion réalisée par jointure externe partielle

	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

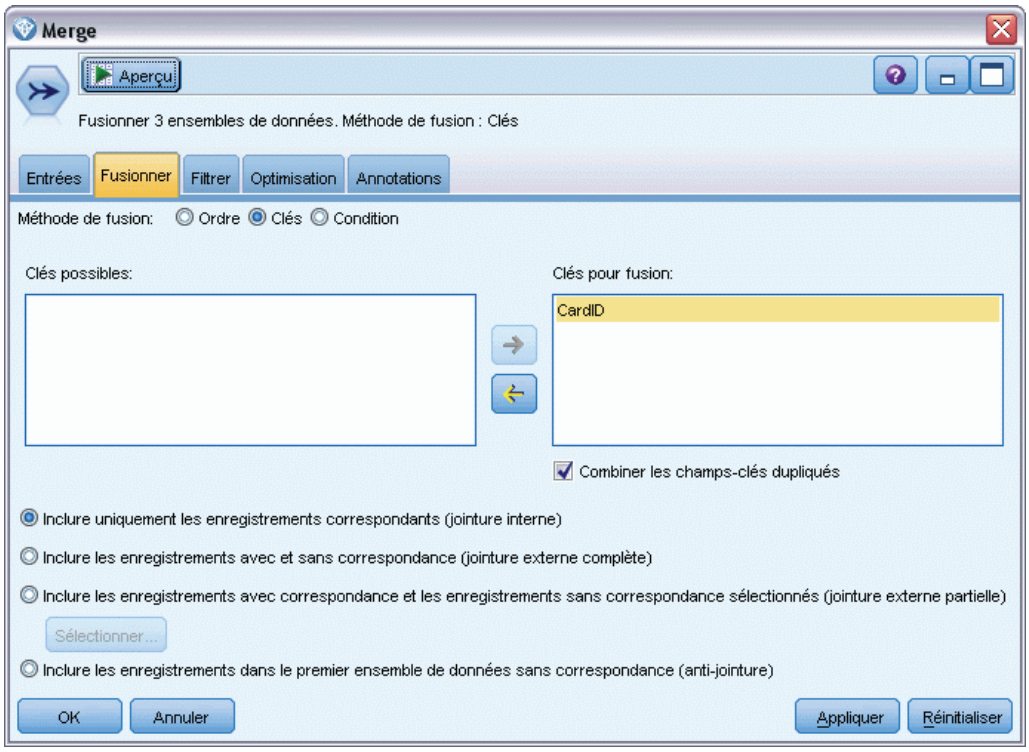
Si vous utilisez l'option d'anti-jointure, la table ne renvoie que les enregistrements sans correspondance issus de la première table d'entrée.

Figure 3-15
Exemple de sortie pour une fusion réalisée par anti-jointure

	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

Spécification d'une méthode de fusion et des clés

Figure 3-16
Utilisation de l'onglet Fusionner pour définir les options de la méthode de fusion



Méthode de fusion. Sélectionnez Ordre ou Clés pour spécifier la méthode de fusion des enregistrements. Lorsque vous sélectionnez Clés, la partie inférieure de la boîte de dialogue est activée.

- **Ordre :** Fusionne les enregistrements dans l'ordre de sorte que le *n*ième enregistrement de chaque entrée soit fusionné pour générer le *n*ième enregistrement de sortie. Lorsqu'un enregistrement n'a plus d'enregistrement d'entrée correspondant, la création d'enregistrements de sortie s'arrête. Ainsi, le nombre d'enregistrements produits est égal au nombre d'enregistrements du plus petit ensemble de données.
- **Clés.** Utilise un champ-clé, tel que *ID transaction*, pour fusionner les enregistrements ayant une valeur identique dans ce champ. Cette opération est équivalente à une « équi-jointure » de base de données. S'il existe plusieurs occurrences d'une valeur-clé, toutes les combinaisons possibles sont renvoyées. Par exemple, si des enregistrements avec la même valeur de champ-clé *A* contiennent des valeurs *B*, *C* et *D* dans d'autres champs, les champs fusionnés produiront un enregistrement distinct pour chaque combinaison de *A* avec la valeur *B*, de *A* avec la valeur *C*, et de *A* avec la valeur *D*.

  Remarque : les valeurs nulles ne sont pas considérées comme identiques dans la méthode de fusion par clés et ne sont pas regroupées.
- **Condition.** Utilisez cette option pour spécifier une condition à la fusion. Pour plus d'informations, reportez-vous à la section Spécification de conditions pour une fusion sur p. 95.

Clés possibles. Répertorie uniquement les champs avec des noms de champs exactement semblables dans toutes les sources de données d'entrées. Sélectionnez un champ dans la liste et utilisez la flèche pour l'ajouter en tant que champ-clé pour la fusion des enregistrements. Vous pouvez utiliser plusieurs champs-clés. Vous pouvez renommer les champs d'entrées qui ne correspondent pas à l'aide du noeud Filtre ou de l'onglet Filtre d'un noeud source.

Clés pour fusion. Affiche tous les champs utilisés pour fusionner les enregistrements de toutes les sources de données d'entrée, sur la base des valeurs des champs-clés. Pour supprimer une clé de la liste, sélectionnez-la et utilisez la flèche pour la renvoyer dans la liste Clés possibles. Lorsque plusieurs champs-clés sont sélectionnés, l'option ci-dessous est activée.

Combiner les champs-clés dupliqués. Lorsque plusieurs champs-clés sont sélectionnés, cette option garantit qu'un seul champ de sortie porte ce nom. Cette option est activée par défaut, excepté lorsque des flux ont été importés de versions antérieures de IBM® SPSS® Modeler. Lorsque cette option est désactivée, les champs-clés en double doivent être renommés ou exclus à l'aide de l'onglet Filtrer de la boîte de dialogue du noeud Fusionner.

Inclure uniquement les enregistrements correspondants (jointure interne). Permet de fusionner uniquement les enregistrements complets.

Inclure les enregistrements avec et sans correspondance (jointure externe complète). Permet d'effectuer une « jointure externe complète ». Cela signifie que même si les valeurs du champ-clé ne sont pas présentes dans toutes les tables d'entrée, les enregistrements incomplets sont néanmoins conservés. La valeur indéfinie ($null$) est ajoutée au champ-clé et incluse dans l'enregistrement de sortie.

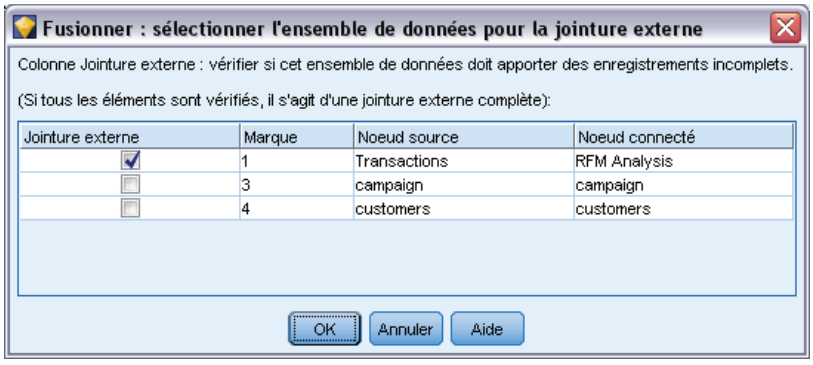
Inclure les enregistrements avec correspondance et les enregistrements sans correspondance sélectionnés (jointure externe partielle). Permet d'effectuer une « jointure externe partielle » des tables que vous sélectionnez dans une sous-boîte de dialogue. Cliquez sur Sélectionner pour indiquer les tables pour lesquelles des enregistrements incomplets seront conservés lors de la fusion.

Inclure les enregistrements dans le premier ensemble de données sans correspondance (anti-jointure). Permet d'effectuer un type d'« anti-jointure », où seuls les enregistrements sans correspondance du premier ensemble de données sont transmis au flux en aval. Vous pouvez indiquer l'ordre des ensembles de données d'entrée à l'aide des flèches de l'onglet Entrées. Ce type de jointure n'inclut pas les enregistrements complets dans l'ensemble de données de sortie. Pour plus d'informations, reportez-vous à la section Types de jointure sur p. 91.

Sélection de données pour des jointures partielles

Pour une jointure externe partielle, vous devez sélectionner les tables pour lesquelles des enregistrements incomplets seront conservés. Par exemple, vous pouvez conserver tous les enregistrements d'une table Client et ne conserver que les enregistrements avec correspondance de la table Prêt hypothécaire.

Figure 3-17
Sélection de données pour une jointure partielle externe

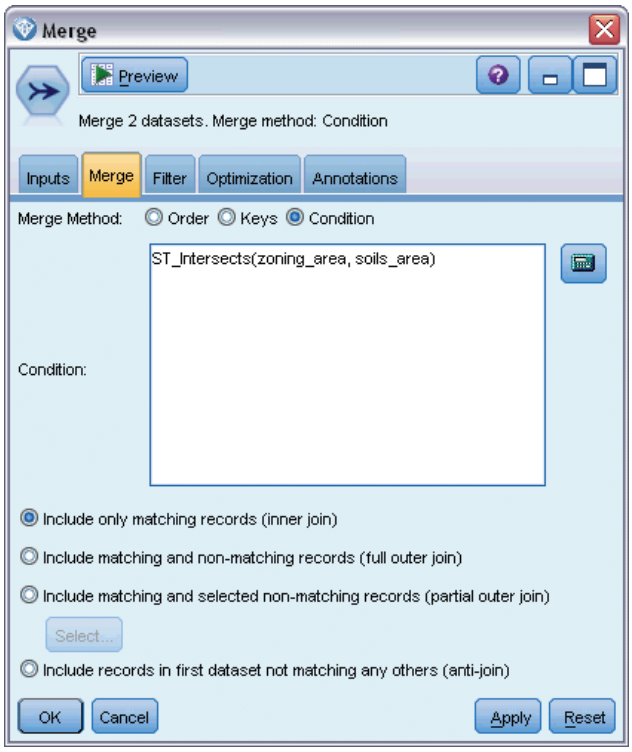


Colonne Jointure externe. Dans la colonne *Jointure externe*, sélectionnez les ensembles de données à inclure dans leur intégralité. Pour une jointure partielle, les enregistrements qui se chevauchent seront conservés, de même que les enregistrements incomplets pour les ensembles de données sélectionnés à ce niveau. Pour plus d'informations, reportez-vous à la section Types de jointure sur p. 91.

Spécification de conditions pour une fusion

En définissant la méthode de fusion sur Condition, vous pouvez spécifier une ou plusieurs conditions à satisfaire qui déterminent si la fusion a lieu ou pas.

Figure 3-18
Définition des conditions d'une fusion

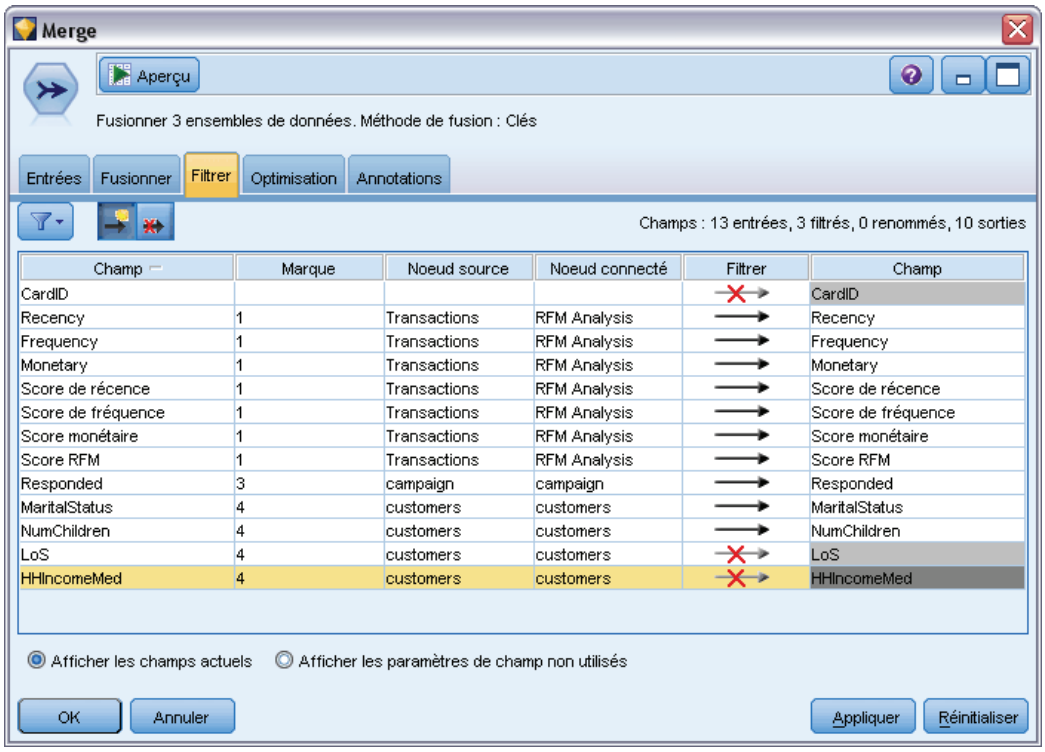


Vous pouvez entrer les conditions directement dans le champ Condition ou les créer à l'aide du Générateur de formules en cliquant sur le bouton en forme de calculatrice situé à droite du champ.

Filtrage des champs à partir du noeud Fusionner

Les noeuds Fusionner permettent de filtrer ou de renommer des champs apparaissant en double à la suite de la fusion de plusieurs sources de données. Cliquez sur l'onglet Filtrer de la boîte de dialogue pour sélectionner les options de filtrage.

Figure 3-19
Filtrage à partir du noeud Fusionner



Les options présentées sont quasi-identiques à celles du noeud Filtrer. Toutefois, des options supplémentaires, non présentées ici, sont disponibles dans le menu Filtrer. Pour plus d'informations, reportez-vous à la section Filtrage ou modification du nom des champs dans le chapitre 4 sur p. 155.

Champ. Affiche les champs d'entrée des sources de données actuellement connectées.

Marque. Affiche le nom (ou le numéro) de la marque associée au lien de la source de données. Cliquez sur l'onglet Entrées pour modifier les liens actifs vers ce noeud Fusionner.

Noeud source. Affiche le noeud source dont les données sont en cours de fusion.

Noeud connecté. Affiche le nom du noeud connecté au noeud Fusionner. Les travaux de Data mining complexes nécessitent souvent plusieurs opérations de fusion ou d'ajout qui peuvent inclure le même noeud source. Le nom du noeud connecté permet de les distinguer.

Filtrer. Affiche les connexions actuelles entre le champ d'entrée et le champ de sortie. Les connexions actives affichent une flèche continue. Les connexions affichant un X rouge indiquent des champs filtrés.

Champ. Affiche les champs de sortie après la fusion ou l'ajout. Les champs en double sont affichés en rouge. Cliquez dans le champ Filtrer pour désactiver les champs en double.

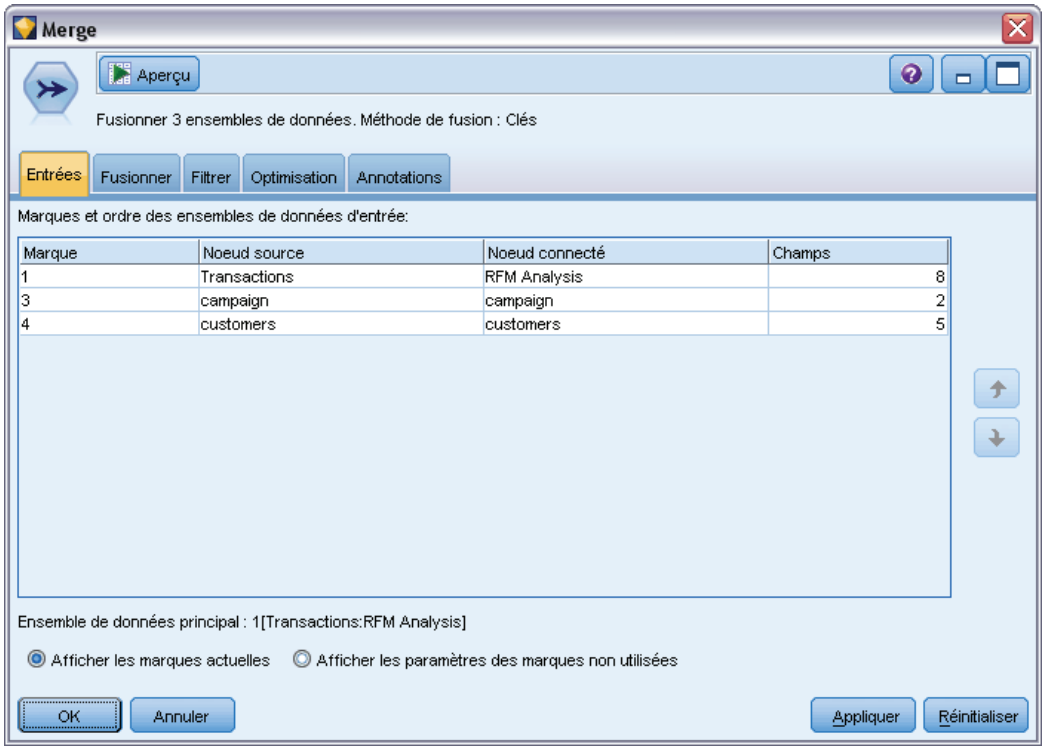
Afficher les champs actuels. Sélectionnez cette option pour afficher des informations sur les champs à utiliser en tant que champs-clés.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher des informations sur les champs actuellement non utilisés.

Définition de l'ordre d'entrée et du marquage

A l'aide de l'onglet Entrées des boîtes de dialogue des noeuds Fusionner et Ajouter, vous pouvez spécifier l'ordre des sources de données d'entrée et modifier le nom de la marque de chaque source.

Figure 3-20
Utilisation de l'onglet Entrées pour spécifier les marques et l'ordre d'entrée



Marques et ordre des ensembles de données d'entrée. Sélectionnez cette option pour fusionner ou ajouter uniquement les enregistrements complets.

- **Marque.** Affiche les noms de marque actuels pour chaque source de données d'entrée. Les noms de marque, également appelés **marques**, permettent d'identifier de façon unique les liens de données pour les opérations de fusion et d'ajout. Imaginons, par exemple, de l'eau provenant de divers conduits, qui débouche en un point, puis circule à travers un conduit unique. Les données dans IBM® SPSS® Modeler circulent de façon similaire et le point de fusion est souvent une interaction complexe entre les différentes sources de données. Les marques permettent de gérer les entrées (« conduits ») vers un noeud Fusionner ou Ajouter de sorte que si le noeud est enregistré ou déconnecté, les liens ne sont pas supprimés et sont facilement identifiables.

Lorsque vous connectez des sources de données supplémentaires à un noeud Fusionner ou Ajouter, des marques par défaut sont automatiquement créées, à l'aide de nombres, afin de représenter l'ordre de connexion des noeuds. Cet ordre n'est pas lié à l'ordre des champs dans

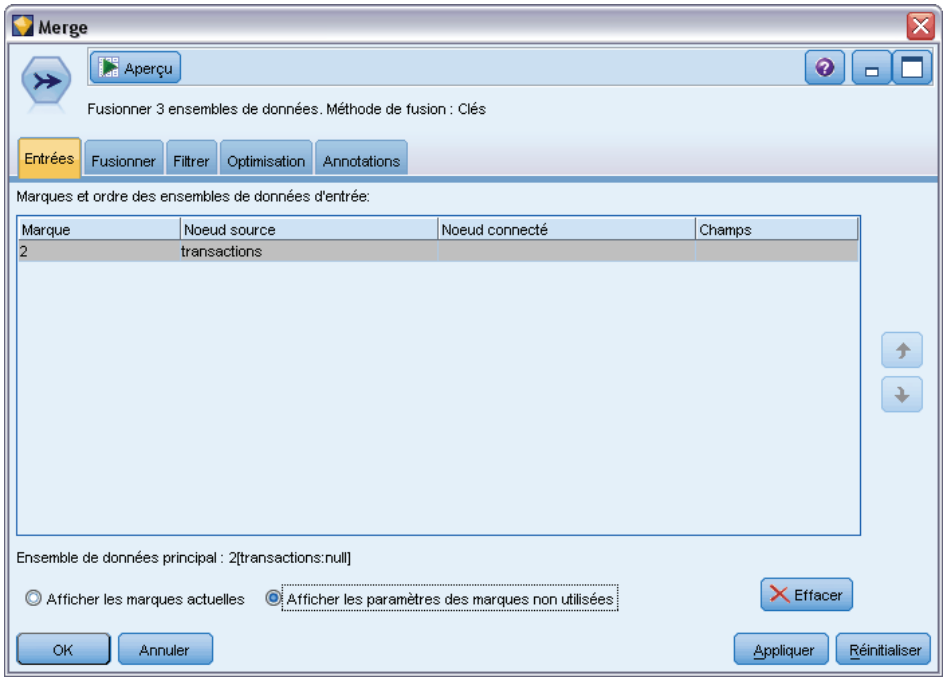
les ensembles de données d'entrée ou de sortie. Vous pouvez modifier la marque par défaut en entrant un nouveau nom dans la colonne *Marque*.

- **Noeud source.** Affiche le noeud source dont les données sont en cours de fusion.
- **Noeud connecté.** Affiche le nom du noeud connecté au noeud Fusionner ou Ajouter. Les travaux de Data mining complexes nécessitent souvent plusieurs opérations de fusion qui peuvent inclure le même noeud source. Le nom du noeud connecté permet de les distinguer.
- **Champ.** Répertorie le nombre de champs dans chaque source de données.

Afficher les marques actuelles. Sélectionnez cette option pour afficher les marques actuellement utilisées par le noeud Fusionner ou Ajouter. En d'autres termes, les marques actuelles identifient les liens vers le noeud à travers lequel circulent des données. En reprenant la métaphore des conduits, les marques actuelles s'apparentent aux conduits dans lesquels circule de l'eau.

Afficher les paramètres des marques non utilisées. Sélectionnez cette option pour afficher les marques, ou liens, précédemment utilisées pour la connexion au noeud Fusionner ou Ajouter, mais qui ne sont pas actuellement connectées à une source de données. Cette représentation s'apparente aux conduits vides au sein d'un réseau de plomberie. Vous pouvez choisir de connecter ces « conduits » à une nouvelle source ou de les supprimer. Pour supprimer du noeud les marques non utilisées, cliquez sur Effacer. Cette action efface en une fois toutes les marques non utilisées.

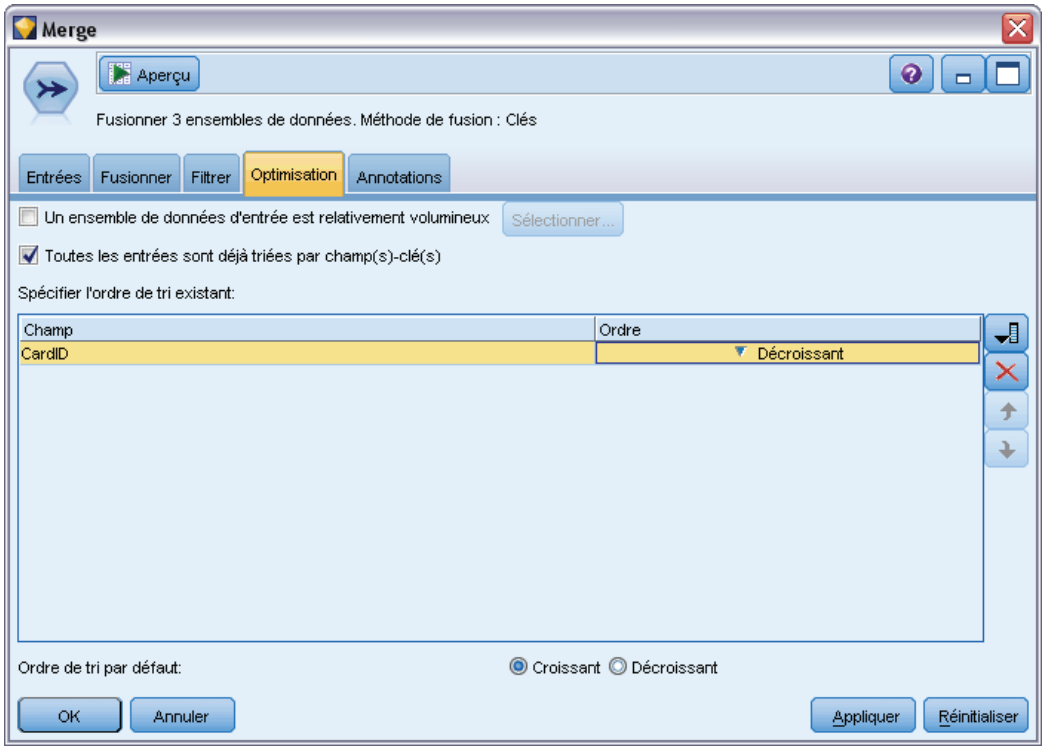
Figure 3-21
Suppression des marques non utilisées à partir du noeud Fusionner



Paramètres d'optimisation de la fusion

Le système comprend deux options qui permettent une fusion plus efficace des données dans certaines situations. Ainsi, grâce à ces options, vous pouvez optimiser l'opération de fusion lorsqu'un ensemble de données d'entrée est nettement plus volumineux que les autres, ou lorsque les données sont déjà triées en fonction de tout ou partie des champs-clés utilisés pour la fusion.

Figure 3-22
Paramètres d'optimisation



Un ensemble de données d'entrée est relativement volumineux. Sélectionnez cette option pour indiquer que l'un des ensembles de données d'entrée est bien plus volumineux que les autres. Le système met alors en mémoire cache les ensembles de données moins volumineux, puis traite, pour la fusion, l'ensemble de données le plus volumineux sans le trier ni le mettre en mémoire cache. Vous utilisez généralement ce type de jointure avec des données organisées selon un schéma en étoile (ou selon tout autre schéma semblable), en présence d'une table centrale volumineuse contenant des données partagées (des données transactionnelles, par exemple). Si vous sélectionnez cette option, cliquez sur Sélectionner pour indiquer l'ensemble de données volumineux. Vous ne pouvez sélectionner qu'*un* ensemble de données volumineux. Le tableau suivant reprend chaque type de jointure et indique pour chacun s'il peut être optimisé à l'aide de cette méthode.

Type de jointure	**Peut être optimisé pour un ensemble de données d'entrée volumineux ?**
Interne	Oui
Partiel	Oui, si l'ensemble de données volumineux ne contient pas d'enregistrements incomplets.

Type de jointure	Peut être optimisé pour un ensemble de données d'entrée volumineux ?
Complet	Non
Anti-jointure	Oui, si l'ensemble de données volumineux constitue la première entrée.

Toutes les entrées sont déjà triées par champ(s)-clé(s). Sélectionnez cette option pour indiquer que les données d'entrée sont déjà triées en fonction d'un ou de plusieurs des champs-clés utilisés pour la fusion. Assurez-vous que *tous* les ensembles de données d'entrée sont triés.

Spécifier l'ordre de tri existant. Indiquez les champs déjà triés. Dans la boîte de dialogue Sélectionner les champs, ajoutez des champs à la liste. Vous ne pouvez sélectionner que les champs-clés utilisés pour la fusion (tels que définis dans l'onglet Fusionner). Dans la colonne *Ordre*, précisez si chaque champ est trié dans l'ordre croissant ou décroissant. Si vous entrez ici plusieurs champs, veillez à les répertorier dans le bon ordre. Cliquez sur les flèches situées à droite de la liste pour trier les champs dans l'ordre souhaité. Si vous définissez un ordre de tri existant incorrect, un message d'erreur apparaît lors de l'exécution du flux, indiquant le numéro d'enregistrement au niveau duquel le tri n'est pas conforme à ce que vous avez indiqué.

En fonction de la sensibilité à la casse de la méthode de collationnement utilisée par la base de donnée, il se peut que l'optimisation ne fonctionne pas correctement lorsqu'une ou plusieurs entrées sont triées par la base de données. Par exemple, si sur deux entrées, l'une est sensible à la casse et l'autre non, les résultats du tri peuvent être différents. L'optimisation de la fusion entraîne le traitement des enregistrements selon leur ordre de tri. En conséquence, si les entrées sont triées à l'aide de méthodes de collationnement différentes, le noeud Fusion fait état d'une erreur et affiche le numéro de l'enregistrement dont le tri a été incohérent. Lorsque toutes les entrées proviennent d'une source unique, ou sont triées à l'aide de collationnements mutuellement inclusifs, les enregistrements peuvent être fusionnés avec succès.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de fusion.

Noeud Ajouter

Les noeuds Ajouter permettent de concaténer des ensembles d'enregistrements. Contrairement au noeud Fusionner, qui joint des enregistrements provenant de sources différentes, le noeud Ajouter lit tous les enregistrements d'une source jusqu'au dernier et les inclut dans le flux en aval. Les enregistrements de la source suivante sont ensuite lus, avec la même structure de données (nombre d'enregistrements, nombre de champs, etc.) que la première source, ou source principale. Si la source principale comporte plus de champs qu'une autre source d'entrée, la chaîne manquante par défaut (`$null$`) est utilisée pour les valeurs incomplètes.

Utilisez un noeud Ajouter pour combiner des ensembles de données dont les structures sont similaires, mais les données différentes. Par exemple, vous pouvez avoir stocké les données relatives à vos transactions dans des fichiers différents selon la période à laquelle elles se rapportent (un fichier pour mars et un autre pour avril, par exemple). Supposons que ces fichiers sont structurés de la même manière (les mêmes champs dans le même ordre), le noeud Ajouter réunira les données dans un même fichier que vous pourrez alors analyser.

Remarque : pour que les fichiers puissent être ajoutés, les niveaux de mesure de champ doivent être identiques. Par exemple, un champ *Nominal* ne peut être ajouté à un champ dont le niveau de mesure est *Continu*.

Figure 3-23
Boîte de dialogue du noeud Ajouter présentant une correspondance de champ par nom

Définition des options du noeud Ajouter

Apparier les champs par. Sélectionnez la méthode de correspondance des champs à ajouter.

- **Position :** Permet d'ajouter des ensembles de données, sur la base de la position des champs dans la source de données principale. Lorsque vous utilisez cette méthode, pensez à trier vos données afin de garantir un ajout adéquat.
- **Nom.** Permet d'ajouter des ensembles de données, sur la base du nom des champs dans les ensembles de données d'entrée. Sélectionnez également Respecter la casse pour activer la distinction des majuscules/minuscules lors de la mise en correspondance des noms de champ.

Champ de sortie. Affiche les noeuds source connectés au noeud Ajouter. Le premier noeud de la liste est la source d'entrée principale. Vous pouvez trier les champs en cliquant sur l'en-tête de la colonne. Ce tri n'a pas pour effet de réorganiser les champs dans l'ensemble de données.

Inclure les champs de. Sélectionnez Premier ensemble de données uniquement pour générer des champs de sortie sur la base des champs de l'ensemble de données principal. L'ensemble de données principal est la première entrée, spécifiée dans l'onglet Entrées. Sélectionnez Tous les ensembles de données pour générer des champs de sortie pour tous les champs dans tous les ensembles de données, qu'un champ correspondant soit présent dans tous les ensembles de données d'entrée ou non.

Marquer les enregistrements en incluant l'ensemble de données dans le champ. Sélectionnez cette option pour ajouter un champ supplémentaire au fichier de sortie, dont les valeurs indiquent l'ensemble de données source pour chaque enregistrement. Indiquez un nom dans le champ de texte. Le nom par défaut du champ est *Entrée*.

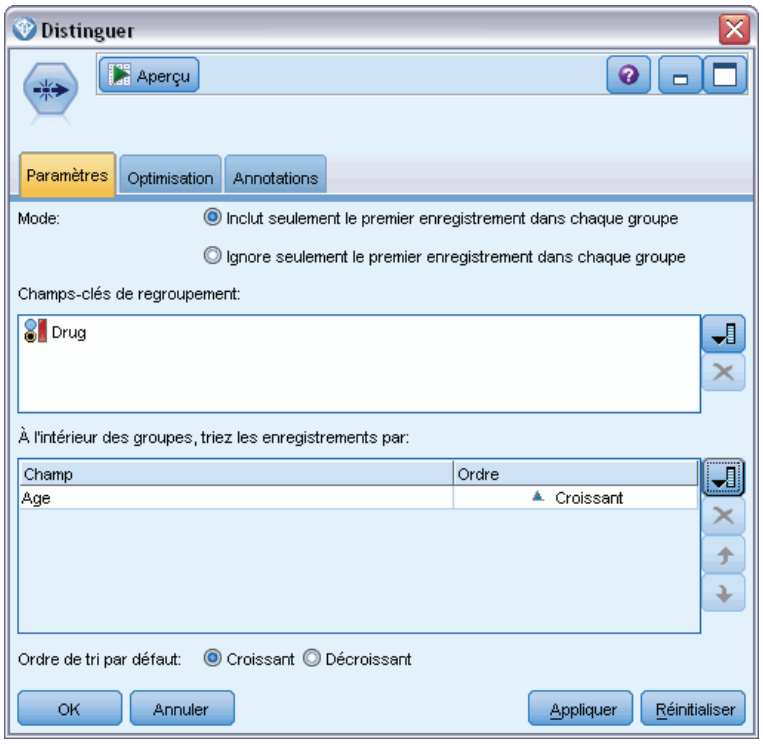
Noeud Distinguer

Les enregistrements en double d'un ensemble de données doivent être supprimés avant le début du Data mining. Par exemple, dans une base de données marketing, certaines personnes peuvent apparaître plusieurs fois avec des adresses différentes ou des informations de contact différentes. Afin de rechercher ou de supprimer les enregistrements en double de votre ensemble de données, utilisez le noeud Distinguer.

A l'aide des noeuds Distinguer, vous pouvez supprimer les enregistrements en double en incluant le premier enregistrement distinct dans le flux de données ou vous pouvez rechercher des enregistrements en double en supprimant le premier enregistrement distinct et en incluant ses doublons dans le flux de données.

En outre, vous pouvez définir un ordre de tri pour chaque valeur de clé distincte pour les résultats retournés. Si vous souhaitez qu'une ligne spécifique soit retournée pour chaque clé distincte, vous devez trier les enregistrements dans le noeud Distinguer plutôt qu'utiliser un noeud Trier en amont (Reportez-vous à "Tri des enregistrements dans le noeud Distinguer", ci-dessous).

Figure 3-24
Boîte de dialogue du noeud Distinguer



Mode. Indiquez si vous souhaitez enlever ou isoler (retirer) le premier enregistrement.

- **Inclure uniquement le premier enregistrement dans chaque groupe.** Permet d'inclure le premier enregistrement distinct dans le flux de données et de supprimer tous ses doublons.
- **Ignorer uniquement le premier enregistrement dans chaque groupe.** Permet d'ignorer le premier enregistrement distinct trouvé et d'inclure ses doublons dans le flux de données. Cette option permet de *détecter* les doublons présents dans les données, afin qu'ils puissent être examinés ultérieurement dans le flux.

Champs-clés pour le regroupement. Répertorie le ou les champs utilisés pour détecter les enregistrements identiques. Vous pouvez :

- Ajouter des champs à cette liste en utilisant le bouton de sélection des champs situé à droite.
- Supprimer des champs de la liste à l'aide du bouton de suppression rouge en forme de X.

Au sein des groupes, trier les enregistrements par. Répertorie les champs utilisés pour déterminer le mode de tri des enregistrements pour chaque valeur de clé distincte et déterminer s'ils sont triés dans l'ordre croissant ou décroissant. Vous pouvez :

- Ajouter des champs à cette liste en utilisant le bouton de sélection des champs situé à droite.
- Supprimer des champs de la liste à l'aide du bouton de suppression rouge en forme de X.
- Déplacer les champs à l'aide des boutons Haut ou Bas, si vous triez en fonction de plusieurs champs.

Ordre de tri par défaut. Spécifiez si par défaut, les enregistrements sont triés dans l'ordre croissant ou décroissant.

Tri des enregistrements dans le noeud Distinguer

Vous pouvez retourner une ligne spécifique pour chaque clé distincte en utilisant l'option Au sein des groupes, trier les enregistrements par dans le noeud Distinguer, de sorte qu'il n'est pas nécessaire d'utiliser un noeud Trier précédent. Par exemple, supposons que nous possédons les données suivantes sur les âges des utilisateurs de médicaments prescrits.

Age	Médicament
50	Médicament A
71	Médicament B
44	Médicament A
65	Médicament X
39	Médicament A
75	Médicament C
72	Médicament Y
57	Médicament X

Age	Médicament
79	Médicament Y
69	Médicament C
74	Médicament B
85	Médicament Y
69	Médicament X

Pour trouver le consommateur le plus âgé de chaque médicament, nous définissons le mode de manière à inclure uniquement le premier enregistrement dans chaque groupe, utilisons le médicament comme le champ-clé, et l'âge comme champ de tri dans l'ordre décroissant. L'ordre d'entrée n'affecte pas les résultats car la sélection de tri spécifie quelle ligne d'un médicament donné sera retournée, et le résultat final doit être le suivant.

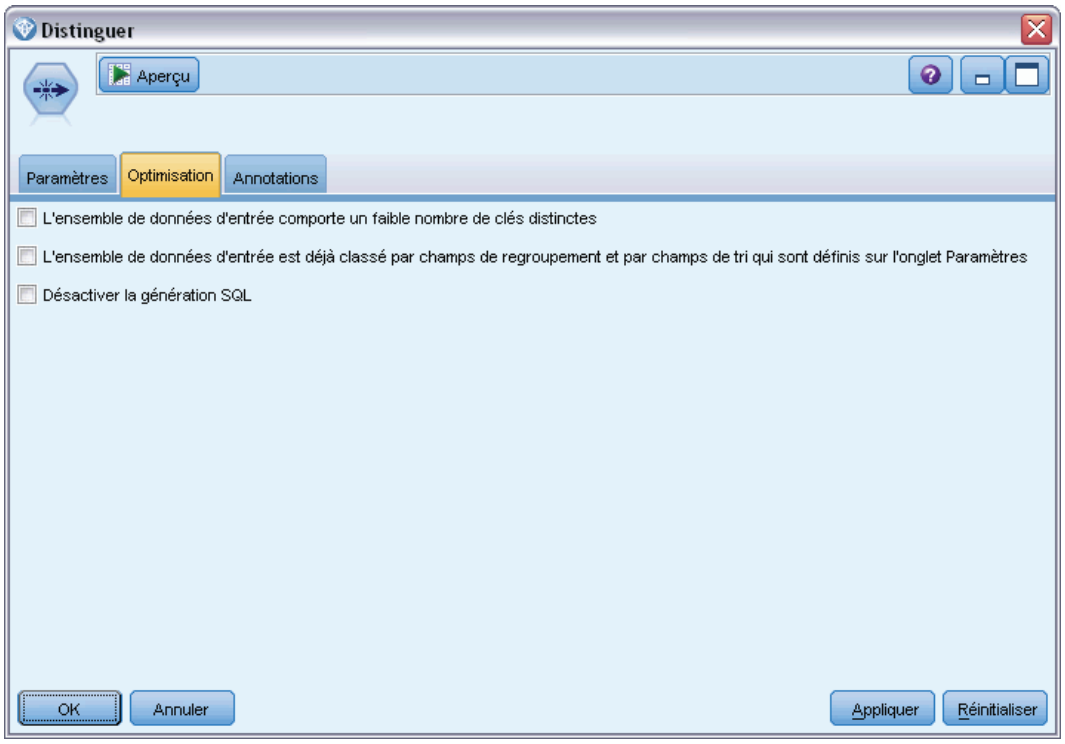
Age	Médicament
50	Médicament A
74	Médicament B
75	Médicament C
69	Médicament X
85	Médicament Y

Paramètres d'optimisation distincts

Si les données sur lesquelles vous travaillez ne contiennent qu'un petit nombre d'enregistrements ou ont déjà été triées, vous pouvez optimiser la manière dont elles sont traitées pour permettre à IBM® SPSS® Modeler de traiter les données de manière plus efficace.

Remarque : Que vous sélectionniez l'option L'ensemble de données d'entrée comporte un petit nombre de clés distinctes. ou utilisiez la génération SQL pour le noeud, toute ligne dans la valeur de clé distincte peut être retournée. Afin de contrôler quelle ligne est retournée pour une clé distincte, vous devez spécifier l'ordre de tri en utilisant les champs Au sein des groupes, trier les enregistrements par sur l'onglet Paramètres. Les options d'optimisation n'affectent pas les résultats retournés par le noeud Distinguer tant qu'un ordre de tri est spécifié sur l'onglet Paramètres.

Figure 3-25
Paramètres d'optimisation



L'ensemble de données d'entrée comporte un petit nombre de clés distinctes. Sélectionnez cette option si vous avez un petit nombre d'enregistrements et/ou un petit nombre de valeurs uniques des champs-clés. Ainsi, vous améliorez les performances.

L'ensemble de données d'entrée est déjà ordonnée en champs de regroupement et en champs de tri sur l'onglet Paramètres. Sélectionnez cette option uniquement si vos données sont déjà triées en fonction de tous les champs Au sein des groupes, trier les enregistrements par répertoriés sur l'onglet Paramètres, et si l'ordre de tri (croissant ou décroissant) est identique pour toutes les données. Ainsi, vous améliorez les performances.

Désactiver la génération SQL. Sélectionnez cette option pour désactiver la génération SQL pour le noeud.

Chapitre
4

Noeuds d'opérations sur les champs

Présentation des opérations sur les champs

Après une première exploration des données, vous devrez peut-être sélectionner, nettoyer ou élaborer des données en vue d'une préparation à l'analyse. La palette Opérations sur les champs contient de nombreux noeuds utiles aux opérations de transformation et de préparation.

Par exemple, à l'aide d'un noeud Calculer, vous pouvez créer un attribut qui n'est pas actuellement représenté dans les données. Vous pouvez également utiliser un noeud Discrétiser pour recoder automatiquement les valeurs de champ de l'analyse cible. Vous aurez certainement recours fréquemment au noeud Typer ; en effet, il permet d'attribuer un niveau de mesure, des valeurs et un rôle de modélisation à chaque champ de l'ensemble de données. Ces opérations sont utiles pour la gestion des valeurs manquantes et la modélisation en aval.

La palette Opérations sur les champs contient les noeuds suivants :

Le noeud de préparation automatisée de données (ADP) peut analyser vos données, identifier des corrections et filtrer des champs qui sont problématiques ou qui sont peu susceptibles d'être utiles. Il peut aussi créer de nouveaux attributs le cas échéant et améliorer la performance au moyen de techniques de filtrage et d'échantillonnage intelligentes. Vous pouvez utiliser le noeud de manière totalement automatisée, en laissant le noeud choisir et appliquer les corrections, ou vous pouvez prévisualiser les modifications avant qu'elles ne soient mises en place et les accepter, les rejeter ou les modifier selon les besoins. Pour plus d'informations, reportez-vous à la section Préparation automatique des données sur p. 109.

Le noeud Typer définit les propriétés et métadonnées de champ. Par exemple, vous pouvez indiquer un niveau de mesure (continu, nominal, ordinal ou booléen) pour chaque champ, définir des options pour la gestion des valeurs manquantes et des valeurs système nulles, spécifier le rôle d'un champ en vue de la modélisation, définir des étiquettes de champ et de valeur, et indiquer les valeurs d'un champ. Pour plus d'informations, reportez-vous à la section Noeud Typer sur p. 136.

Le noeud Filtrer filtre (supprime) les champs, les renomme et les mappe entre un noeud source et un autre. Pour plus d'informations, reportez-vous à la section Filtrage ou modification du nom des champs sur p. 155.

Le noeud Calculer modifie les valeurs de données ou crée des nouveaux champs à partir d'un ou de plusieurs champs existants. Il crée des champs de type formule, booléen, ensemble, nominal, statistiques, comptage et conditionnel. Pour plus d'informations, reportez-vous à la section Noeud Calculer sur p. 166.

Le noeud Ensemble combine deux ou plusieurs nuggets de modèles pour obtenir des prévisions plus précises que celles acquises à partir d'un modèle quelconque. Pour plus d'informations, reportez-vous à la section Noeud Ensemble sur p. 162.

© Copyright IBM Corporation 1994, 2012.

Le noeud Remplacer permet de remplacer les valeurs de champ et de modifier le type de stockage. Vous pouvez décider de remplacer les valeurs reposant sur une condition CLEM, telle que @BLANK(@FIELD). Vous pouvez également choisir de remplacer tous les blancs ou toutes les valeurs nulles par une valeur précise. Un noeud Remplacer est souvent associé à un noeud Typer pour remplacer les valeurs manquantes. Pour plus d'informations, reportez-vous à la section Noeud Remplacer sur p. 178.

Le noeud Anonymiser transforme la façon dont les noms et les valeurs des champs sont représentés en aval, masquant ainsi les données d'origine. Cela peut s'avérer utile si vous souhaitez permettre à d'autres utilisateurs de générer des modèles utilisant des données confidentielles, par exemple des noms de clients ou autre. Pour plus d'informations, reportez-vous à la section Noeud Anonymiser sur p. 182.

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser. Pour plus d'informations, reportez-vous à la section Noeud Recoder sur p. 186.

Le noeud Discrétiser crée automatiquement des champs nominaux (ensemble) sur la base des valeurs d'un ou de plusieurs champs continus (intervalle numérique) existants. Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus comme écarts par rapport à la moyenne. Une fois les intervalles du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure. Pour plus d'informations, reportez-vous à la section Noeud Discrétiser sur p. 191.

Le noeud Analyse RFM (Récence, Effectif, Monétaire) permet de déterminer de façon quantitative les clients susceptibles d'être les meilleurs par l'étude de leur dernier achat (récence), l'effectif de leurs achats (effectif), et la somme dépensée lors de toutes les transactions (monétaire). Pour plus d'informations, reportez-vous à la section Noeud Analyse RFM sur p. 203.

Le noeud Partitionner génère un champ de partition qui répartit les données dans des sous-ensembles distincts pour les étapes d'apprentissage, de test et de validation de la création d'un modèle. Pour plus d'informations, reportez-vous à la section Noeud Partitionner sur p. 207.

Le noeud Binariser calcule plusieurs champs booléens en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux. Pour plus d'informations, reportez-vous à la section Noeud Binariser sur p. 210.

Le noeud Restructurer convertit un champ nominal ou un champ booléen en un groupe de champs renseignés à partir des valeurs d'un autre champ. Par exemple, si l'on considère un champ nommé *type de paiement*, qui comporte les valeurs *crédit*, *liquide* et *débit*, trois champs sont alors créés (*crédit*, *liquide*, *débit*), chacun contenant la valeur du paiement réel effectué. Pour plus d'informations, reportez-vous à la section Noeud Restructurer sur p. 211.

Le noeud Transposer fait passer les données des lignes vers les colonnes (et réciproquement) de sorte que les enregistrements deviennent des champs et les champs des enregistrements. Pour plus d'informations, reportez-vous à la section Noeud Transposer sur p. 214.

Le noeud Intervalle de temps définit des intervalles et crée, si nécessaire, des étiquettes pour la modélisation des séries temporelles. Si les valeurs ne sont pas espacées de manière égale, ce noeud peut les étoffer ou les agréger, selon les besoins, pour générer un intervalle uniforme entre les enregistrements. Pour plus d'informations, reportez-vous à la section Noeud Intervalles de temps sur p. 219.

Le noeud Historiser crée des champs contenant des données provenant de champs d'enregistrements antérieurs. Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Avant d'utiliser un noeud Historiser, vous pouvez trier les données à l'aide d'un noeud Trier. Pour plus d'informations, reportez-vous à la section Noeud Historiser sur p. 240.

Le noeud Re-trier définit l'ordre naturel utilisé pour afficher les champs situés en aval. Cet ordre a une incidence sur l'affichage des champs en différents endroits : tableaux, listes et sélecteur de champs. Cette opération est utile lorsque vous utilisez des ensembles de données volumineux pour rendre plus visibles les champs intéressants. Pour plus d'informations, reportez-vous à la section Noeud Re-trier sur p. 242.

Certains de ces noeuds peuvent être générés directement à partir du rapport d'audit créé par un noeud Audit données. Pour plus d'informations, reportez-vous à la section Génération d'autres noeuds en vue d'une préparation de données dans le chapitre 6 sur p. 435.

Préparation automatique des données

La préparation des données pour l'analyse est une des étapes les plus importantes des projets—et généralement, l'une de celles qui prend le plus de temps. La préparation automatique des données (ADP) s'occupe de cette tâche à votre place, analyse vos données, identifie les corrections, supprime les champs problématiques ou inutiles, dérive de nouveaux attributs si nécessaire et améliore les performances grâce à des techniques d'analyse intelligentes. Vous pouvez utiliser l'algorithme en mode complètement **automatique**, le laissant choisir et appliquer les corrections ou vous pouvez utiliser son mode **interactif** qui prévoit les modifications avant qu'elles ne soient effectuées vous laissant libre de les accepter ou de les refuser.

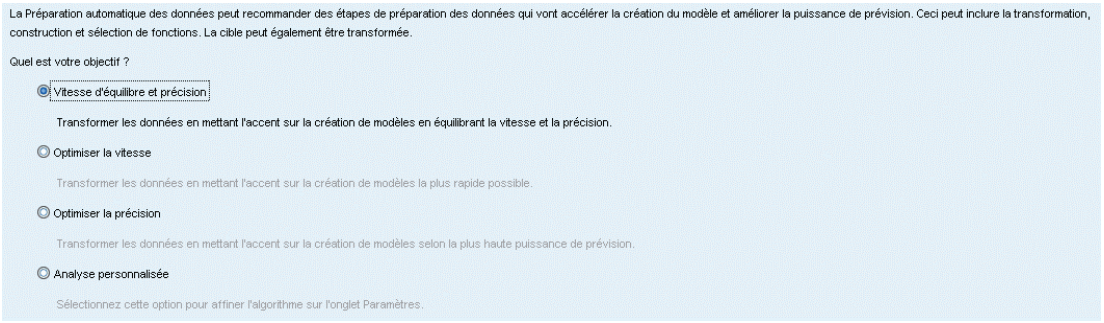
L'utilisation de l'ADP vous permet de préparer facilement et rapidement vos données pour la création de modèle, sans qu'il soit nécessaire de maîtriser les concepts de statistiques utilisés. Les modèles seront alors créés et les scores déterminés plus rapidement ; de plus, l'utilisation de l'ADP améliore la robustesse des processus de modélisation automatique, tels que les actualisations de modèles et les champion / challenger.

Remarque : lorsque la préparation automatique des données prépare un champ pour l'analyse, elle crée un nouveau champ contenant les ajustements ou les transformations, au lieu de remplacer les valeurs et les propriétés existantes de l'ancien champ. L'ancien champ n'est pas utilisé pour l'analyse, son rôle est défini sur Aucun.

Exemple : Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour signaler des réclamations suspectes et potentiellement frauduleuses. Avant de construire le modèle, il est nécessaire de préparer les données à l'aide de la préparation automatique des données. La compagnie souhaitant être capable de consulter et modifier les transformations avant de les appliquer, elle utilise la préparation automatique des données de manière interactive.

Un groupe automobile suit les ventes de véhicules automobiles personnels divers. Afin d'être en mesure d'identifier les modèles dont les ventes sont très satisfaisantes et ceux pour lesquels elles le sont moins, des responsables du groupe souhaitent établir une relation entre les ventes de véhicules et les caractéristiques des véhicules. Ils utilisent la préparation automatique des données pour cette analyse afin de construire des modèles à l'aide des données « avant » et « après » la préparation et de pouvoir en comparer les résultats.

Figure 4-1
Onglet Objectif de la préparation automatique des données



Quel est votre objectif ? La préparation automatique des données recommande des étapes de préparation de données qui amélioreront la vitesse de création de modèles par les autres algorithmes et la puissance de prédiction de ces modèles. Cela peut comprendre la transformation, la construction et la sélection de fonctionnalités. La cible peut également être transformée. Vous pouvez spécifier les priorités de création de modèle sur lesquelles le processus de préparation des données doit se concentrer.

- **Équilibrer la vitesse et la précision.** Cette option prépare les données à accorder la même importance à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle et à la précision des prévisions.
- **Optimiser la vitesse.** Cette option prépare les données à accorder la priorité à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle. Lorsque vous travaillez avec de très grands ensembles de données ou que vous recherchez une réponse rapide, sélectionnez cette option.
- **Optimiser la précision.** Cette option prépare les données à accorder la priorité à la précision des prédictions produites par les algorithmes de création de modèle.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement l'algorithme dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'un des autres objectifs.

Formation du noeud

Le noeud ADP est mis en œuvre en tant que noeud de processus et fonctionne de la même façon qu'un noeud Type ; la **formation** du noeud ADP correspond à l'instanciation du noeud Type. Lorsque l'analyse est terminée, les transformations spécifiées sont appliquées aux données sans analyse supplémentaire tant que le modèle des données en amont ne change pas. Tout comme les noeuds Type et Filtre, si le noeud ADP est déconnecté, il se souvient du modèle de données et des transformations et n'a pas besoin d'être de nouveau formé lorsqu'il est reconnecté. Cela vous

permet de le former sur un sous-ensemble de données standard puis de le déployer ou de le copier pour l'utiliser avec des données en direct aussi souvent que possible.

Utilisation de la barre d'outils

La barre d'outils permet d'exécuter et de mettre à jour l'affichage de l'analyse des données et de générer des noeuds pouvant être utilisés en conjonction avec les données d'origine.

Figure 4-2
Préparation automatique des données : barre d'outils

- **Générer** Depuis ce menu, vous pouvez générer un noeud Filtre ou un noeud Calculer. Veuillez noter que ce menu est uniquement disponible lorsqu'une analyse apparaît dans l'onglet Analyse.

  Le noeud Filtre supprime les champs d'entrée transformés. Si vous configurez le noeud ADP pour qu'il laisse les champs d'entrée d'origine dans l'ensemble de donnés, cela restaure l'ensemble d'entrées d'origine et vous permet d'interpréter le champ des scores en terme d'entrées. Par exemple, cela peut être utile si vous souhaitez produire un diagramme du champ de scores par rapport aux différentes entrées.

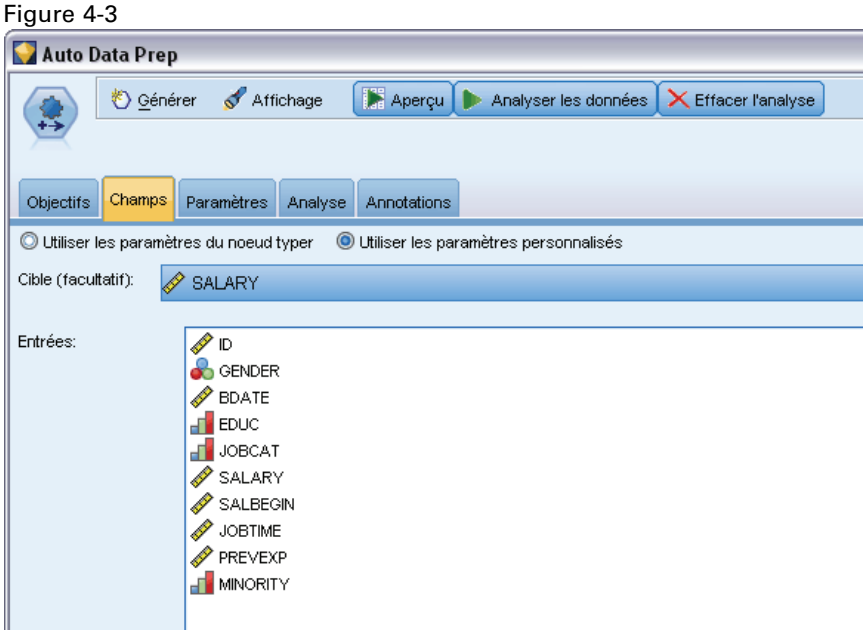
  Le noeud Calculer peut restaurer l'ensemble de données d'origine et les unités cibles. Vous ne pouvez générer un noeud Calculer que lorsque le noeud ADP contient une analyse qui rééchelonne une cible plage (c'est-à-dire que le rééchelonnement de Box-Cox est sélectionné dans le panneau Préparer les entrées & la cible). Vous ne pouvez pas générer de noeud Calculer si la cible n'est pas une plage, ou si le rééchelonnement de Box-Cox n'est pas sélectionné. Pour plus d'informations, reportez-vous à la section Génération d'un noeud Calculer sur p. 134.
- **Afficher** Contient des options qui contrôlent ce qui apparaît dans l'onglet Analyse. Cela comprend les commandes de modification des diagrammes et les sélections d'affichage à la fois pour le panneau principal et les vues liées.
- **Aperçu** Affiche un échantillon des transformations qui seront appliquées aux données d'entrée.
- **Analyser les données** Démarre une analyse avec les paramètres actuels et affiche les scores dans l'onglet Analyse.
- **Effacer l'analyse** Supprime l'analyse existante (disponible uniquement lorsqu'une analyse en cours existe).

Statut du noeud

Le statut du noeud ADP sur le canevas IBM® SPSS® Modeler est indiqué par une flèche ou par une graduation sur l'icône qui indique si l'analyse a eu lieu ou pas.

Onglet Champs

Figure 4-3



Avant de construire un modèle, vous devez indiquer les champs à utiliser en tant que cibles et en tant qu'entrées. A quelques exceptions près, tous les noeuds de modélisation utilisent les informations de champ d'un nœud Typer en amont. Si vous utilisez un nœud Typer pour sélectionner les champs d'entrée et les champs cible, vous n'avez aucune modification à apporter dans cet onglet.

Utiliser les paramètres du noeud Typer. Cette option indique au nœud d'utiliser les informations du champ à partir d'un nœud Typer en amont. Il s'agit de la valeur par défaut.

Utiliser les paramètres personnalisés. Cette option indique au nœud d'utiliser les informations du champ spécifiées ici au lieu des informations données dans un nœud Typer en amont. Une fois cette option sélectionnée, renseignez les champs ci-dessous.

Cible. Pour les modèles qui nécessitent un ou plusieurs champs cible, sélectionnez le ou les champs cible. Cela revient à définir le rôle du champ sur la valeur *Cible* dans un nœud Typer.

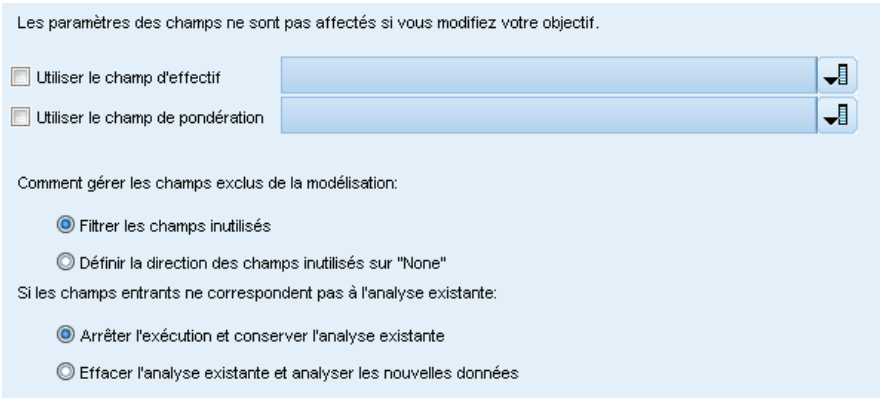
Entrées. Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un nœud Typer.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec les autres objectifs, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option Personnaliser l'analyse.

Paramètres des champs

Figure 4-4
Préparation automatique des données - Paramètres des champs



Utiliser un champ d'effectif. Cette option permet de sélectionner un champ en tant que pondération d'effectif. Utilisez cette option si les enregistrements de vos données d'apprentissage représentent chacun plus d'une seule unité, par exemple si vous utilisez des données agrégées. Les valeurs des champs doivent être égales au nombre d'unités représentées par chaque enregistrement.

Utiliser un champ de pondération. Cette option permet de sélectionner un champ en tant que pondération d'observation. Les pondérations d'observation sont utilisées pour représenter les différences de variance dans les niveaux du champ de résultat.

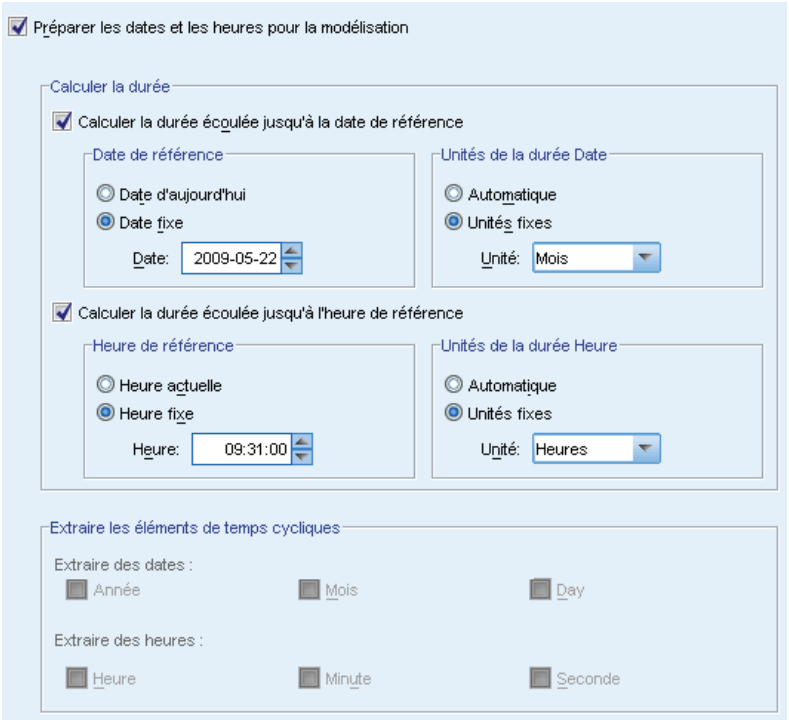
Comment traiter les champs exclus de la modélisation. Spécifiez la manière de traiter les champs exclus. Vous pouvez choisir de les filtrer des données ou de simplement définir leur *Rôle* sur Aucun.

Si les champs d'entrée ne correspondent pas à l'analyse existante. Spécifiez ce qui se passe si un ou plusieurs champs d'entrée requis sont manquants de l'ensemble de données entrant, lors de l'exécution d'un noeud ADP d'apprentissage.

- **Arrêter l'exécution et conserver l'analyse existante.** Cette option interrompt l'exécution, conserve les informations de l'analyse en cours et affiche une erreur.
- **Effacer l'analyse existante et analyser les nouvelles données.** Cette option efface l'analyse existante, analyse les données d'entrée et applique les transformations recommandées aux données.

Préparer les dates & les heures

Figure 4-5
Paramètres Dates & Heures de la préparation automatique des données



De nombreux algorithmes de modélisation ne peuvent pas traiter directement les informations sur la date et l'heure. Ces paramètres vous permettent de calculer de nouvelles données de durée qui peuvent être utilisées comme entrées de modèle à partir des dates et des heures de vos données existantes. Les champs contenant les dates et les heures doivent être prédéfinis à l'aide des types de stockage de dates et d'heures. Il n'est pas recommandé de définir les champs de date et d'heure d'origine comme entrées de modèle après la préparation automatique des données.

Préparer les dates et les heures pour la modélisation. En désélectionnant cette option, vous désactivez tous les autres contrôles Préparer les dates et les heures, tout en conservant les sélections.

Calculer la durée écoulée jusqu'à la date de référence. Cette option génère le nombre d'années/mois/jours depuis une date de référence pour chaque variable qui contient des dates.

- **Date de référence.** Spécifier la date à partir de laquelle la durée sera calculée en fonction des informations sur la date dans les données d'entrée. Sélectionner Date d'aujourd'hui signifie que la date du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une date spécifique, sélectionnez Date fixe et saisissez la date désirée. La date actuelle est automatiquement saisie dans le champ Date fixe lorsque le noeud est créé.
- **Unités de la durée Date.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Date ou choisir dans les unités fixes des Années, Mois ou Jours.

Calculer la durée écoulée jusqu'à l'heure de référence. Cette option génère le nombre d'heures/minutes/secondes depuis une heure de référence pour chaque variable qui contient des heures.

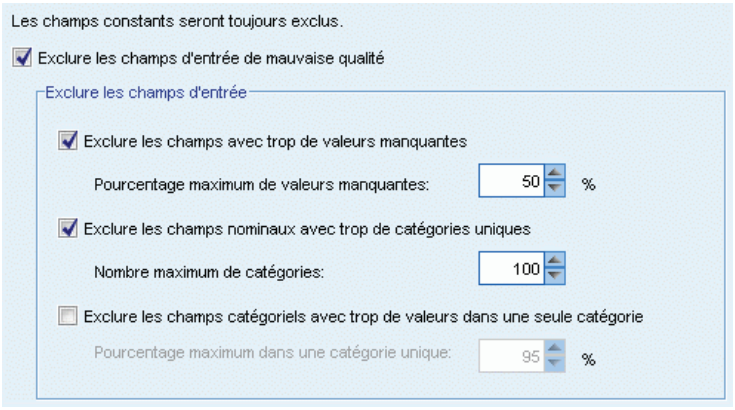
- **Heure de référence.** Spécifier l'heure à partir de laquelle la durée sera calculée en fonction des informations sur l'heure dans les données d'entrée. Sélectionner Heure actuelle signifie que l'heure du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une heure spécifique, sélectionnez Heure fixe et saisissez l'heure désirée. L'heure actuelle est automatiquement saisie dans le champ Heure fixe lorsque le noeud est créé.
- **Unités de la durée Heure.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Heure ou choisir dans les unités fixes des Heures, Minutes ou Secondes.

Extraire les éléments de temps cycliques. Utilisez ces paramètres pour scinder un champ de date ou d'heure en un ou plusieurs autres champs. Par exemple, si vous sélectionnez les trois cases de date, le champ de date d'entrée "1954-05-23" est divisé en trois champs : 1954, 5 et 23, chacun utilisant le suffixe défini dans le panneau Noms des champ et le champ de date d'origine est ignoré.

- **Extraire des dates.** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des années, des mois, des jours ou une des combinaisons possibles.
- **Extraire des heures.** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des heures, des minutes ou des secondes ou une des combinaisons possibles.

Exclure les champs

Figure 4-6
Paramètres Exclure les champs de la préparation automatique des données



Les données de mauvaise qualité peuvent affecter la précision de vos prédictions. Par conséquent, vous pouvez spécifier le niveau de qualité acceptable des caractéristiques d'entrée. Tous les champs constants ou avec 100% de valeurs manquantes sont automatiquement exclus.

Exclure les champs d'entrée de mauvaise qualité. En désélectionnant cette option, vous désactivez tous les autres contrôles Exclure les champs, tout en conservant les sélections.

Exclure les champs avec trop de valeurs manquantes. Les champs ayant plus que le pourcentage spécifié de valeurs manquantes sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à désélectionner cette option, et inférieure ou égale à 100, puisque

les champs qui ne contiennent que des valeurs manquantes sont exclus automatiquement. La valeur par défaut est 50.

Exclure les champs nominaux avec trop de modalités uniques. Les champs nominaux ayant plus que le nombre spécifié de modalités sont supprimés de l'analyse. Spécifiez un nombre entier positif. La valeur par défaut est 100. Cette option est utile pour supprimer automatiquement de la modélisation les champs contenant des informations d'enregistrement unique, tels que l'ID, l'adresse ou le nom.

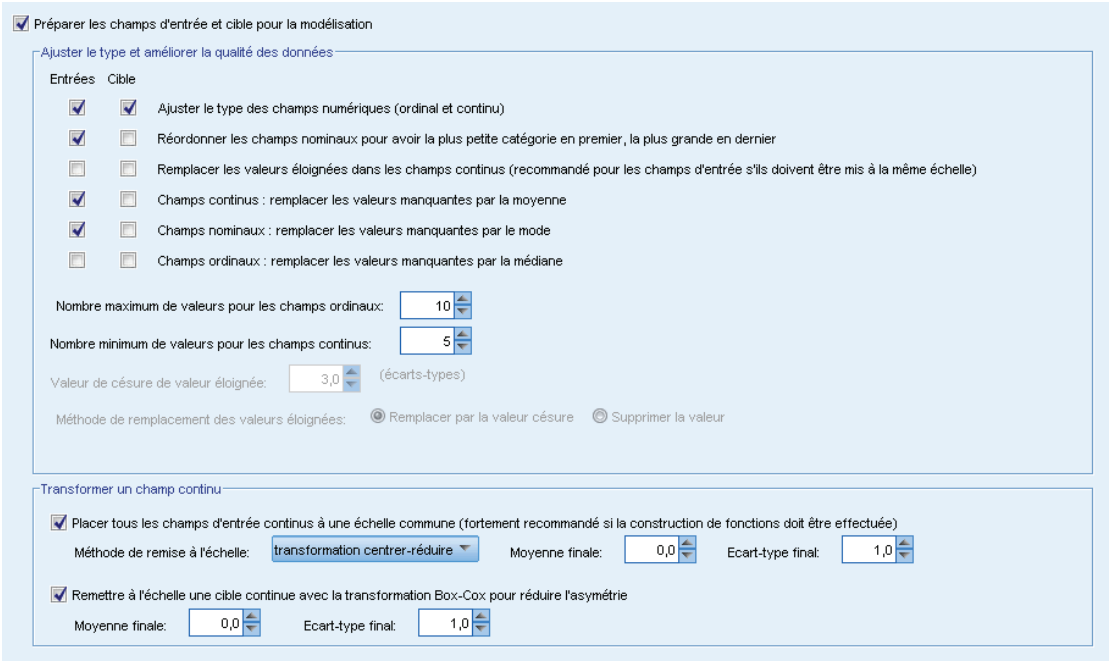
Exclure les champs qualitatifs avec trop de valeurs dans une seule modalité. Les champs ordinaux et nominaux avec une modalité contenant plus que le pourcentage spécifié d'enregistrements sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à désélectionner cette option, et inférieure ou égale à 100, puisque les champs constants sont exclus automatiquement. La valeur par défaut est 95.

Préparation des entrées et des cibles

Aucune donnée n'étant jamais dans un parfait état avant le traitement, vous voudrez sans doute ajuster certains paramètres avant d'exécuter une analyse. Par exemple, vous voudrez supprimer les valeurs éloignées, spécifier la manière de traiter les valeurs manquantes ou encore ajuster le type.

Remarque : Si vous modifiez les valeurs de ce volet, l'onglet Objectifs se met à jour automatiquement et sélectionne l'option Analyse personnalisée.

Figure 4-7
Préparation automatique des données - Paramètres d'entrée et de cible



Préparez les champs d'entrée et les champs cible pour la modélisation. Active ou désactive tous les champs du volet.

Ajuster le type et améliorer la qualité des données. Pour les entrées et la cible, il est possible de spécifier plusieurs transformations de données de manière séparée, si vous ne souhaitez pas modifier les valeurs de la cible. Par exemple, une prévision de revenu en dollars est plus significative qu'une prévision mesurée en log(dollars). En outre, si la cible contient des valeurs manquantes, il n'existe pas de gain prévu pour remplacer les valeurs manquantes, alors que le remplacement des valeurs manquantes en entrée peut permettre à certains algorithmes de traiter des informations qui auraient été perdues.

Des paramètres supplémentaires pour ces transformations, par exemple la valeur de césure des valeurs éloignées, sont communs aux entrées et à la cible.

Vous pouvez sélectionner les paramètres suivants pour les entrées ou la cible ou pour les deux à la fois :

- **Ajuster le type des champs numériques.** Sélectionnez cette option pour déterminer si les champs numériques avec un niveau de mesure *Ordinal* peuvent être convertis en champs *Continus*, et vice versa. Vous pouvez spécifier les valeurs minimale et maximale du seuil pour contrôler la conversion.
- **Réorganiser les champs nominaux.** Sélectionnez cette option pour trier les champs nominaux (ensemble) de la plus petite modalité à la plus grande.
- **Remplacer les valeurs éloignées dans les champs continus.** Spécifiez si vous souhaitez remplacer les valeurs éloignées. Utilisez cette option en conjonction avec les options Méthode de remplacement des valeurs éloignées ci-dessous.
- **Champs continus : remplacer les valeurs manquantes par la moyenne.** Sélectionnez cette option pour remplacer les valeurs manquantes des caractéristiques continues (plage).
- **Champs nominaux : remplacer les valeurs manquantes par le mode.** Sélectionnez cette option pour remplacer les valeurs manquantes des caractéristiques nominales (ensemble).
- **Champs ordinaux : remplacer les valeurs manquantes par la médiane.** Sélectionnez cette option pour remplacer les valeurs manquantes des caractéristiques ordinales (ensemble ordonné).

Le nombre maximum de valeurs pour les champs ordinaux. Spécifiez le seuil pour la redéfinition des champs ordinaux (ensemble ordonné) en champs continus (plage). La valeur par défaut est 10. Si un champ ordinal contient plus de 10 catégories, il est redéfini en champ continu (plage).

Le nombre minimum de valeurs pour les champs continus. Spécifiez le seuil pour la redéfinition des champs continus ou d'échelle (plage) en champs ordinaux (ensemble ordonné). La valeur par défaut est 5. Si un champ continu contient plus de 5 valeurs, il est redéfini en champ ordinal (ensemble ordonné).

Valeur de césure des valeurs éloignées. Spécifiez la valeur de césure des valeurs éloignées, mesurée dans les écarts-types. La valeur par défaut est 3.

Méthode de remplacement des valeurs éloignées. Choisissez si les valeurs éloignées doivent être remplacées (en les tronquant de force) par la valeur de césure ou supprimées et définies comme valeurs manquantes. Les valeurs éloignées définies comme valeurs manquantes suivent les paramètres de traitement des valeurs manquantes sélectionnés ci-dessus.

Attribuer la même échelle à tous les champs d'entrée continus. Pour normaliser les champs d'entrée continus, cochez cette case et choisissez la méthode de normalisation. La méthode par défaut est transformation en score z, pour laquelle vous pouvez spécifier la moyenne finale, dont la valeur

par défaut est 0, et l'écart -type final, dont la valeur par défaut est 1. Sinon, vous pouvez choisir d'utiliser l'option Transformation min/max et spécifier les valeurs minimum et maximum, dont les valeurs par défaut respectives sont 0 et 100.

Ce champ est particulièrement utile lorsque vous sélectionnez l'option Exécuter la construction des caractéristiques. dans le volet Construire et Sélectionner les caractéristiques.

Redimensionner une cible continue avec la transformation de Box-Cox. Pour normaliser un champ cible continu (d'échelle ou de plage), cochez cette case. La transformation de Box-Cox possède une valeur par défaut de 0 pour la moyenne finale et de 1 pour l'écart-type final.

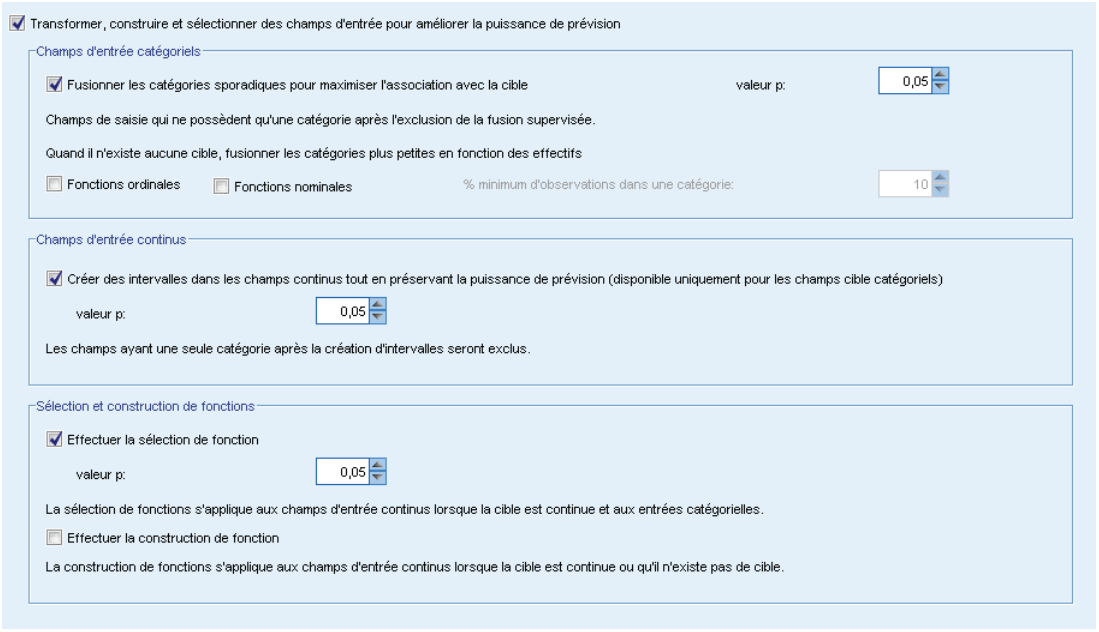
Remarque : Si vous choisissez de normaliser la cible, sa dimension sera transformée. Dans ce cas, vous pourriez avoir besoin de générer un noeud Calculer pour appliquer une transformation inverse et redonner un format reconnaissable aux unités transformées pour un traitement ultérieur. Pour plus d'informations, reportez-vous à la section Génération d'un noeud Calculer sur p. 134.

Construction et sélection des caractéristiques

Pour améliorer la puissance de prédiction de vos données, vous pouvez transformer les champs d'entrées ou en construire de nouveaux basés sur les champs existants.

Remarque : Si vous modifiez les valeurs de ce volet, l'onglet Objectifs se met à jour automatiquement et sélectionne l'option Analyse personnalisée.

Figure 4-8
Préparation automatique des données - Paramètres de transformation, de construction et de sélection



Transformation, construction et sélection de champs d'entrée pour améliorer la puissance de prédiction. Active ou désactive tous les champs du volet.

Fusionner les modalités éparpillées pour optimiser l'association avec une cible. Sélectionner cette option pour créer un modèle plus petit en réduisant le nombre de variables à traiter en association avec la cible. Si nécessaire, modifiez la valeur de probabilité dont la valeur par défaut est de 0,05.

Remarque : si toutes les catégories sont fusionnées en une seule, les versions d'origine et dérivées du champ sont exclues car elles n'ont pas de valeur de variable prédite

Lorsqu'il n'existe aucune cible, fusionner les modalités éparpillées en fonction de leur nombre. Si l'ensemble de données n'a pas de cible, vous pouvez choisir de fusionner les modalités éparpillées des champs ordinaux (ensemble ordonné) ou nominaux (ensemble) ou des deux à la fois. Spécifiez le pourcentage minimum d'observations, ou d'enregistrements dans les données, qui identifie les catégories à fusionner. La valeur par défaut est 10.

Les catégories sont fusionnées en utilisant les règles suivantes :

- La fusion n'est pas réalisée sur les champs binaires.
- Lorsqu'il n'y a que deux catégories à fusionner, la fusion est interrompue.
- La fusion est interrompue s'il n'existe pas de catégorie d'origine, ni de catégorie créée durant la fusion, avec un pourcentage d'observations inférieur au pourcentage minimum spécifié.

Regrouper les champs continus tout en conservant la puissance de prédiction. Si l'ensemble de données comprend une cible qualitative, vous pouvez regrouper les entrées continues ayant de fortes associations pour améliorer les performances du traitement. Si nécessaire, modifiez la valeur de probabilité des sous-ensembles homogènes dont la valeur par défaut est de 0,05.

Si l'opération de regroupement génère un regroupement unique pour un champ spécifique, les versions d'origine et regroupées du champ sont exclues car elles n'ont pas de valeur de variable indépendante.

Remarque : Le regroupement dans l'ADP est différent du regroupement optimal utilisé dans les autres parties de IBM® SPSS® Modeler. Le regroupement optimal utilise des informations d'entropie pour convertir une variable continue en une variable qualitative ; il doit trier les données et les stocker dans la mémoire. L'ADP utilise des sous-ensembles homogènes pour regrouper une variable continue. Cela signifie que le regroupement ADP n'a pas besoin de trier les données et ne stocke pas toutes les données dans une mémoire. L'utilisation de la méthode des sous-ensembles homogènes pour regrouper une variable continue signifie que le nombre de modalités après le regroupement est toujours inférieur ou égal au nombre de modalités de la cible.

Exécuter la sélection des caractéristiques. Sélectionnez cette option pour supprimer les caractéristiques dont le coefficient de corrélation est faible. Si nécessaire, modifiez la valeur de probabilité dont la valeur par défaut est de 0,05.

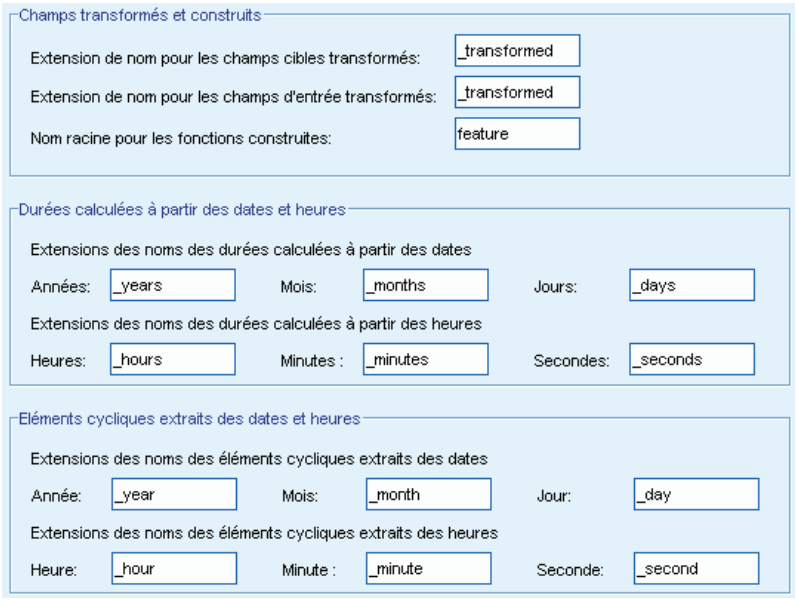
Cette option s'applique uniquement aux caractéristiques d'entrée continues où la cible est continue et aux caractéristiques d'entrée qualitatives.

Exécuter la construction des caractéristiques. Sélectionner cette option pour dériver de nouvelles caractéristiques d'une combinaison de plusieurs caractéristiques existantes (qui sont ensuite supprimées de la modélisation).

Cette option s'applique uniquement aux caractéristiques d'entrée continues où la cible est continue ou lorsqu'il n'y a pas de cible.

Noms de champ

Figure 4-9
Paramètres Nommer les champs de la préparation automatique des données



Pour identifier facilement les caractéristiques nouvelles et transformées, l'ADP crée et applique de nouveaux noms, préfixes ou suffixes de base. Vous pouvez modifier ces noms pour qu'ils soient plus adaptés à vos propres besoins et données. Si vous souhaitez spécifier d'autres étiquettes, vous devez le faire dans le noeud Type en aval.

Champs transformés et construits. Spécifiez les extensions de nom à appliquer aux champs cibles et d'entrées transformés.

Veuillez noter que le noeud ADP définissant les champs de chaîne pour qu'ils soient vides, peut provoquer une erreur en fonction du traitement accordé aux champs non utilisés. Si Comment traiter les champs exclus de la modélisation est défini sur Eliminer les champs non utilisés dans le panneau Paramètres des champs de l'onglet Paramètres, les extensions de nom des entrées et de la cible peuvent être définies sur rien. Les champs d'origine sont éliminés et les champs transformés sont enregistrés à leur place. Dans ce cas, les nouveaux champs transformés auront le même nom que vos champs d'origine.

Cependant, si vous choisissez de paramétrer Définir la direction des champs non utilisés sur 'Aucune', alors les extensions de nom de la cible et des entrées nulles ou vides provoqueront une erreur car vous essaierez de créer des noms de champ en doublon.

En outre, spécifiez le nom du préfixe à appliquer aux caractéristiques construites à l'aide des paramètres Sélectionner et Construire. Le nouveau nom est créé en ajoutant un suffixe numérique à ce nom racine du préfixe. Le format du nombre dépend du nombre de nouvelles caractéristiques dérivées, par exemple :

- si 1 à 9 caractéristiques sont construites, elles seront nommées : caractéristique1 à caractéristique9.
- si 10 à 99 caractéristiques sont construites, elles seront nommées : caractéristique01 à caractéristique99.
- si 100 à 999 caractéristiques sont construites, elles seront nommées : caractéristique001 à caractéristique999, etc.

Cela permet que les caractéristiques construites soient triées dans un ordre cohérent quel que soit leur nombre.

Durées calculées à partir des dates et heures. Spécifier les extensions de nom à appliquer aux durées calculées à partir des dates et des heures.

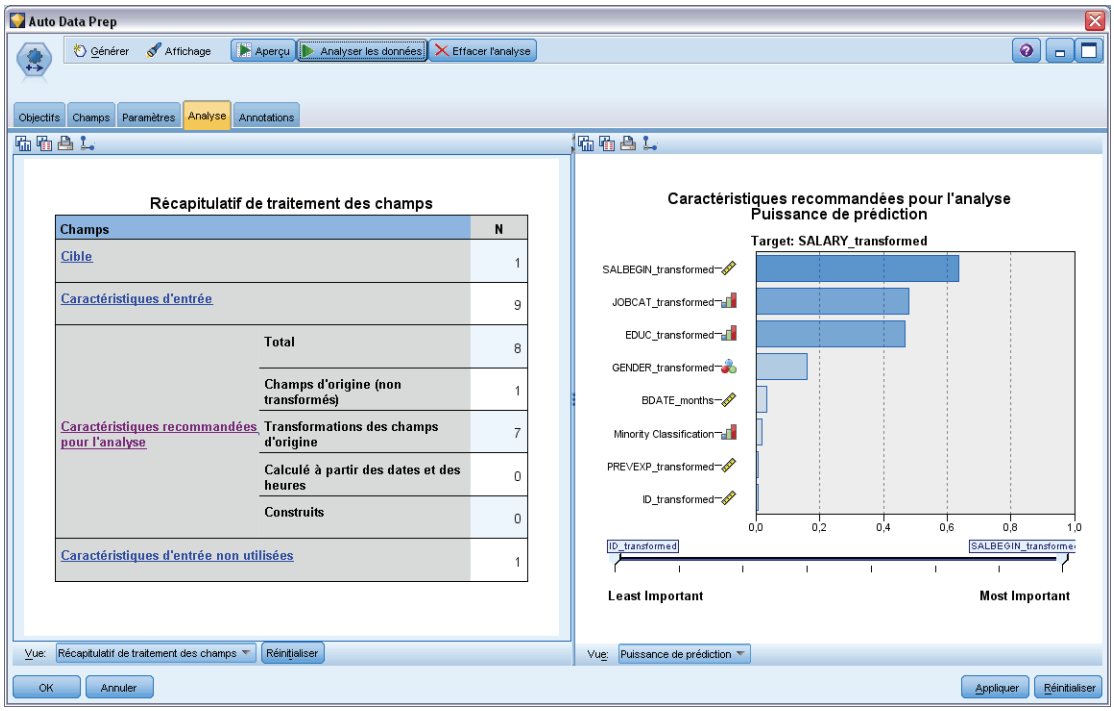
Eléments cycliques extraits des dates et heures. Spécifier les extensions de nom à appliquer aux éléments cycliques extraits des dates et des heures.

Onglet Analyse

▶ Lorsque les paramètres d'ADP vous conviennent, y compris les modifications effectuées dans les onglets Objectif, Champs et Paramètres, cliquez sur Analyser les données. L'algorithme applique les paramètres aux entrées de données et affiche les résultats dans l'onglet Analyse.

L'onglet Analyse contient à la fois des résultats en tableaux et des résultats graphiques qui résument le traitement de vos données et affichent les recommandations sur la façon de modifier ou d'améliorer les données pour l'évaluation. Vous pouvez ensuite revoir puis accepter ou refuser ces recommandations.

Figure 4-10
Onglet Analyse de la préparation automatique des données



L'onglet Analyse est composé de deux panneaux, la vue principale à gauche et la vue liée, ou auxiliaire, à droite. Il existe trois vues principales :

- Récapitulatif de traitement des champs (par défaut). Pour plus d'informations, reportez-vous à la section Récapitulatif de traitement des champs sur p. 123.
- Champs. Pour plus d'informations, reportez-vous à la section Champs sur p. 124.
- Récapitulatif des actions. Pour plus d'informations, reportez-vous à la section Récapitulatif des actions sur p. 126.

Il existe quatre vues liées/auxiliaires :

- Puissance de prédiction (par défaut). Pour plus d'informations, reportez-vous à la section Puissance de prédiction sur p. 127.
- Tableau des champs. Pour plus d'informations, reportez-vous à la section Tableau des champs sur p. 128.
- Détails des champs. Pour plus d'informations, reportez-vous à la section Détails des champs sur p. 129.
- Détails des actions. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 131.

Liens entre les vues

Dans la vue principale, le texte souligné dans les tableaux contrôle ce qui apparaît dans la vue liée. Si vous cliquez sur ces parties de texte, vous obtenez des détails sur un champ, un ensemble de champs ou une étape de traitement spécifique. Le lien que vous avez sélectionné en dernier apparaît en une couleur plus foncée qui permet d'identifier la connexion entre les contenus des deux panneaux de la vue.

Réinitialisation des vues

Pour afficher de nouveau les recommandations d'analyse d'origine et abandonner les modifications effectuées sur les vues Analyse, cliquez sur Réinitialiser au bas du panneau de la vue principale.

Récapitulatif de traitement des champs

Figure 4-11
Récapitulatif du traitement de champ

Récapitulatif de traitement des champs

Champs		N
Cible		1
Caractéristiques d'entrée		9
Caractéristiques recommandées pour l'analyse	Total	8
	Champs d'origine (non transformés)	1
	Transformations des champs d'origine	7
	Calculé à partir des dates et des heures	0
	Construits	0
Caractéristiques d'entrée non utilisées		1

Le tableau Récapitulatif de traitement des champs fournit un instantané de l'impact du traitement général projeté, y compris les modifications de l'état des caractéristiques et le nombre de caractéristiques construites.

Veuillez noter que le modèle est bien construit, et que par conséquent il n'y a pas de mesure ou de diagramme de la modification de la puissance prédictive générale avant et après la préparation des données. Par contre, vous pouvez afficher les diagrammes de la puissance de prédiction des variables indépendantes individuelles recommandées.

Le tableau affiche les informations suivantes :

- le nombre de champs cibles.

- Le nombre de valeurs prédites d'entrée d'origine.
- Les valeurs prédites recommandées pour l'analyse et la modélisation. Cela comprend le nombre total de champs recommandés, le nombre de champs non transformés d'origine recommandés, le nombre de champs transformés recommandés (à l'exclusion des versions intermédiaires de champ, des champs calculés à partir des valeurs prédites date/heure et des valeurs prédites construites), le nombre de champs recommandés dérivés des champs date/heure et le nombre de valeurs prédites construites.
- Le nombre de valeurs prédites d'entrée non recommandées quelle que soit leur forme, que ce soit sous leur forme d'origine, comme champ dérivé, ou comme entrée d'une valeur prédite construite.

Lorsque des informations sur les champs sont soulignées, cliquez pour afficher plus de détails dans une vue liée. Les détails de la Cible, des Caractéristiques d'entrée, et des Caractéristiques d'entrée non utilisées apparaissent dans la vue liée Tableau des champs. Pour plus d'informations, reportez-vous à la section Tableau des champs sur p. 128. Les Caractéristiques recommandées pour l'analyse apparaissent dans la vue liée Puissance de prédiction. Pour plus d'informations, reportez-vous à la section Puissance de prédiction sur p. 127.

Champs

Figure 4-12
Champ

La vue principale Champs affiche les champs traités et si l'ADP recommande de les utiliser dans les modèles en aval. Vous pouvez ignorer les recommandations pour n'importe quel champ ; par exemple, exclure les caractéristiques construites ou inclure les caractéristiques que l'ADP recommande d'exclure. Si un champ a été transformé, vous pouvez décider d'accepter ou non la transformation suggérée ou d'utiliser ou non la version d'origine.

La vue Champs est composée de deux tableaux, un pour la cible et un pour les valeurs prédites qui ont été traitées ou créées.

Tableau Cible

Le tableau Cible n'apparaît que si une cible est définie dans les données.

Ce tableau contient deux colonnes :

- **Nom.** C'est le nom ou l'étiquette du champ cible ; le nom d'origine est toujours utilisé, même si le champ a été transformé.
- **Niveau de mesure.** Ceci affiche l'icône représentant le niveau de mesure. Placez la souris sur l'icône pour afficher une étiquette (continu, ordinal, nominal, etc.) qui décrive les données.

  Si la cible a été transformée, la colonne Niveau de mesure reflète la version transformée finale. *Remarque* : vous ne pouvez pas désactiver les transformations pour la cible.

Tableau Valeurs prédites

Le tableau Valeurs prédites est affiché en permanence. Chaque ligne du tableau représente un champ. Les lignes sont triées par défaut dans l'ordre décroissant de la puissance de prédiction.

Pour les caractéristiques ordinaires, le nom d'origine est toujours utilisé comme nom de ligne. Les versions d'origine et dérivée des champs date/heure apparaissent dans le tableau (dans des lignes séparées) ; le tableau contient également des valeurs prédites construites.

Veuillez noter que les versions transformées des champs apparaissant dans le tableau représentent toujours les versions finales.

Par défaut, seuls les champs recommandés sont affichés dans le tableau des valeurs prédites. Pour afficher les champs restants, sélectionnez la boîte de dialogue Inclure les champs non recommandés dans le tableau au-dessus du tableau ; ces champs sont ensuite affichés au bas du tableau.

Le tableau contient les colonnes suivantes :

- **Version à utiliser.** Affiche une liste déroulante qui contrôle l'utilisation d'un champ en aval et s'il faut utiliser les transformations recommandées. Par défaut, la liste déroulante reflète les recommandations.

  Pour les valeurs prédites ordinaires qui ont été transformées, la liste déroulante contient trois choix : Transformée, D'origine, et Ne pas utiliser.

  Pour les valeurs prédites non transformées ordinaires, les choix sont : D'origine et Ne pas utiliser.

Pour les champs dérivés date/heure et les valeurs prédites construites, les choix sont : Transformée et Ne pas utiliser.

Pour les champs de date d'origine, la liste déroulante est désactivée et définie sur Ne pas utiliser.

Remarque : Pour les valeurs prédites contenant à la fois les versions d'origine et transformées, passer des versions d'origine aux versions transformées met automatiquement à jour les paramètres Niveau de mesure et Puissance de prédiction pour ces caractéristiques.

- **Nom.** Chaque nom de champ est un lien. Cliquez sur un nom pour afficher plus d'informations sur le champ dans la vue liée. Pour plus d'informations, reportez-vous à la section Détails des champs sur p. 129.
- **Niveau de mesure.** Affiche l'icône représentant le type de données ; passez la souris sur l'icône pour afficher une étiquette (continu, ordinal, nominal, etc.) qui décrive les données.
- **Puissance de prédiction.** La puissance de prédiction est affichée uniquement pour les champs recommandés par l'ADP. Cette colonne n'apparaît pas si aucune cible n'est définie. La puissance de prédiction est comprise entre 0 et 1, les valeurs les plus élevées, indiquant des variables indépendantes de « meilleure » qualité. En général, la puissance de prédiction est utile pour comparer les variables indépendantes dans une analyse ADP, mais les valeurs de la puissance de prédiction ne doivent pas être comparées entre des analyses différentes.

Récapitulatif des actions

Figure 4-13
Récapitulatif des actions

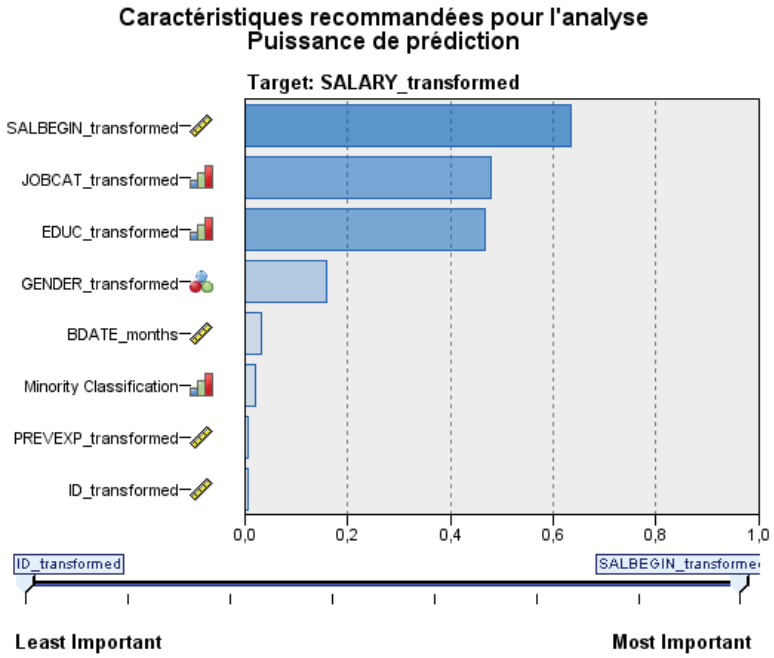
Pour chaque action effectuée par la préparation automatique des données, les valeurs prédites d'entrée sont transformées et/ou supprimées ; les champs qui survivent à une action sont utilisés à la suivante. Les champs qui survivent jusqu'à la dernière étape sont ensuite recommandés pour la modélisation, alors que les entrées des valeurs prédites transformées et construites sont supprimées.

Le Récapitulatif des actions est un simple tableau qui répertorie les actions effectuées par l'ADP. Lorsqu'une Action est soulignée, vous pouvez cliquer dessus pour afficher plus de détails sur les actions effectuées dans une vue liée. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 131.

Remarque : Seules les versions d'origine et transformées finales de chaque champ sont affichées, et pas les versions intermédiaires utilisées pendant l'analyse.

Puissance de prédiction

Figure 4-14
Puissance de prédiction



Affichée par défaut au début de l'analyse ou lorsque vous sélectionnez Valeurs prédites recommandées pour l'analyse dans la vue principale Récapitulatif du traitement des champs, le diagramme affiche la puissance de prédiction des valeurs prédites recommandées. Les champs sont triés par puissance de prédiction, avec le champ ayant la plus haute valeur apparaissant en premier.

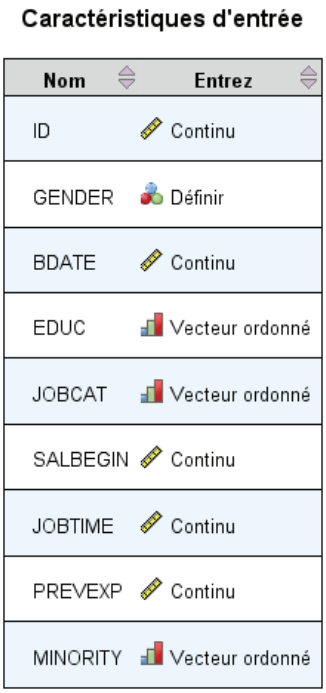
Pour les versions transformées des valeurs prédites ordinaires, le nom des champs reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

Les icônes de niveau de mesure sont affichées après les noms de champ individuels.

La puissance de prédiction de chaque valeur prédite recommandée est calculée à partir d'une régression linéaire ou d'un modèle de Naïve Bayes selon que la cible est continue ou qualitative.

Tableau des champs

Figure 4-15
Tableau des champs



La vue Tableau des champs est un simple tableau qui répertorie les caractéristiques importantes et qui apparaît lorsque vous cliquez sur Cible, Valeurs prédites, ou Valeurs prédites non utilisées dans la vue principale Récapitulatif du traitement des champs.

Ce tableau contient deux colonnes :

- **Nom.** Nom de la valeur prédite.

  Pour les cibles, l'étiquette ou le nom d'origine du champ est utilisé, même si la cible a été transformée.

  Pour les versions transformées des valeurs prédites ordinaires, le nom reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

  Pour les champs dérivés des dates et des heures, le nom de la version transformée finale est utilisé ; par exemple : *bdate_years*.

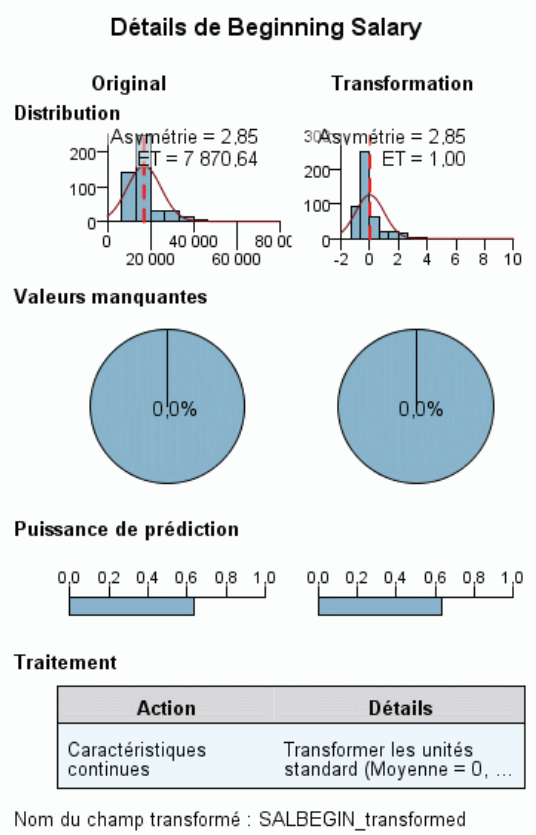
Pour les valeurs prédites construites, le nom de la valeur prédite construite est utilisée ; par exemple : *Valeur prédite1*.

- **Niveau de mesure.** Affiche l'icône représentant le type de données.

  Pour la cible, le Niveau de mesure reflète toujours la version transformée (si la cible a été transformée), par exemple, changée d'ordinale (ensemble ordonné) à continue (plage, échelle) et vice versa.

Détails des champs

Figure 4-16
Détails des champs



La vue Détails des champs contient les diagrammes de distribution, des valeurs manquantes et de la puissance de prédiction (le cas échéant) pour le champ sélectionné et s'affiche lorsque vous cliquez sur un Nom de la vue principale Champs. De plus, l'historique du traitement pour le champ et le nom du champ transformé apparaissent également (le cas échéant).

Pour chaque ensemble de diagrammes, deux versions apparaissent côte à côte pour comparer le champ avec et sans transformations appliquées ; si aucune version transformée du champ n'existe, un diagramme apparaît pour la version d'origine uniquement. Pour les champs de date ou d'heure dérivés et les valeurs prédites construites, les diagrammes n'apparaissent que pour la nouvelle valeur prédite.

Remarque : Si un champ est exclu parce qu'il contient trop de modalités, seul l'historique de traitement apparaît.

Diagramme de distribution

La distribution des champs continus apparaît dans un histogramme, avec une courbe normale superposée et une ligne de référence verticale pour la valeur moyenne ; les champs qualitatifs apparaissent sous forme de diagramme en bâtons.

Les histogrammes sont étiquetés pour montrer l'écart-type et l'asymétrie, toutefois l'asymétrie n'apparaît pas si le nombre des valeurs est inférieur ou égal à 2 ou si la variance du champ d'origine est inférieure à 10-20.

Passez la souris sur le diagramme pour afficher la moyenne des histogrammes ou le nombre et le pourcentage du nombre total d'enregistrements des modalités dans les diagrammes en bâtons.

Diagramme des valeurs manquantes

Les diagrammes en secteurs comparent le pourcentage des valeurs manquantes avec et sans transformations appliquées ; les étiquettes de diagramme indiquent le pourcentage.

Si l'ADP traite les valeurs manquantes, le diagramme en secteurs après la transformation comprend la valeur de remplacement comme étiquette, c'est-à-dire la valeur utilisée à la place des valeurs manquantes.

Passez la souris sur le diagramme pour afficher le nombre des valeurs manquantes et le pourcentage du nombre total d'enregistrements.

Diagramme de puissance de prédiction

Pour les champs recommandés, les diagrammes en bâtons affichent la puissance de prédiction avant et après la transformation. Si la cible a été transformée, la puissance de prédiction calculée tient compte de la cible transformée.

Remarque : Les diagrammes de puissance de prédiction ne sont pas affichés si aucune cible n'est définie, ou si la cible est atteinte depuis le panneau de la vue principale.

Passez la souris sur le diagramme pour afficher la valeur de la puissance de prédiction.

Tableau des historiques du traitement

Ce tableau indique la façon dont la version transformée d'un champ a été dérivée. Les actions entreprises par l'ADP sont répertoriées dans l'ordre dans lequel elles ont été exécutées ; mais, pour certaines étapes, plusieurs actions ont pu être exécutées pour un champ particulier.

Remarque : Ce tableau n'apparaît pas pour les champs qui n'ont pas été transformés.

Les informations du tableau sont divisées en trois colonnes :

- **Action.** Le nom de l'action. Par exemple, Valeurs prédites continues. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 131.
- **Détails.** La liste des traitements effectués. Par exemple, Transformer en unités standard.
- **Fonction.** Apparaît uniquement pour les valeurs prédites construites et affiche la combinaison linéaire de champs d'entrée, par exemple, 0,06*âge + 1,21*hauteur.

Détails des actions

Figure 4-17
Analyse ADP - Détails des actions

Étape 9 : Caractéristiques continues

Transformation	Nombre de caractéristiques	Critères	
		Moyenne	SD
Transformer en unités standard	5	0	1

Construction d'espace de caractéristiques	N
Caractéristiques construites	0
Caractéristiques exclues en raison d'une faible association avec une cible	1
Caractéristiques exclues parce qu'elles étaient constantes après le regroupement	0

La vue liée Détails des actions apparaît lorsque vous cliquez sur Action dans la vue principale Récapitulatif des actions. La vue liée Détails des actions affiche des informations relatives aux actions et des informations communes pour chaque étape de traitement effectuée. Les détails relatifs à chaque action spécifique apparaissent d'abord.

La description de chaque action est utilisée comme titre en haut de la vue liée. Les détails relatifs à chaque action sont affichés sous le titre, et peuvent contenir des détails sur le nombre de valeurs prédites dérivées, de champs reconvertis, de transformations de cible, de modalités fusionnées ou réorganisées et de valeurs prédites construites ou exclues.

Au cours du traitement des actions, le nombre de valeurs prédites utilisées pour le traitement peut varier, par exemple lorsque des valeurs prédites sont exclues ou fusionnées.

Remarque : Si une action est désactivée ou qu'aucune cible n'est spécifiée, un message d'erreur apparaît à la place des détails de l'action lorsque vous cliquez dessus dans la vue principale Récapitulatif des actions.

Il existe neuf actions possibles, toutefois, toutes ne sont pas nécessairement actives pour chaque analyse.

tableau Champs de texte

Ce tableau affiche le nombre :

- d'espaces de droite éliminés.
- de valeurs prédites exclues de l'analyse.

Tableau Valeurs prédites de date et d'heure

Ce tableau affiche le nombre :

- des durées dérivées des valeurs prédites de date et d'heure.
- d'éléments Date et heure.
- de valeurs prédites de date et d'heure dérivées, au total.

La date ou heure de référence est affichée comme note de bas de page si des durées de date ont été calculées.

Tableau Filtrage des valeurs prédites

Ce tableau affiche le nombre des valeurs prédites suivantes exclues du traitement :

- constantes.
- valeurs prédites avec trop de valeurs manquantes.
- valeurs prédites avec trop d'observations dans une seule modalité.
- Champs nominaux (ensembles) avec trop de modalités.
- valeurs prédites supprimées, au total.

Vérifier le tableau de niveau de mesure

Ce tableau affiche le nombre de champs reconvertis, répartis selon les catégories suivantes :

- Champs ordinaux (ensembles ordonnés) reconvertis en champs continus.
- Champs continus reconvertis en champs ordinaux.
- Nombre total des champs reconvertis.

Si aucun champ d'entrée (cible ou valeurs prédites) n'est continu ou ordinal, cela est indiqué en note de bas de page.

tableau Valeurs éloignées

Ce tableau affiche le nombre de valeurs éloignées traitées.

- soit le nombre de champs continus pour lesquels des valeurs éloignées ont été recherchées et éliminées, ou le nombre de champs continus pour lesquels les valeurs éloignées ont été recherchées et définies sur manquantes, en fonction de vos paramètres dans le panneau Préparer les entrées & la cible dans l'onglet Paramètres.
- le nombre de champs continus exclus parce qu'ils étaient constants après le traitement des valeurs éloignées.

Une note de bas de page indique la valeur de césure des valeurs éloignées et une autre note de bas de page apparaît si aucun champ d'entrée (cible ou de valeurs prédites) n'est continu.

tableau Valeurs manquantes

Ce tableau affiche le nombre de champs qui contenaient des valeurs manquantes remplacées, selon les catégories suivantes :

- Cible. Cette ligne n'apparaît pas si aucune cible n'est spécifiée.
- Valeurs prédites. Elles sont divisées en nombre de champs nominaux (ensemble), ordinaux (ensemble ordonné) et continus.
- Le nombre total de valeurs manquantes remplacées.

Tableau Cible

Ce tableau indique si la cible a été transformée :

- transformation de Box-Cox en normalité. Cette catégorie est elle-même divisée en colonnes qui indiquent le critère spécifié (moyenne et écart-type) et le Lambda.
- modalités cibles réorganisées pour améliorer la stabilité.

Tableau des valeurs prédites catégorielles

Ce tableau affiche le nombre de valeurs prédites catégorielles :

- dont les modalités ont été réorganisées de Faible à Elevé pour améliorer la stabilité.
- dont les modalités ont été fusionnées pour optimiser l'association avec la cible.
- dont les modalités ont été fusionnées pour traiter les modalités éparpillées.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après la fusion.

Une note de bas de page apparaît si aucune valeur prédite catégorielle n'existe.

Tableau des valeurs prédites continues

Il existe deux tableaux. Le premier affiche une des transformations suivantes :

- les valeurs des variables prédites transformées en unités standard. De plus, il indique le nombre de valeurs prédites transformées, la moyenne spécifiée et l'écart-type.
- Les valeurs des variables prédites mappées sur un intervalle commun. De plus, il indique le nombre de valeurs prédites transformées utilisant une transformation min-max, ainsi que les valeurs minimum et maximum spécifiées.
- les valeurs des valeurs prédites et le nombre de valeurs prédites regroupées.

Le deuxième tableau affiche les détails de construction de l'espace des valeurs prédites, sous la forme du nombre de valeurs prédites :

- construites.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après le regroupement.
- exclues parce qu'elles étaient constantes après la construction.

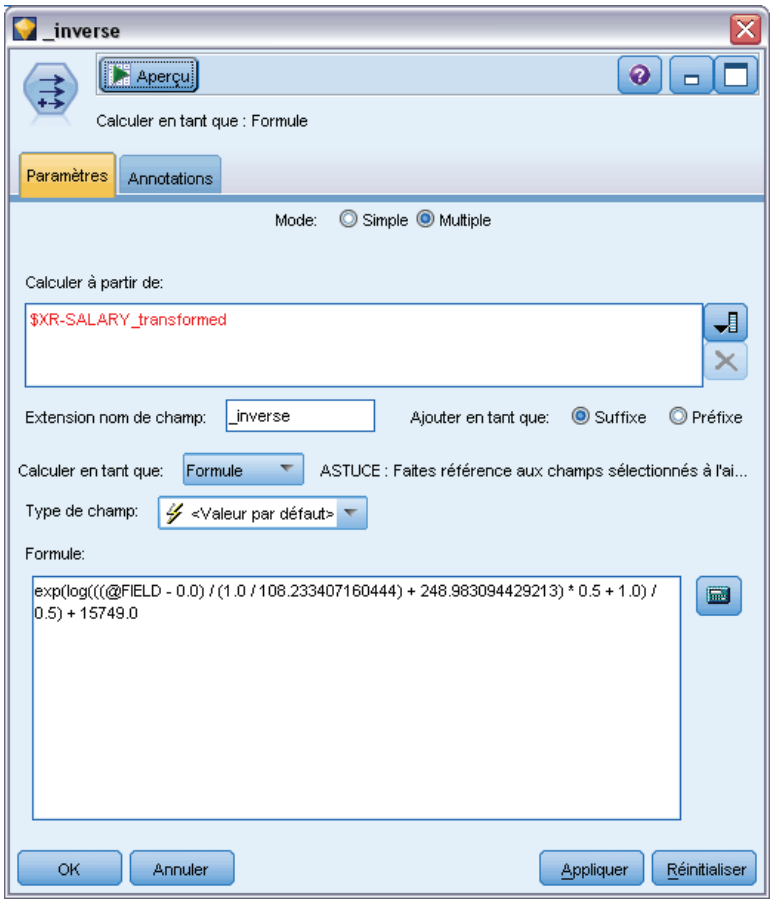
Une note de bas de page apparaît si aucune valeur prédite continue n'a été saisie.

Génération d'un noeud Calculer

Lorsque vous générez un noeud Calculer, il applique la transformation inversée de la cible au champ de scores. Par défaut, le noeud entre le nom du champ de scores qui serait produit par un logiciel de modélisation automatique (comme Auto Classifier ou Auto Numeric) ou par le noeud Ensemble. Si une cible d'échelle (intervalle) a été transformée, le champ de scores apparaît en unités transformées ; par exemple, log($) à la place de $. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine.

Remarque : Vous ne pouvez générer un noeud Calculer que lorsque le noeud ADP contient une analyse qui rééchelonne une cible plage (c'est-à-dire que le rééchelonnement de Box-Cox est sélectionné dans le panneau Préparer les entrées & la cible). Vous ne pouvez pas générer de noeud Calculer si la cible n'est pas une plage, ou si le rééchelonnement de Box-Cox n'est pas sélectionné.

Figure 4-18
Noeud Calculer généré à partir du noeud Préparation automatique des données



Le noeud Calculer est créé en mode Multiple et utilise @FIELD dans l'expression pour que vous puissiez ajouter la cible transformée si nécessaire. Par exemple, en utilisant les informations suivantes :

- Nom de champ cible : réponse
- Nom de champ cible transformé : response_transformed
- Nom du champ de scores : $XR-response_transformed

Le noeud Calculer créerait un nouveau champ : $XR-response_transformed_inverse.

Remarque : Si vous n'utilisez pas de logiciel de modélisation automatique ou de noeud Ensemble, vous devrez modifier le noeud Calculer pour transformer le bon champ de scores pour votre modèle.

Cibles continues normalisées

Par défaut, si vous sélectionnez la case Rééchelonner une cible continue avec la transformation de Box-Cox dans le panneau Préparer les entrées & la cible, cela transforme la cible et vous créez un nouveau champ qui sera la cible pour la création de votre modèle. Par exemple, si votre cible d'origine était *response*, la nouvelle cible sera *response_transformed*; les modèles en aval du noeud ADP choisiront automatiquement cette nouvelle cible.

Mais, cela peut provoquer des problèmes, en fonction de la cible d'origine. Par exemple, si la cible était *Age*, les valeurs de la nouvelle cible ne seront pas *Années*, mais une version transformée de *Années*. Cela signifie que vous ne pouvez pas consulter les scores et les interpréter car ils ne sont pas présentés en unités reconnaissables. Dans ce cas, vous pouvez appliquer une transformation inverse qui reconvertira vos unités transformées en ce qu'elles devaient être. Pour ce faire :

- Après avoir cliqué sur Analyser les données pour effectuer l'analyse ADP, sélectionnez le *noeud Calculer* dans le menu *Générer*.
- Placez le noeud Calculer après votre nugget sur le canevas des modèles.

Le noeud Calculer restaurera le champ de scores aux dimensions d'origine afin que la prédiction soit effectuée en des valeurs *Années* d'origine.

Par défaut, le noeud Calculer transforme le champ de scores généré par un logiciel de modélisation automatique ou un modèle combiné. Si vous construisez un modèle individuel, vous devez modifier le noeud Calculer pour dériver à partir de votre champ de scores actuel. Si vous souhaitez évaluer votre modèle, vous devez ajouter la cible transformée au champ Calculer à partir de dans le noeud Calculer. Cela applique la même transformation inverse à la cible et les noeuds en aval Evaluation ou Analyse utiliseront les données transformées correctement tant que vous modifiez ces noeuds pour qu'ils utilisent des noms de champs à la place des métadonnées.

Si vous voulez également restaurer le nom d'origine, vous pouvez utiliser un noeud Filtre pour supprimer le champ cible d'origine s'il existe encore et renommer la cible et les champs de scores.

Noeud Typer

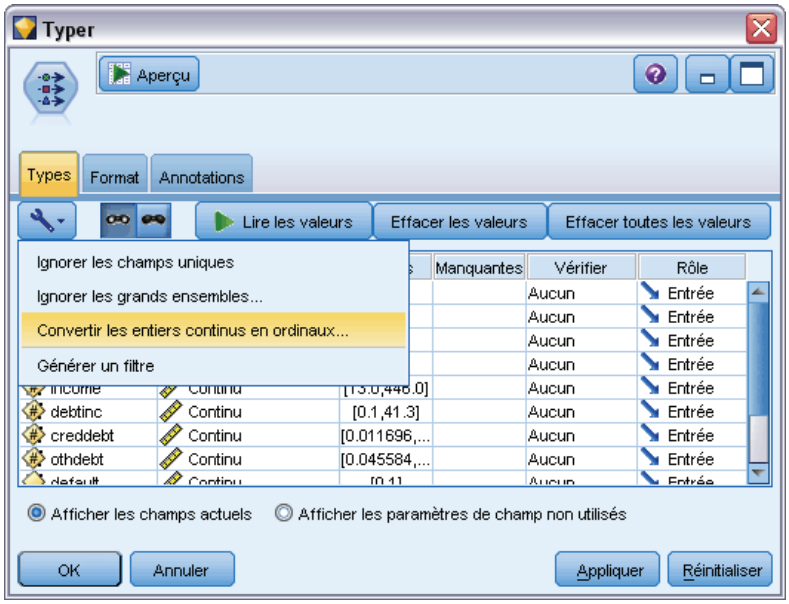
Les propriétés de champ peuvent être indiquées dans un noeud source ou dans un noeud Typer distinct. Les fonctionnalités sont similaires dans les deux noeuds. Les propriétés suivantes sont disponibles :

- **Champ.** Double-cliquez sur un nom de champ pour indiquer les étiquettes de valeur et de champ des données de IBM® SPSS® Modeler. Par exemple, vous pouvez consulter ou modifier ici les métadonnées de champ importées à partir de IBM® SPSS® Statistics. De même, vous pouvez créer des étiquettes pour les champs et leurs valeurs. La présence des étiquettes indiquées ici dans SPSS Modeler dépend des sélections effectuées dans la boîte de dialogue Propriétés du flux.
- **Mesure.** Il s'agit d'un niveau de mesures qui permet de décrire les caractéristiques des données d'un champ précis. Si tous les détails d'un champ sont connus, il est dit **complètement instancié**. Pour plus d'informations, reportez-vous à la section Niveaux de mesure sur p. 138.

Remarque : Le niveau de mesure d'un champ est différent de son type de stockage, qui indique si les données sont stockées sous forme de chaînes, d'entiers, de nombres réels, de dates, d'heures ou d'horodatages.

- **Valeurs :** Cette colonne vous permet de spécifier des options pour la lecture de valeurs de données à partir de l'ensemble de données ou d'utiliser l'option Spécifier afin de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte. Vous pouvez également choisir de transférer les champs sans lire leurs valeurs. Pour plus d'informations, reportez-vous à la section Valeurs de données sur p. 143.
- **Manquant :** Permet de spécifier le traitement des valeurs manquantes du champ. Pour plus d'informations, reportez-vous à la section Définition de valeurs manquantes sur p. 149.
- **Vérifier.** Dans cette colonne, vous pouvez définir des options pour garantir que les valeurs de champ sont conformes aux intervalles ou aux valeurs spécifiés. Pour plus d'informations, reportez-vous à la section Vérification des valeurs de type sur p. 149.
- **Rôle.** Permet d'indiquer aux noeuds de modélisation si les champs sont des champs d'entrée (champs prédicteurs) ou de cible (champs prédits) pour un processus d'apprentissage automatique. Sont également disponibles les rôles Les deux et Aucun, et l'option Partition. Cette dernière signale les champs utilisés pour partitionner les enregistrements en échantillons distincts à des fins d'apprentissage, de test et de validation. La valeur Diviser spécifie que des modèles séparés seront construits pour chaque valeur possible du champ. Pour plus d'informations, reportez-vous à la section Définition du rôle du champ sur p. 150.

Figure 4-19
Options du noeud Typer



Plusieurs autres options peuvent également être spécifiées dans la fenêtre du noeud Typer :

- Les options du menu Outils permettent d'ignorer les champs uniques une fois le nœud Typer instancié (via vos spécifications, la lecture des valeurs ou l'exécution du flux). Si vous choisissez d'ignorer les champs uniques, les champs comportant une seule valeur sont automatiquement ignorés.

- Les options du menu Outils permettent d'ignorer les grands ensembles une fois le noeud Typer instancié. Si vous choisissez d'ignorer les grands ensembles, les ensembles dont le nombre de membres est élevé sont automatiquement ignorés.
- Les options du menu Outils vous permettent de choisir de Convertir les entiers continus en ordinaux une fois le noeud Typer instancié. Pour plus d'informations, reportez-vous à la section Conversion de données continues sur p. 141.
- Les options du menu Outils permettent de créer un noeud Filtrer pour exclure les champs sélectionnés.
- A l'aide des boutons bascule représentant des lunettes de soleil, vous pouvez définir le paramètre par défaut Lire ou Transférer pour tous les champs. L'onglet Types du noeud source transmet les champs par défaut, alors que le noeud Typer lit les valeurs par défaut.
- A l'aide du bouton Effacer les valeurs, vous pouvez supprimer les modifications apportées aux valeurs de champ de ce noeud (valeurs non héritées) et relire les valeurs des opérations effectuées en amont. Cette option est utile pour réinitialiser les changements que vous avez apportés à certains champs en amont.
- A l'aide du bouton Effacer toutes les valeurs, vous pouvez réinitialiser les valeurs de **tous** les champs lus dans le noeud. Cette option paramètre la colonne *Valeurs* de tous les champs sur **Lire**. Cette option est utile pour réinitialiser les valeurs de tous les champs, et relire les valeurs et les types des opérations effectuées en amont.
- Dans le menu contextuel, vous pouvez choisir de copier les attributs d'un champ à l'autre. Pour plus d'informations, reportez-vous à la section Copie d'attributs de type sur p. 152.
- A l'aide de l'option Afficher les paramètres de champ non utilisés, vous pouvez afficher les paramètres de type des champs qui ne figurent plus dans les données ou qui étaient auparavant connectés à ce noeud Typer. Cette option est utile lorsque vous réutilisez un noeud Typer pour des ensembles de données qui ont été modifiés.

Niveaux de mesure

Le niveau de mesure (auparavant appelé « type de données » ou « type d'utilisation ») décrit l'utilisation des champs de données dans IBM® SPSS® Modeler. Le niveau de mesure peut être spécifié sur l'onglet Types d'une source ou d'un noeud Typer. Par exemple, vous pouvez définir le niveau de mesure de nombre entier comportant les valeurs 1 et 0 comme étant un champ *booléen*. En général, 1 correspond à la valeur *True (vrai)* et 0 à la valeur *False (faux)*.

Stockage et mesure. Le type niveau de mesure d'un champ est différent de son type de stockage, lequel indique si les données sont stockées sous la forme d'une chaîne, d'un entier, d'un nombre réel, d'une date, d'une heure ou d'un horodatage. Si les types de données peuvent être modifiés en tout point d'un flux à l'aide d'un noeud Typer, le stockage doit, quant à lui, être déterminé au niveau de la source dans SPSS Modeler (il peut cependant être modifié ultérieurement à l'aide d'une fonction de conversion). Pour plus d'informations, reportez-vous à la section Définition du stockage et du formatage des champs dans le chapitre 2 sur p. 32.

Certains noeuds de modélisation indiquent les types de niveau de mesure autorisés pour leurs champs d'entrée et de sortie à l'aide d'icônes sur leur onglet Champs.

Icones de niveau de mesure

Icône	Le niveau de mesure
	Default
	Continu
	Catégoriel
	Flag
	Nominal
	Ordinal
	Sans type

Les niveaux de mesure suivants sont disponibles :

- **Défaut :** Les données dont le type de stockage et les valeurs sont inconnus (par exemple, parce qu'elles n'ont pas encore été lues) sont affichées en tant que <Par défaut>.
- **Continue.**Permet de décrire les valeurs numériques, par exemple un intervalle de 0 à 100 ou de 0,75 à 1,25. Une valeur continue peut être un entier, un nombre réel ou une date/heure.
- **Qualitatives :** Utilisé pour les valeurs de chaîne lorsque le nombre exact de valeurs distinctes est inconnu. Il s'agit d'un type de données **non instancié**, ce qui signifie que toutes les informations possibles sur le stockage et l'utilisation des données ne sont pas encore connues. Une fois les données lues, le niveau de mesure sera *Booléen*, *Nominal* ou *Sans type*, en fonction du nombre maximal spécifié de membres pour les champs nominaux dans la boîte de dialogue Propriétés du flux.
- **Booléen.**Utilisé pour les données comportant deux valeurs distinctes qui indiquent la présence ou l'absence d'un trait telle que true et false, Yes et No ou 0 et 1. Les valeurs utilisées peuvent varier, mais une d'elles doit toujours être désignée comme valeur « vrai » et l'autre comme valeur « faux ». Vous pouvez représenter les données sous forme de texte, d'entier, de nombre réel, de date, d'heure ou d'horodatage.
- **Nominal :** Utilisé pour décrire les données ayant plusieurs valeurs distinctes, chacune étant traitée en tant que membre d'un ensemble, tel que small/medium/large. Les données nominales peuvent bénéficier de n'importe quel stockage—numérique, chaînes ou date/heure Le fait de définir le niveau de mesure *Nominal* n'a pas pour effet de convertir automatiquement les valeurs en stockage de chaîne.
- **Ordinal :** Utilisé pour décrire les données comportant plusieurs valeurs distinctes ayant un ordre inhérent. Par exemple, les catégories de salaire ou l'indice de satisfaction peuvent avoir le type données ordinales. L'ordre est défini par l'ordre de tri naturel des éléments des données. Par exemple, 1, 3, 5 est l'ordre de tri par défaut d'un ensemble d'entiers, alors que HIGH, LOW, NORMAL (tri alphabétique croissant) est l'ordre d'un ensemble de chaînes. Le niveau de mesure ordinal vous permet de définir un ensemble de données catégorielles comme des données ordinales, pour la visualisation, la création de modèles et l'exportation vers d'autres applications (telles que IBM® SPSS® Statistics) qui reconnaissent les données ordinales comme un type distinct. Vous pouvez utiliser le champ ordinal partout où un champ

nominal peut être utilisé. De plus, les champs de n'importe quel type de stockage (réel, entier, chaîne, date, heure, etc.) peuvent être définis comme ordinal.

- **Sans type.** Utilisé pour les données qui ne sont pas conformes aux types ci-dessus, pour des champs avec une valeur unique ou pour des données nominales où l'ensemble possède davantage de membres que le maximum défini. Ce type s'avère pratique dans les cas où autrement le niveau de mesure serait un ensemble avec de nombreux membres (par exemple, un numéro de compte). Lorsque vous sélectionnez Sans type pour un champ, le rôle est automatiquement défini sur Aucun, avec ID d'enregistrement comme seule alternative. La taille maximale par défaut des ensembles est de 250 valeurs uniques. Ce nombre peut être ajusté ou désactivé dans l'onglet Options de la boîte de dialogue Propriétés du flux, à laquelle vous pouvez accéder à partir du menu Outils.

Vous pouvez indiquer manuellement des niveaux de mesure, ou laisser le logiciel lire les données et déterminer le niveau de mesure en fonction des valeurs lues.

Vous pouvez aussi sélectionner une option, si vous avez plusieurs champs de données continues qui doivent être traitées comme des données catégorielles, afin de les convertir. Pour plus d'informations, reportez-vous à la section Conversion de données continues sur p. 141.

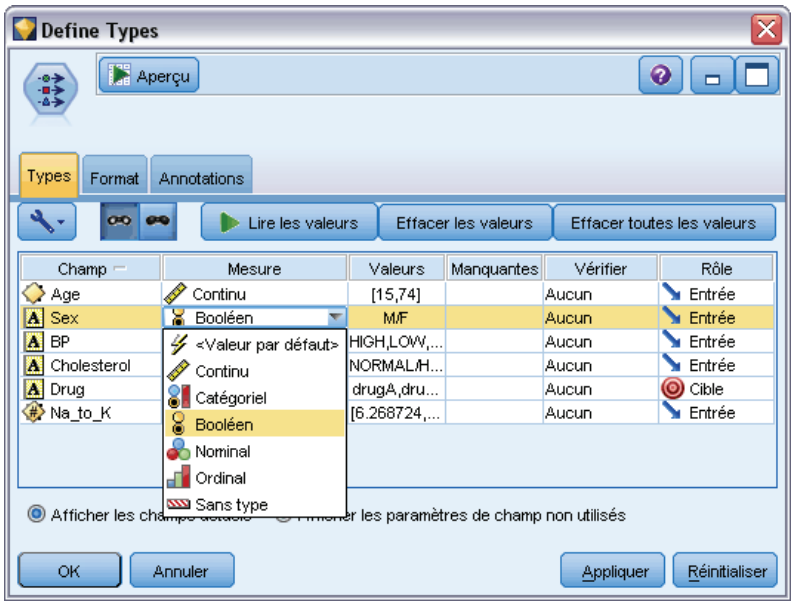
Pour utiliser la définition automatique du type

- Dans un noeud Typer ou dans l'onglet Types d'un noeud source, définissez la colonne *Valeurs* sur <Lire> pour les champs voulus. Les métadonnées sont ainsi disponibles pour tous les noeuds situés en aval. Vous pouvez rapidement paramétrer tous les champs sur <Lire> ou <Transférer> à l'aide des boutons représentant des lunettes de soleil dans la boîte de dialogue.
- Cliquez sur Lire les valeurs pour lire les valeurs directement à partir de la source de données.

Pour définir manuellement le niveau de mesure d'un champ

- Sélectionnez un champ dans le tableau.
- Dans la liste déroulante de la colonne *Mesure*, sélectionnez le niveau de mesure du champ.
- Vous pouvez également utiliser la combinaison Ctrl+A ou Ctrl+clic pour sélectionner plusieurs champs avant de choisir un niveau de mesure dans la liste déroulante.

Figure 4-20
Définition manuelle des niveaux de mesure

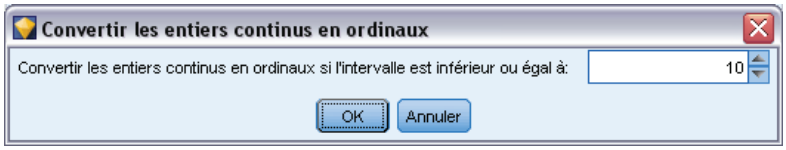


Conversion de données continues

Le traitement de données catégorielles en tant que données continues peut avoir des effets importants sur la qualité d'un modèle, surtout s'il s'agit du champ cible, comme par exemple, la création d'un modèle de régression plutôt que d'un modèle binaire. Pour éviter cet inconvénient, vous pouvez convertir des intervalles d'entiers en des types catégoriels tels que *Ordinal* ou *Booléen*.

▶ Dans le bouton du menu Opérations et Générer (comportant le symbole d'outil), sélectionnez Convertir des entiers continus en ordinaux. La boîte de dialogue des valeurs de conversion s'affiche.

Figure 4-21
Boîte de dialogue Valeurs de conversion



▶ Spécifiez la taille de l'intervalle qui sera automatiquement converti ; cela s'applique à n'importe quel intervalle jusqu'à la taille (incluse) que vous avez saisie.

▶ Cliquez sur OK. Les intervalles concernés sont convertis soit en *Booléen* ou en *Ordinal* et sont affichés dans l'onglet Types du noeud Typer.

Résultats de la conversion

- Lorsqu'un champ *Continu* avec stockage d'entiers est modifié en *Ordinal*, les valeurs inférieures et supérieures sont étendues afin d'inclure toutes les valeurs entières, de la plus basse à la plus élevée. Par exemple, si l'intervalle est 1, 5, l'ensemble des valeurs est 1, 2, 3, 4, 5.
- Lorsque le champ *Continu* est converti en un champ *Booléen*, les valeurs inférieures et supérieures deviennent des valeurs false (faux) et true (vrai) du champ booléen.

Qu'est-ce que l'instanciation ?

L'**instanciation** est le processus qui consiste à lire ou à spécifier des informations, telles que le type de stockage ou les valeurs d'un champ de données. Afin d'optimiser les ressources système, l'instanciation est gérée par l'utilisateur— Celui-ci demande au logiciel de lire les valeurs en spécifiant des options dans l'onglet Types d'un noeud source ou en exécutant des données via un noeud Typer.

- Les données dont le type est inconnu sont par ailleurs désignées comme **non instanciées**. Les données dont les valeurs et le type de stockage sont inconnus figurent dans la colonne *Mesure* de l'onglet Types sous la forme <Par défaut>.
- Lorsque vous disposez d'informations sur le stockage d'un champ (valeur numérique ou chaîne, par exemple), les données sont dites **partiellement instanciées**. Les types Catégoriel et Continu sont des mesures de niveau partiellement instanciés. Par exemple, le type Catégoriel indique que le champ est symbolique, mais vous ne savez pas s'il s'agit du type nominal, ordinal ou booléen.
- Lorsque tous les détails sur un type sont connus, y compris les valeurs, le niveau de mesure **entièrement instancié** —nominal, ordinal, booléen ou continu—est affiché dans cette colonne. *Remarque* : Le type *continu* est utilisé aussi bien pour les champs de données partiellement instanciés que pour ceux entièrement instanciés. Les données continues peuvent être des entiers ou des nombres réels.

Pendant l'exécution d'un flux de données avec un noeud Typer, les types non instanciés deviennent immédiatement partiellement instanciés, en fonction des valeurs de données initiales. Une fois que toutes les données sont passées dans le noeud, elles deviennent complètement instanciées sauf si les valeurs ont été définies sur <Transférer>. Si l'exécution est interrompue, les données demeurent partiellement instanciées. Une fois l'onglet Types instancié, les valeurs des champs sont statiques à cet endroit du flux. Autrement dit, tout changement intervenant en amont n'affectera pas les valeurs d'un champ particulier, même si vous exécutez de nouveau le flux. Pour modifier ou mettre à jour les valeurs en fonction de nouvelles données ou de manipulations supplémentaires, vous devez les éditer directement dans l'onglet Types ou paramétrer la valeur des champs sur <Lire> ou <Lire +>.

Moment d'instanciation

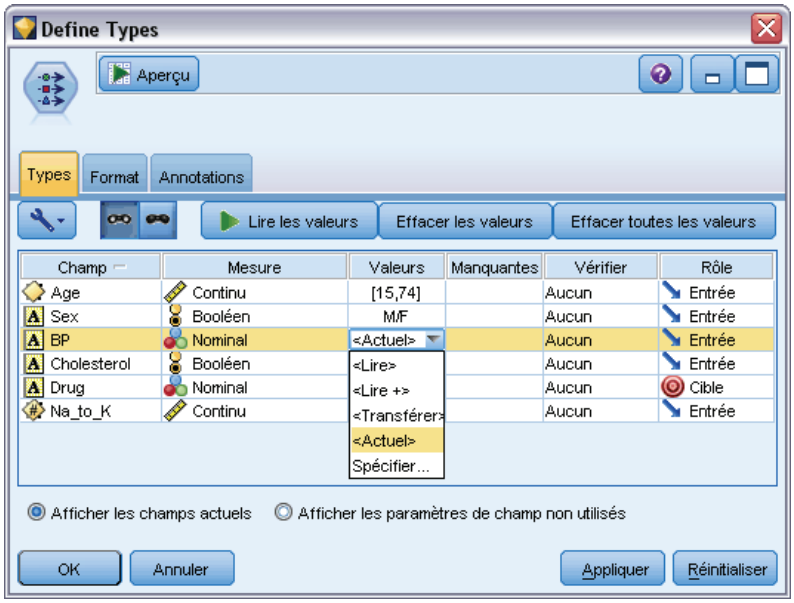
En général, si votre ensemble de données n'est pas trop volumineux et si vous ne prévoyez pas d'ajouter des champs au flux par la suite, l'instanciation au niveau du noeud source est la méthode la plus pratique. Cependant, l'instanciation dans un autre noeud Typer est utile lorsque :

- L'ensemble de données est volumineux et le flux filtre un sous-ensemble avant le noeud Typer.
- Des données ont été filtrées dans le flux.
- Des données ont été fusionnées ou ajoutées dans le flux.
- De nouveaux champs de données sont calculés au cours du traitement.

Valeurs de données

La colonne *Valeurs* de l'onglet Types vous permet de lire automatiquement des valeurs à partir des données, ou de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte.

Figure 4-22
Sélection des méthodes de lecture, de transfert ou de spécification des valeurs de données



Les options disponibles dans cette liste déroulante fournissent les instructions de définition automatique du type suivantes :

Option	Fonction
<Lire>	Les données sont lues lors de l'exécution du noeud.
<Lire +>	Les données sont lues et ajoutées aux données actuelles (le cas échéant).
<Pass>	Aucune donnée n'est lue.
<Current>	Les valeurs des données actuelles sont conservées.
Indiquer...	Une boîte de dialogue distincte est ouverte pour que vous puissiez spécifier des valeurs et des options de niveau de mesure.

Si vous exécutez un noeud Typer ou que vous cliquez sur Lire les valeurs, une définition automatique du type a lieu et les valeurs sont lues à partir de votre source de données, en fonction de votre sélection. Vous pouvez également spécifier ces valeurs manuellement en utilisant l'option Spécifier ou en double-cliquant sur une cellule de la colonne *Champ*.

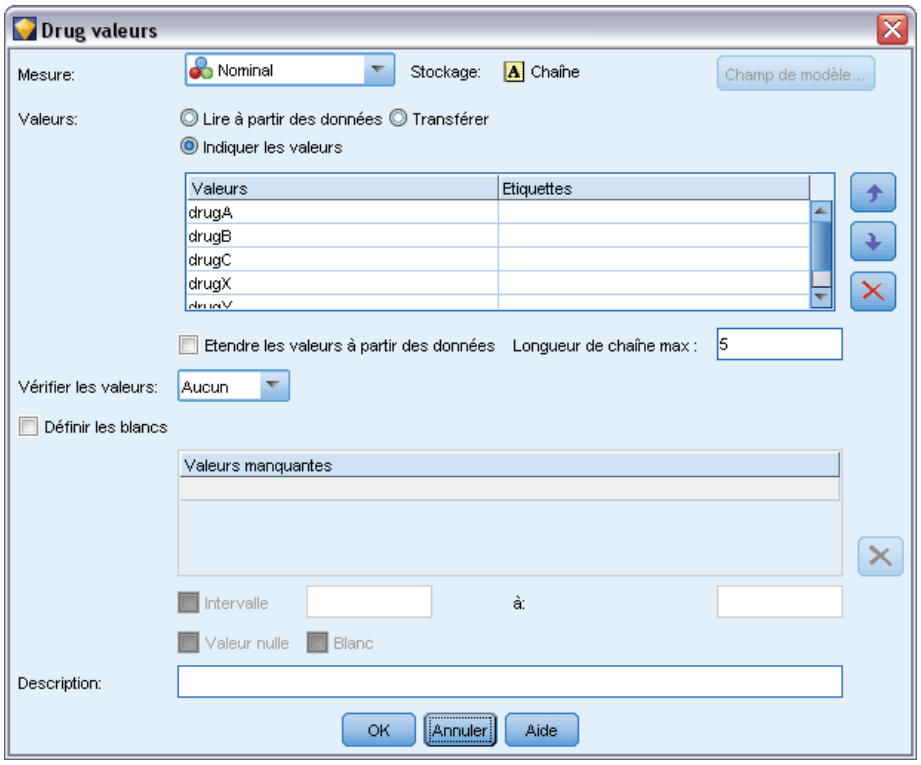
Une fois les champs du noeud Typer modifiés, vous pouvez réinitialiser les informations concernant les valeurs à l'aide des boutons suivants de la barre d'outils de la boîte de dialogue :

- A l'aide du bouton Effacer les valeurs, vous pouvez supprimer les modifications apportées aux valeurs de champ de ce noeud (valeurs non héritées) et relire les valeurs des opérations effectuées en amont. Cette option est utile pour réinitialiser les changements que vous avez apportés à certains champs en amont.
- A l'aide du bouton Effacer toutes les valeurs, vous pouvez réinitialiser les valeurs de **tous** les champs lus dans le noeud. Cette option paramètre la colonne *Valeurs* de tous les champs sur **Lire**. Cette option est utile pour réinitialiser les valeurs de tous les champs et relire les valeurs et les niveaux de mesure des opérations effectuées en amont.

Utilisation de la boîte de dialogue Valeurs

Cliquez sur la colonne *Valeurs* ou *Manquantes* de l'onglet Types affiche une liste déroulante des valeurs prédéfinies. Choisissez l'option *Spécifier* dans cette liste pour ouvrir une boîte de dialogue distincte permettant de définir les options de lecture, de spécification, d'étiquetage et de gestion des valeurs du champ sélectionné.

Figure 4-23
Paramétrage des options des valeurs de données



La majeure partie des contrôles sont communs à tous les types de données. Ces contrôles communs sont abordés ici.

Mesure. Affiche le type de niveau de mesure actuellement sélectionné. Vous pouvez modifier le paramètre pour indiquer la façon dont vous souhaitez utiliser les données. Par exemple, si un champ appelé *jour_de_la_semaine* contient des chiffres représentant des jours particuliers, vous souhaiterez peut-être modifier ceci en des données nominales afin de créer un noeud Proportion analysant chaque catégorie individuellement.

Stockage. Affiche le type de stockage, s'il est connu. Les types de stockage ne sont pas affectés par le niveau de mesure que vous choisissez. Pour modifier le type de stockage, vous pouvez utiliser l'onglet Données des noeuds source Fixe et Délimité, ou la fonction de conversion d'un noeud Remplacer.

Champ de modèle. Pour les champs générés par suite du scoring d'un nugget de modèle, les détails du champ de modèle peuvent être également affichés. Ceci comprend le nom du champ cible ainsi que la fonction du champ dans la modélisation (que ce soit une valeur prédite, une probabilité ou une propension, etc.).

Valeurs : Sélectionnez la méthode permettant de déterminer les valeurs du champ sélectionné. Ces sélections annulent celles faites précédemment à partir de la colonne *Valeurs* de la boîte de dialogue du noeud Typer. Les choix de lecture des valeurs sont les suivants :

- **Lire à partir des données.** Permet de lire les valeurs lorsque le noeud est exécuté. Cette option est identique à <Lire>.
- **Transférer.** Permet de ne pas lire les données du champ actuel. Cette option est identique à <Transférer>.
- **Indiquer les valeurs.** Ici, les options sont utilisées pour indiquer des valeurs et des étiquettes pour le champ sélectionné. Utilisée avec la vérification des valeurs, cette option vous permet de spécifier des valeurs en fonction de vos connaissances sur le champ actuel. Elle active des contrôles propres à chaque type de champ. Les options des valeurs et des étiquettes sont abordées une par une dans les rubriques suivantes. *Remarque* : Vous ne pouvez pas spécifier des valeurs ou des étiquettes pour un champ dont le niveau de mesure est *Sans type* ou<*Par défaut*>.
- **Etendre les valeurs à partir des données.** Permet d'ajouter aux données actuelles les valeurs que vous entrez ici. Par exemple, si l'intervalle du *champ_1* est compris entre 0 et 10 (0,10), et que vous saisissez l'intervalle de valeurs (8,16), l'intervalle est augmenté via l'ajout de la valeur 16, sans que la valeur minimale d'origine soit supprimée. Le nouvel intervalle est (0,16). Si vous choisissez cette option, l'option de définition automatique du type est automatiquement paramétrée sur <Lire +>.

Vérifier les valeurs. Sélectionnez une méthode de conversion forcée des valeurs pour qu'elles soient conformes aux valeurs continues, booléennes ou nominales spécifiées. Cette option correspond à la colonne *Vérifier* de la boîte de dialogue du noeud Typer ; les paramétrages effectués ici annulent ceux de la boîte de dialogue. Lorsque vous l'utilisez avec l'option Indiquer les valeurs, la vérification des valeurs vous permet de conformer les valeurs des données aux valeurs théoriques. Par exemple, si vous indiquez les valeurs 1, 0, l'option Supprimer vous permet de supprimer tous les enregistrements contenant des valeurs autres que 1 ou 0.

Définir les blancs. Sélectionnez cette option pour activer les commandes ci-dessous qui vous permettent d'indiquer des valeurs manquantes ou des blancs dans vos données.

- **Table Valeurs manquantes.** Permet de déterminer que des valeurs spécifiques, telles que 99 ou 0, sont des blancs. La valeur doit être adaptée au type de stockage du champ.
- **Intervalle :** Utilisé pour indiquer un intervalle de valeurs manquantes, comme les âges compris entre 1–17 ou supérieurs à 65. Si une valeur de limite est vide, l'intervalle est illimité ; par exemple, si une limite inférieure de 100 est indiquée sans limite supérieure, toutes les valeurs supérieures ou égales à 100 sont définies comme manquantes. Les valeurs de limite sont inclusives ; par exemple, un intervalle dont la limite inférieure est 5 et la limite supérieure est 10 comprend 5 et 10 dans la définition de l'intervalle. Vous pouvez définir un intervalle de valeurs manquantes pour tous les types de stockage, y compris de date/heure et de chaîne (dans ce cas, l'ordre de tri alphabétique est utilisé pour déterminer si une valeur fait partie de l'intervalle).
- **Nul/Blanc.** Vous pouvez également déterminer comme blancs les **valeurs système nulles** (affichées dans les données sous la forme $null$) et les **espaces blancs** (valeurs de chaîne comportant des caractères non visibles). Le noeud Typer traite également les chaînes vides comme des espaces blancs à des fins d'analyse, bien qu'elles soient stockées différemment en interne et gérées différemment dans certains cas.

Remarque : pour coder les blancs comme des valeurs non définies ou $null$, vous devez utiliser le noeud Remplacer.

Description : Utilisez cette zone de texte pour indiquer une étiquette de champ. Ces étiquettes apparaissent dans divers emplacements, tels que des graphiques, des tableaux, des résultats et des navigateurs de modèle, selon les éléments sélectionnés dans la boîte de dialogue Propriétés du flux.

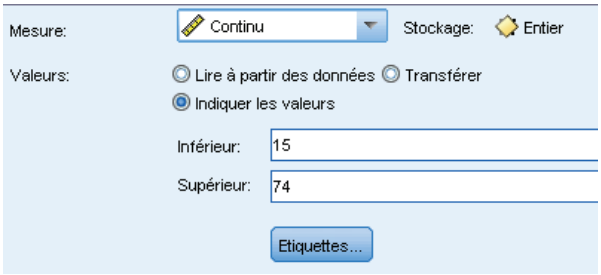
Spécification de valeurs et d'étiquettes pour des données continues

Le niveau de mesure *Continu* permet de mesurer des champs numériques. Il existe trois types de stockages pour des données continues :

- Réel
- Entier
- Date/heure

La même boîte de dialogue permet d'éditer tous les champs continus. Le type de stockage est affiché uniquement à titre de référence.

Figure 4-24
Options permettant de spécifier les valeurs continues et leurs étiquettes



Spécification de valeurs

Les commandes suivantes sont propres aux champs continus et sont utilisées pour indiquer un intervalle de valeurs :

Inférieur. Indiquez la limite inférieure de l'intervalle des valeurs.

Supérieur. Indiquez la limite supérieure de l'intervalle des valeurs.

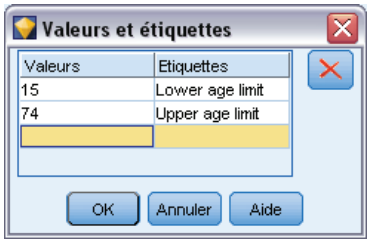
Spécification d'étiquettes

Vous pouvez spécifier des étiquettes pour toutes les valeurs d'un champ d'intervalle. Cliquez sur le bouton Etiquettes pour ouvrir une boîte de dialogue distincte permettant de spécifier les étiquettes de valeur.

Sous-boîte de dialogue Valeurs et étiquettes

Cliquez sur Etiquettes dans la boîte de dialogue Valeurs d'un champ d'intervalle pour ouvrir une nouvelle boîte de dialogue permettant de spécifier des étiquettes pour les valeurs de votre choix dans l'intervalle.

Figure 4-25
Définition d'étiquettes (facultatif) pour les valeurs d'intervalle



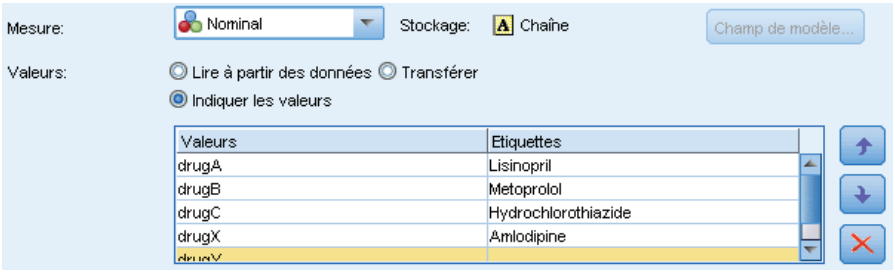
Vous pouvez utiliser les colonnes *Valeurs* et *Etiquettes* de ce tableau pour définir des paires de valeurs et d'étiquettes. Les paires actuellement définies sont indiquées ici. Pour ajouter de nouvelle paires d'étiquettes, cliquez sur une cellule vide et entrez une valeur et son étiquette. *Remarque* : L'ajout de paires valeur/valeur-étiquette à ce tableau n'engendre l'ajout d'aucune nouvelle valeur au champ. Cela crée simplement des métadonnées pour la valeur du champ.

Les étiquettes indiquées dans le noeud Typer s'affichent dans de nombreux emplacements (sous forme d'info-bulles, d'étiquettes de sortie, etc.), selon les éléments sélectionnés dans la boîte de dialogue Propriétés du flux.

Spécification des valeurs et des étiquettes pour des données nominales et ordinales

Les niveaux de mesure nominaux (ensemble) et ordinaux (ensemble ordonné) indiquent que les valeurs de données sont utilisées discrètement en tant que membres de l'ensemble. Les ensembles disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Figure 4-26
Options permettant de spécifier les valeurs nominales et les étiquettes



Les commandes suivantes sont propres aux champs nominaux et ordinaux et sont utilisées pour indiquer les valeurs et les étiquettes :

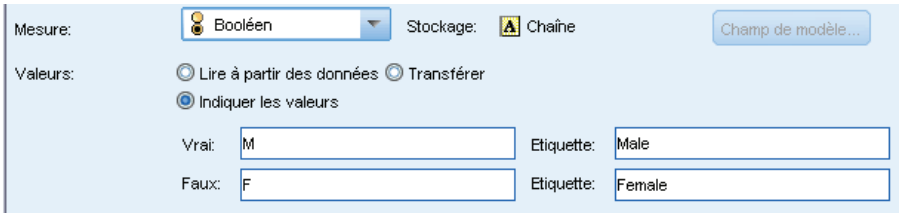
Valeurs : La colonne *Valeurs* du tableau vous permet de spécifier des valeurs, selon la connaissance que vous avez du champ actuel. Grâce à ce tableau, vous pouvez saisir des valeurs théoriques pour le champ et vérifier la conformité de l'ensemble de données par rapport à ces valeurs à l'aide de la liste déroulante Vérifier les valeurs. A l'aide des flèches et du bouton Supprimer, vous pouvez modifier, réorganiser ou supprimer les valeurs existantes.

Etiquettes : La colonne *Etiquettes* permet de spécifier des étiquettes pour chaque valeur de l'ensemble. Ces étiquettes apparaissent dans divers emplacements, tels que des graphiques, des tableaux, des résultats et des navigateurs de modèle, selon vos sélections dans la boîte de dialogue Propriétés du flux.

Spécification des valeurs d'un champ booléen

Les champs booléens servent à afficher les données possédant deux valeurs distinctes. Les booléens disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Figure 4-27
Options permettant de spécifier les valeurs des champs booléens



Vrai. Spécifiez la valeur booléenne du champ lorsque la condition est respectée.

Faux. Spécifiez la valeur booléenne du champ lorsque la condition n'est pas respectée.

Etiquettes : Spécifiez des étiquettes pour chaque valeur du champ booléen. Ces étiquettes apparaissent dans divers emplacements, tels que des graphiques, des tableaux, des résultats et des navigateurs de modèle, selon vos sélections dans la boîte de dialogue Propriétés du flux.

Définition de valeurs manquantes

La colonne Manquant de l'onglet Types indique si le traitement des valeurs manquantes a été défini pour un champ. Les paramètres possibles sont :

Activé (*). Indique que le traitement des valeurs manquantes est défini pour ce champ. Ceci peut être effectué à l'aide d'un noeud Remplacer en aval, ou à travers une spécification explicite avec l'option Spécifier (voir ci-dessous).

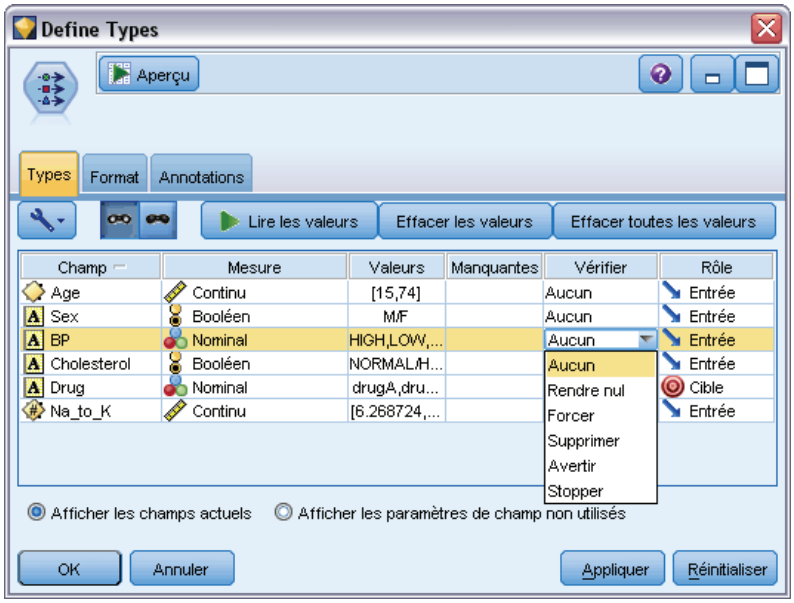
Désactivé. Le champ n'a pas de traitement des valeurs manquantes défini.

Spécifier. Choisissez cette option pour afficher une boîte de dialogue dans laquelle vous pouvez déclarer des valeurs explicites à considérer comme des valeurs manquantes pour ce champ.

Vérification des valeurs de type

Activez l'option Vérifier de chaque champ pour examiner toutes les valeurs de ce champ, et déterminer si elles sont conformes aux paramètres de type actuels ou aux valeurs spécifiées dans la boîte de dialogue Indiquer les valeurs. Cette option est pratique pour nettoyer les ensembles de données et réduire leur taille en une seule opération.

Figure 4-28
Sélection des options Vérifier du champ sélectionné



Le paramètre de la colonne *Vérifier* de la boîte de dialogue du noeud Typer détermine ce qui se produit si une valeur hors limites est découverte. Pour modifier les paramètres Vérifier d'un champ, utilisez la liste déroulante correspondante dans la colonne *Vérifier*. Pour définir les paramètres Vérifier de tous les champs, cliquez dans la colonne *Champ* et appuyez sur Ctrl+A. Utilisez ensuite la liste déroulante de n'importe quel champ de la colonne *Vérifier*.

Les paramètres Vérifier suivants sont disponibles :

Aucune. Les valeurs sont transmises sans être vérifiées. Il s'agit du paramètre par défaut.

Rendre nul. Convertit les valeurs hors limites en valeurs système nulles ($null$).

Forcer. Une recherche des valeurs situées hors de l'intervalle indiqué est effectuée sur les champs dont les niveaux de mesure sont complètement instanciés. Ces valeurs sont converties en valeurs adaptées au niveau de mesure selon les règles suivantes :

- Pour les booléens, les valeurs autres que « true » et « false » sont converties en valeurs « false ».
- Pour les ensembles (nominaux et ordinaux), les valeurs inconnues sont converties en la valeur du premier membre des valeurs de l'ensemble.
- Les nombres supérieurs à la limite supérieure d'un intervalle sont remplacés par la valeur de cette limite.
- Les nombres inférieurs à la limite inférieure d'un intervalle sont remplacés par la valeur de cette limite.
- Les valeurs nulles d'un intervalle prennent la valeur médiane de cet intervalle.

Supprimer. Lorsque des valeurs incorrectes sont trouvées, l'intégralité de l'enregistrement est supprimé.

Avertir. Le nombre d'éléments incorrects est calculé et reporté dans la boîte de dialogue des propriétés du flux une fois toutes les données lues.

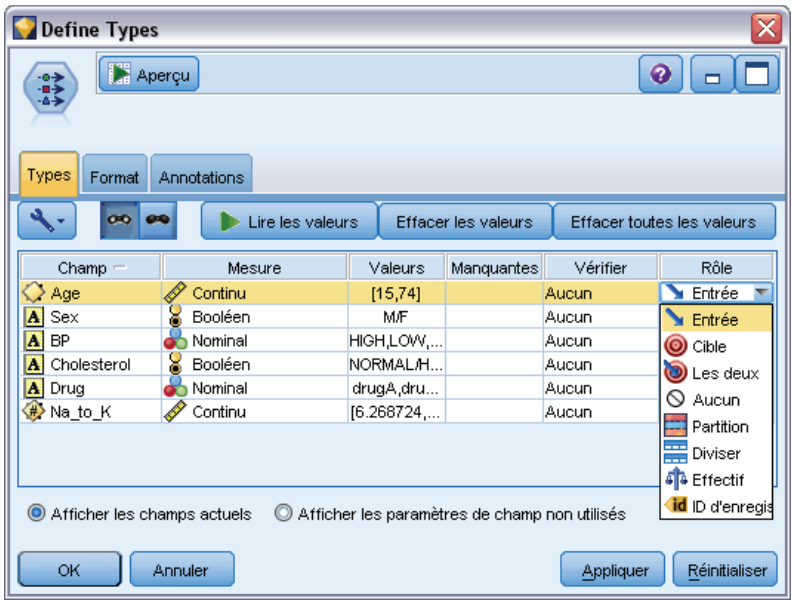
Stopper. La première valeur incorrecte rencontrée met fin à l'exécution du flux. L'erreur est reportée dans la boîte de dialogue des propriétés du flux.

Définition du rôle du champ

Le rôle d'un champ indique comment il est utilisé lors de la création de modèles, par exemple, s'il s'agit d'un champ d'entrée ou d'un champ cible (chose prévue).

Remarque : Les rôles de partition, de fréquence et d'ID d'enregistrement peuvent être chacun appliqués à un seul champ.

Figure 4-29
Paramétrage des options de rôle de champ du noeud Typer



Les rôles suivants sont disponibles :

Entrée. Le champ est utilisé comme entrée pour l'apprentissage automatique (champ variable indépendante).

Cible. Le champ est utilisé comme sortie ou cible pour l'apprentissage automatique (l'un des champs que le modèle essaie de prédire).

Les deux. Le champ est utilisé comme entrée et sortie par le noeud Apriori. Tous les autres noeuds de modélisation ignorent ce champ.

Aucune. Le champ est ignoré par l'apprentissage automatique. Les champs dont le niveau de mesure est défini sur Sans type sont automatiquement définis sur Aucun dans la colonne *Rôle*.

Partition. Indique un champ utilisé pour partitionner les données en échantillons distincts pour l'apprentissage, le test et la validation (facultatif). Le champ doit être un type d'ensemble instancié avec deux ou trois valeurs possibles (telles qu'elles sont définies dans la boîte de dialogue Valeurs de champ). La première valeur représente l'échantillon d'apprentissage, le second l'échantillon de test et le troisième (s'il existe) l'échantillon de validation. Toutes les valeurs supplémentaires sont ignorées et les champs booléens ne peuvent pas être utilisés. Pour utiliser la partition dans une analyse, vous devez l'activer dans l'onglet Options de modèle du noeud de création de modèle ou d'analyse approprié. Les enregistrements du champ de partition comportant des valeurs nulles sont exclus de l'analyse lorsque la fonction de partition est activée. Si plusieurs champs de partition ont été définis dans le flux, un champ de partition unique doit être indiqué dans l'onglet Champs de chaque noeud de modélisation applicable. Si aucun champ adapté n'existe encore dans vos données, vous pouvez en créer un via un noeud Partitionner ou Calculer. Pour plus d'informations, reportez-vous à la section Noeud Partitionner sur p. 207.

Scission. (Champs nominaux, ordinaux et booléens) Spécifie qu'un modèle doit être construit pour chaque valeur possible du champ.

Effectif. (Champs numériques uniquement) La définition de ce rôle permet d'utiliser la valeur du champ comme un facteur de pondération de fréquence pour l'enregistrement. Cette caractéristique est uniquement prise en charge par les modèles C&R Tree, CHAID, QUEST et linéaires ; tous les autres noeuds ignorent ce rôle. La pondération de fréquence est activée au moyen de l'option Utiliser la pondération de fréquence de l'onglet Champs de ces noeuds de modélisation qui prennent en charge cette caractéristique.

ID d'enregistrement. Le champ est utilisé comme identificateur d'enregistrement unique. Cette fonction est ignorée par la plupart des noeuds ; cependant, elle est prise en charge par les modèles linéaires et requise pour les noeuds d'exploration de la base de données Netezza IBM.

Copie d'attributs de type

Vous pouvez facilement copier les attributs d'un type, tels que les valeurs, les options de vérification et les valeurs manquantes, d'un champ à l'autre :

- Cliquez avec le bouton droit de la souris sur le champ dont vous souhaitez copier les attributs.
- Dans le menu contextuel, choisissez Copier.
- Cliquez avec le bouton droit de la souris sur les champs dont vous souhaitez changer les attributs.
- Dans le menu contextuel, sélectionnez Collage spécial. *Remarque* : Vous pouvez sélectionner plusieurs champs en appuyant sur Ctrl tout en cliquant sur les champs ou en choisissant l'option Sélectionner les champs dans le menu contextuel.

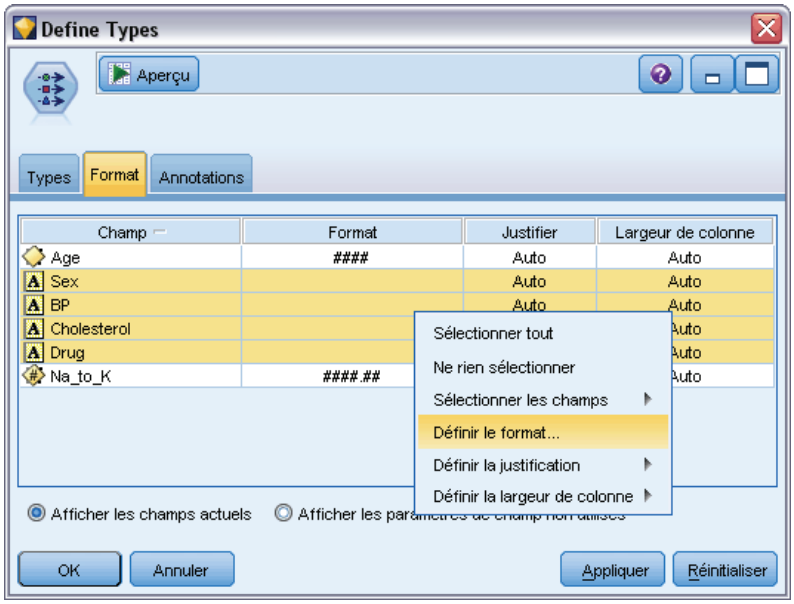
  Une nouvelle boîte de dialogue apparaît ; elle vous permet de sélectionner les attributs spécifiques que vous souhaitez coller. Si vous collez des attributs dans plusieurs champs, les options sélectionnées ici s'appliquent à tous les champs cible.

  Coller les attributs suivants. Sélectionnez parmi les options ci-dessous les attributs à coller d'un champ à l'autre.

  - **Type.** Sélectionnez cette option pour coller le niveau de mesure.
  - **Valeurs.** Sélectionnez cette option pour coller les valeurs de champ.
  - **Manquantes.** Sélectionnez cette option pour coller les paramètres des valeurs manquantes.
  - **Vérifier.** Sélectionnez cette option pour coller les options de vérification des valeurs.
  - **Rôle.** Sélectionnez cette option pour coller le rôle d'un champ.

Onglet Paramètres du champ

Figure 4-30
Noeud Typer, onglet Format



L'onglet Format des noeuds Table et Typer répertorie les champs actuels et non utilisés, ainsi que les options de formatage de chaque champ. Les colonnes du tableau de formatage des champs sont décrites ci-dessous :

Champ. Indique le nom du champ sélectionné.

Format : Double-cliquez sur une cellule de cette colonne pour spécifier le formatage de chacun des champs à l'aide de la boîte de dialogue appelée. Pour plus d'informations, reportez-vous à la section Paramétrage des options de formatage des champs sur p. 154. Le formatage indiqué ici remplace celui indiqué dans les propriétés générales du flux.

Remarque : Les noeuds Export Statistics et Sortie Statistics exportent des fichiers *.sav* comportant dans leurs métadonnées un formatage par champ. Si l'un des formats par champ indiqués n'est pas pris en charge par le format de fichier IBM® SPSS® Statistics*.sav* , le noeud utilise le format SPSS Statistics par défaut.

Justifier. Utilisez cette colonne pour indiquer le mode de justification des valeurs dans les colonnes du tableau. Le paramètre par défaut est Auto : il justifie les valeurs symboliques vers la gauche et les valeurs numériques vers la droite. Vous pouvez remplacer ce paramètre par défaut en sélectionnant Gauche, Droite ou Au milieu.

Largeur de colonne. Par défaut, les largeurs de colonne sont automatiquement calculées sur la base des valeurs du champ. Pour remplacer le calcul automatique de la largeur, cliquez sur une cellule du tableau et utilisez la liste déroulante pour sélectionner une nouvelle largeur. Pour entrer une largeur personnalisée non répertoriée ici, ouvrez la sous-boîte de dialogue Format de champ en double-cliquant sur une cellule de la colonne *Champ* ou *Format* dans le tableau. Vous pouvez également cliquer avec le bouton droit de la souris sur une cellule et sélectionner Définir le format.

Afficher les champs actuels. Par défaut, la boîte de dialogue contient la liste des champs actifs. Pour afficher la liste des champs inutilisés, sélectionnez Afficher les paramètres de champ non utilisés.

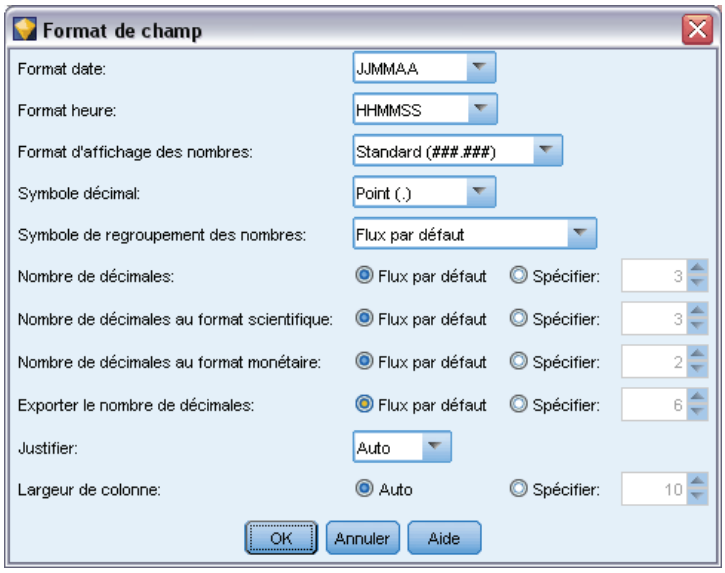
Menu Contexte. Le menu Contexte de cet onglet contient des options de sélection et de mise à jour des paramètres.

- **Sélectionner tout.** Sélectionne l'ensemble des champs.
- **Ne rien sélectionner.** Supprime la sélection.
- **Sélectionner les champs.** Sélectionne des champs sur la base de leur type ou de leur caractéristique de stockage. Les options disponibles sont les suivantes : Sélectionner catégoriel, Sélectionner continu (données numériques), Sélectionner Sans Type, Sélectionner Chaînes, Sélectionner Nombres ou Sélectionner Date/Heure. Pour plus d'informations, reportez-vous à la section Niveaux de mesure sur p. 138.
- **Définir le format.** Ouvre une sous-boîte de dialogue permettant de spécifier les options de date, d'heure et décimales par champ.
- **Définir la justification.** Définit le mode de justification des champs sélectionnés. Les options sont les suivantes : Auto, Au milieu, Gauche ou Droite.
- **Définir la largeur de colonne.** Définit la largeur des champs sélectionnés. Indiquez Automatique pour lire la largeur dans les données. Vous pouvez également définir la largeur du champ sur 5, 10, 20, 30, 50, 100 ou 200.

Paramétrage des options de formatage des champs

Le formatage des champs est spécifié dans une sous-boîte de dialogue disponible à partir de l'onglet Format des noeuds Typer et Table. Si vous avez sélectionné plusieurs champs avant d'ouvrir cette boîte de dialogue, les paramètres du premier champ de la sélection sont utilisés pour tous les champs. Cliquez sur OK une fois les spécifications définies ici pour appliquer ces paramètres à tous les champs sélectionnés dans l'onglet Format.

Figure 4-31
Définition des options de formatage d'un ou de plusieurs champs



Les options suivantes sont disponibles par champ. Vous pouvez également spécifier la plupart de ces paramètres dans la boîte de dialogue Propriétés du flux. Tous les paramètres définis au niveau du champ remplacent les paramètres par défaut indiqués pour le flux.

Format date. Sélectionnez le format de date à utiliser pour les champs de stockage de date ou lorsque les chaînes sont interprétées comme des dates par les fonctions de date CLEM.

Format heure. Sélectionnez le format d'heure à utiliser pour les champs de stockage d'heure ou lorsque les chaînes sont interprétées comme des heures par les fonctions d'heure CLEM.

Format d'affichage des nombres. Vous pouvez sélectionner les formats d'affichage standard (####.###), scientifique (#.###E+##) ou monétaire (### ## €).

Symbole décimal. Sélectionnez la virgule (,) ou le point (.) comme séparateur décimal.

Symbole de regroupement. Pour les formats d'affichage des nombres, sélectionnez le symbole permettant de regrouper des valeurs (par exemple, l'espace dans *3 000,00*). Vous avez le choix entre les options suivantes : aucun, point, virgule, espace et paramètres régionaux définis (auquel cas la valeur par défaut des paramètres régionaux actuels est utilisée).

Nombre de décimales (au format standard, scientifique ou monétaire, ou à exporter). Pour les formats d'affichage des nombres, indique le nombre de décimales à utiliser pour l'affichage, l'impression ou l'exportation des nombres réels. Cette option apparaît séparément pour chaque format d'affichage. Le format d'exportation s'applique uniquement aux champs dont le stockage est réel.

Justifier. Indique le mode de justification des valeurs dans la colonne. Le paramètre par défaut est Auto : il justifie les valeurs symboliques vers la gauche et les valeurs numériques vers la droite. Vous pouvez remplacer ce paramètre par défaut en sélectionnant Gauche, Droite ou Au milieu.

Largeur des colonnes. Par défaut, les largeurs de colonne sont automatiquement calculées sur la base des valeurs du champ. Vous pouvez spécifier des largeurs personnalisées par intervalles de cinq à l'aide des flèches situées à droite de la zone de liste.

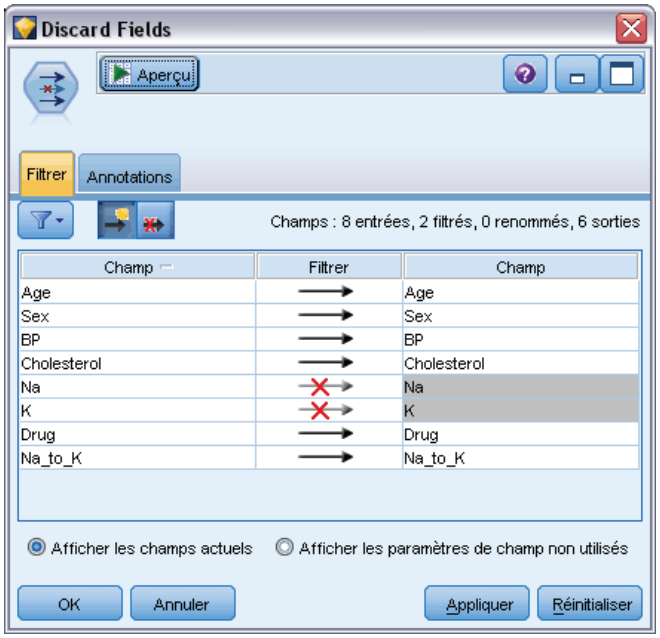
Filtrage ou modification du nom des champs

Vous pouvez renommer ou exclure des champs à tout stade d'un flux. Par exemple, en tant que chercheur en médecine, vous n'êtes peut-être pas intéressé par le niveau de potassium (données de niveau champ) des patients (données de niveau enregistrement) ; vous pouvez donc filtrer le champ *K* correspondant. Vous pouvez réaliser ceci à l'aide d'un noeud Filtrer distinct ou d'un onglet Filtrer sur un noeud source ou de sortie. Cette fonctionnalité est identique quel que soit le noeud à partir duquel vous y accéder.

- Depuis les noeuds source, tels que les noeuds Délimité, Fixe, Statistics et XML, vous pouvez renommer ou filtrer les champs au moment où les données sont lues dans IBM® SPSS® Modeler.
- Le noeud Filtrer permet de renommer ou de filtrer les champs en tout point du flux.
- A partir des noeuds Export Statistics, Transformation Statistics, Modèle Statistics et Sortie Statistics, vous pouvez filtrer ou renommer les champs pour respecter les conventions de dénomination IBM® SPSS® Statistics. Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics dans le chapitre 8 sur p. 513.

- Vous pouvez utiliser l'onglet Filtrer dans l'un des noeuds susmentionnés pour définir ou modifier des ensembles de réponses multiples. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples sur p. 159.
- Finalement, vous pouvez utiliser un noeud Filtrer pour mapper les champs entre un noeud source et un autre.

Figure 4-32
Paramétrage des options du noeud Filtrer



Paramétrage des options de filtrage

Le tableau utilisé dans l'onglet Filtrer affiche le nom des champs dès qu'ils entrent ou sortent du noeud. Vous pouvez utiliser les options de ce tableau pour renommer ou filtrer les champs qui sont en double ou inutiles pour les opérations en aval.

- **Champ.** Affiche les champs d'entrée des sources de données actuellement connectées.
- **Filtrer.** Affiche l'état de filtrage de tous les champs d'entrée. Les champs filtrés comportent un X rouge dans cette colonne, ce qui indique qu'ils ne seront pas transférés en aval. Cliquez dans la colonne *Filtrer* d'un champ sélectionné pour activer ou désactiver le filtrage. Vous pouvez également sélectionner des options pour plusieurs champs en même temps, en utilisant la méthode de sélection Maj+clic.
- **Champ.** Affiche les champs lorsqu'ils quittent le noeud Filtrer. Les noms en double sont affichés en rouge. Pour éditer les noms des champs, cliquez dans la colonne et saisissez un nouveau nom. Vous pouvez également supprimer les champs en cliquant dans la colonne *Filtrer* pour désactiver les champs en double.

Vous pouvez trier toutes les colonnes du tableau en cliquant sur l'en-tête de la colonne.

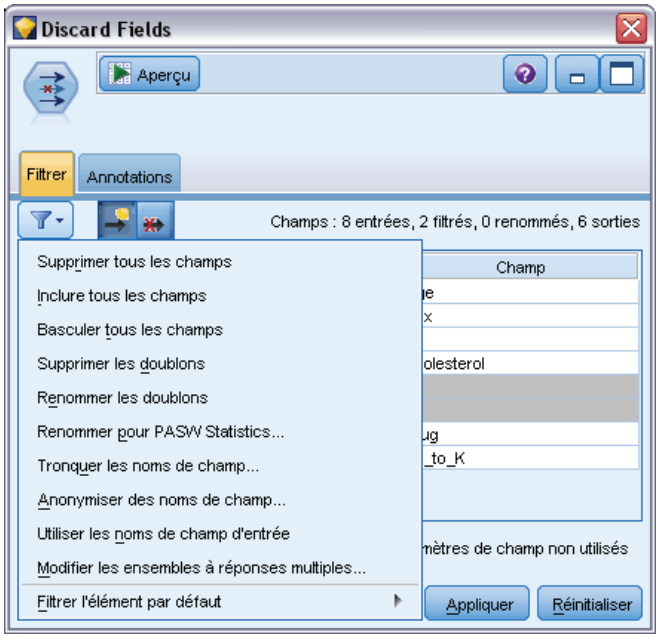
Afficher les champs actuels. Sélectionnez cette option pour afficher les champs des ensembles de données connectés au noeud Filtrer. Cette méthode standard d'utilisation des noeuds Filtrer est sélectionnée par défaut.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher les champs des ensembles de données qui étaient auparavant connectés au noeud Filtrer. Cette option est utile lorsque vous copiez des noeuds Filtrer d'un flux à un autre, ou lorsque vous enregistrez ou rechargez des noeuds Filtrer.

Menu du bouton Filtrer

Cliquez sur le bouton Filtrer en haut à gauche de la boîte de dialogue pour accéder à un menu qui propose un certain nombre de raccourcis et d'autres options.

Figure 4-33
Options du menu Filtrer



Vous pouvez :

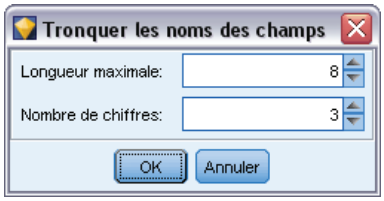
- Supprimer tous les champs.
- Inclure tous les champs.
- Basculer tous les champs.
- Supprimer les doublons. *Remarque* : la sélection de cette option entraîne la suppression de toutes les occurrences du nom en double, y compris la première.
- Renommer les champs et les ensembles de réponses multiples pour être en conformité avec d'autres applications . Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics dans le chapitre 8 sur p. 513.
- Tronquer les noms de champ.
- Anonymiser les noms de champs et d'ensembles de réponses multiples.

- Utiliser les noms de champ d'entrée.
- Modifier les ensembles de réponses multiples. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples sur p. 159.
- Définir l'état de filtrage par défaut.

Vous pouvez également utiliser les boutons bascule représentant une flèche, en haut de la boîte de dialogue, pour indiquer si vous souhaitez, par défaut, inclure ou ignorer les champs. Ces boutons sont particulièrement utiles lorsque vous travaillez avec des ensembles de données volumineux dans lesquels seuls quelques champs doivent être inclus en aval. Par exemple, vous pouvez sélectionner uniquement les champs que vous souhaitez conserver et indiquer que tous les autres doivent être ignorés (au lieu de sélectionner chaque champ à ignorer).

Troncation des noms de champ

Figure 4-34
Boîte de dialogue Tronquer les noms de champ



A partir du menu du bouton Filtrer (en haut à gauche de l'onglet Filtrer), vous pouvez choisir de tronquer des noms de champs.

Longueur maximale. Limitez la longueur des noms de champ en indiquant un nombre de caractères.

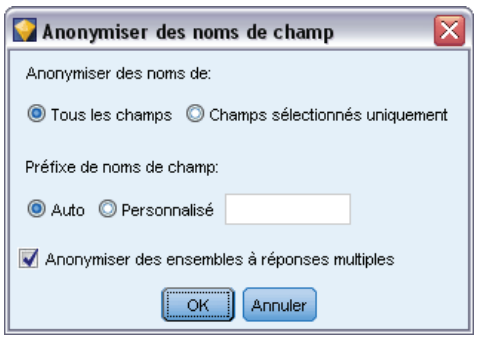
Nombre de chiffres. Si, une fois raccourci, le nom d'un champ n'est plus unique, il est de nouveau raccourci et assorti d'un chiffre permettant de le différencier des autres. Vous pouvez indiquer le nombre de chiffres à utiliser. Utilisez les flèches pour rectifier ce nombre.

Par exemple, le tableau ci-dessous indique comment les noms de champ d'un ensemble de données médicales sont raccourcis en fonction des paramètres par défaut (Longueur maximale = 8 et Nombre de chiffres = 2).

Noms de champ	Noms de champ raccourcis
Entrée patient 1	Patien01
Entrée patient 2	Patien02
Rythme cardiaque	RythmeCa
TA	TA

Anonymisation des noms de champ

Figure 4-35
Boîte de dialogue Anonymiser des noms de champ



Vous pouvez anonymiser des noms de champs à partir d'un noeud quelconque qui comporte un onglet Filtrer en cliquant sur le menu du bouton Filtrer et en choisissant Anonymiser des noms de champ. Les noms de champ anonymisés sont formés d'un préfixe de chaîne suivi d'une valeur numérique unique.

Anonymiser des noms de. Choisissez Champs sélectionnés uniquement pour n'anonymiser que les noms des champs déjà sélectionnés dans l'onglet Filtrer. La valeur par défaut est Tous les champs, qui anonymise tous les noms de champ.

Préfixe de noms de champ. Le préfixe par défaut des noms de champ anonymisés est anon_. Si vous souhaitez le modifier, choisissez Personnalisé et saisissez votre propre préfixe.

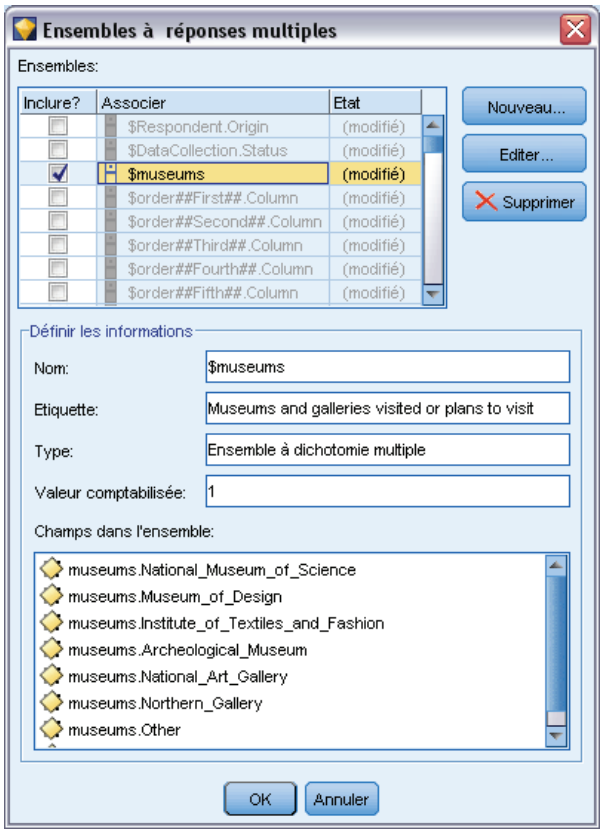
Anonymiser des ensembles de réponses multiples. Anonymise le nom des ensembles de réponses multiples de la même manière que les champs. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples sur p. 159.

Pour restaurer les noms de champ d'origine, choisissez Utiliser les noms de champ d'entrée dans le menu Filtrer.

Modification des ensembles de réponses multiples

Vous pouvez ajouter ou modifier des ensembles de réponses multiples à partir d'un noeud quelconque qui comporte un onglet Filtrer en cliquant sur le menu du bouton Filtrer et en choisissant Editer des ensembles de réponses multiples.

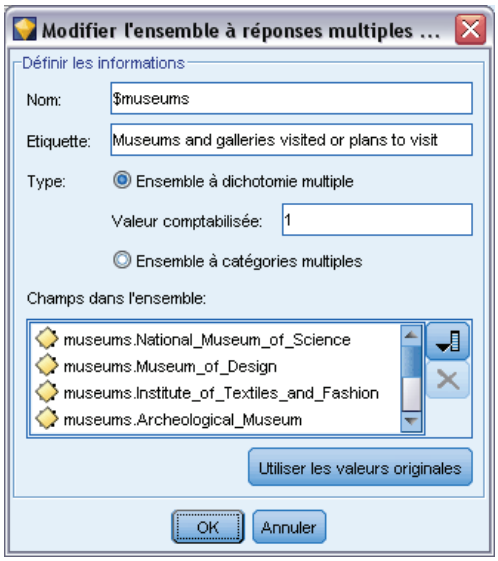
Figure 4-36
Boîte de dialogue Ensembles à réponses multiples



Les ensembles de réponses multiples permettent de consigner des données pouvant comporter plusieurs valeurs pour chaque cas—par exemple, lorsque les personnes sondées sont interrogées sur les musées qu'ils ont visités ou les magazines qu'ils ont lus. Il est possible d'importer des ensembles de réponses multiples dans IBM® SPSS® Modeler à l'aide d'un noeud source Data Collection ou Statistics. En outre, vous pouvez les définir dans SPSS Modeler à l'aide d'un noeud Filtrer.

▶ Cliquez sur Nouveau pour créer un nouvel ensemble de réponses multiples ou sur Modifier pour les modifier.

Figure 4-37
Modification d'un ensemble de réponses multiples



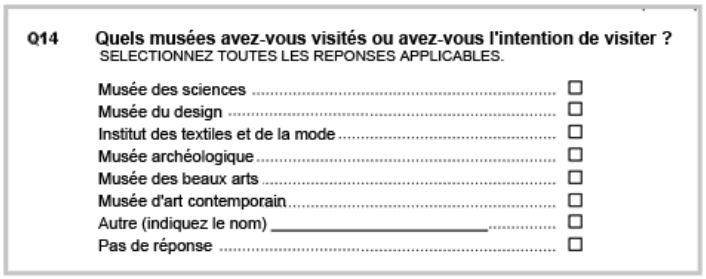
Nom et étiquette. Indique le nom et la description de l'ensemble.

Type. Les questions à réponses multiples peuvent être traitées de l'une des deux manières suivantes :

- **Ensemble de dichotomies multiples** Un champ booléen distinct est créé pour chaque réponse possible. Par conséquent, pour 10 magazines, il y a 10 champs booléens, dont chacun comprend des valeurs telles que 0 ou 1 pour *vrai* ou *faux*. La valeur calculée permet de préciser celle qui est considérée comme étant 'vrai'. Cette méthode est utile pour permettre aux personnes interrogées de choisir toutes les options applicables.
- **Ensemble de catégories multiples.** Un champ nominal est créé pour chaque réponse jusqu'au nombre maximum de réponses d'une personne interrogée donnée. Chaque champ nominal comprend des valeurs qui réprésentent les réponses possibles, telles que *1* pour *Temps*, *2* pour *Newsweek* et *3* pour *PC Week*. Cette méthode est très utile lorsque vous souhaitez limiter le nombre de réponses—par exemple, lorsque les personnes sondées sont interrogées sur les trois magazines les plus souvent lus.

Champs dans l'ensemble. Les icônes de droite permettent d'ajouter ou de supprimer des champs.

Figure 4-38
Question à réponses multiples



Commentaires

- Tous les champs inclus dans un ensemble de réponses multiples doivent disposer du même stockage.
- Les ensembles sont distincts des champs qu'ils comportent. Par exemple, la suppression d'un ensemble n'entraîne pas celle des champs inclus dans celui-ci, simplement des liens entre ces champs. L'ensemble reste visible en amont du point de suppression mais pas en aval.
- Si des champs sont renommés à l'aide d'un noeud Filtrer (directement sur l'onglet ou en choisissant les options Renommer pour IBM® SPSS® Statistics, Tronquer ou Anonymiser sur le menu Filtrer), toute référence à ces champs utilisée dans plusieurs ensembles de réponses est également mise à jour. Toutefois, tout champ dans un ensemble de réponses multiples qui est supprimé par le noeud Filtrer ne l'est pas de l'ensemble de réponses multiples. De tels champs, bien que masqués dans le flux, sont encore référencés par l'ensemble de réponses multiples. Ceci peut être pris en compte lors de l'exportation, par exemple.

Noeud Ensemble

Le noeud Ensemble combine deux ou plusieurs nuggets de modèles pour obtenir des prévisions plus précises que celles acquises à partir des modèles individuels. En combinant les prévisions à partir de plusieurs modèles, il est possible d'éviter les limitations dans les modèles individuels. Ce qui entraîne une plus grande précision globale. Les modèles combinés de cette manière fonctionne généralement aussi bien, sinon mieux, que les modèles individuels.

Cette combinaison de noeuds se produit automatiquement dans les noeuds de modélisation automatisée : Classificateur automatique, Numérisation automatique et Classification automatique.

Après avoir utilisé un noeud Ensemble, vous pouvez utiliser un noeud Analyse ou Evaluation pour comparer la précision des résultats combinés avec chacun des modèles d'entrée. Pour ce faire, assurez-vous que l'option Filtrer les champs générés par des modèles combinés n'est pas sélectionnée dans l'onglet Paramètres du noeud Ensemble.

Champs de sortie

Chaque noeud Ensemble génère un champ contenant les scores combinés. Le nom est basé sur le champ cible spécifié et comporte le préfixe *$XF_*, *$XS_* ou *$XR_*, selon le niveau de mesure de champ (Booléen, nominal (ensemble) ou continu (intervalle), respectivement). Par exemple, si la cible est un champ booléen nommé *réponse*, le champ de sortie est *$XF_response*.

Champs de confiance ou de propension. Pour les champs booléens et nominaux, d'autres champs de confiance ou de propension sont créés selon la méthode d'ensemble, comme illustré dans le tableau suivant :

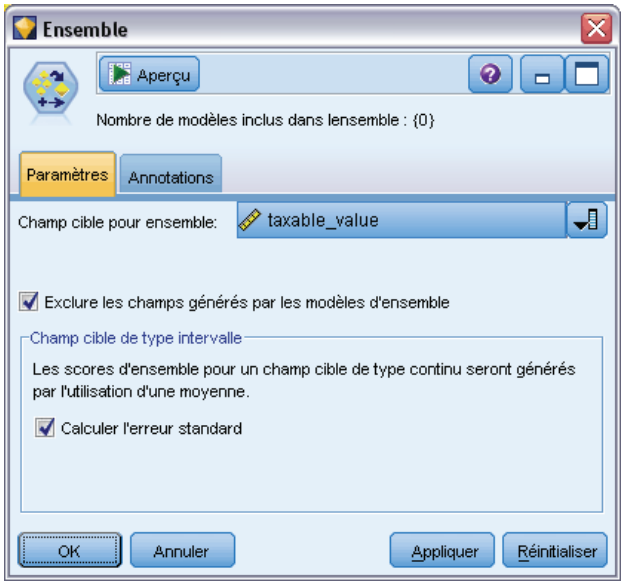
Méthode d'ensemble	Nom de champ
Voting Vote pondéré par la confiance Vote pondéré par la propension brute Vote pondéré - propension ajustée La confiance la plus élevée l'emporte	Champ *$XFC_<>*
Propension brute moyenne	Champ *$XFRP_<>*
Propension brute moyenne ajustée	Champ *$XFAP_<>*

Paramètres du noeud Ensemble

Champ cible pour ensemble. Sélectionnez un champ unique qui est utilisé comme cible pour deux ou plusieurs modèles en amont. Les modèles en amont peuvent utiliser des champs cible Booléen, Nominal ou Continu. Toutefois, au moins deux des modèles doivent partager la même cible afin de combiner les scores.

Filtrer des champs générés par des modèles combinés. Supprime de la sortie tous les champs supplémentaires générés par les modèles individuels qui sont intégrés au noeud Ensemble. Cochez cette case si seul le score combiné de tous les modèles d'entrée vous intéresse. Assurez-vous que cette option est désélectionnée si, par exemple, vous voulez utiliser un noeud Analyse ou Evaluation pour comparer la précision du score combiné avec chacun des modèles d'entrée individuels.

Figure 4-39
Noeud Ensemble avec un champ continu sélectionné comme cible



Les paramètres disponibles dépendent du niveau de mesure de champ sélectionné comme cible.

Cibles continues

Pour une cible continue, la moyenne des scores est effectuée. Il s'agit de la seule méthode disponible pour la combinaison des scores.

Lorsque vous effectuez la moyenne des scores ou des évaluations, le noeud Ensemble utilise un calcul d'erreur standard pour déterminer la différence entre les valeurs mesurées ou estimées et les valeurs réelles et pour montrer la correspondance proche de ces évaluations. Le calcul d'erreur standard est généré par défaut pour les nouveaux modèles ; vous pouvez néanmoins décocher la case des modèles existants, par exemple s'ils doivent être régénérés.

Cibles catégorielles

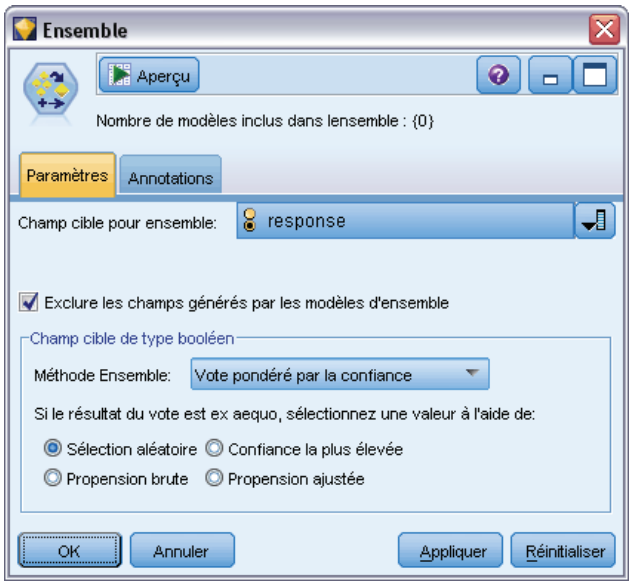
Pour ce type de cible, un certain nombre de méthodes sont prises en charge, dont le **vote**, qui fonctionne en comptant le nombre de fois où chaque valeur prédite possible est choisie et en sélectionnant la valeur avec le total le plus élevé. Par exemple, si trois modèles sur cinq prédisent *oui* et que les deux autres prédisent *non*, le*oui* remporte par un vote de 3 contre 2. Les votes peuvent être aussi **pondérés** selon la valeur de confiance ou de propension pour chaque prédiction. La somme des pondérations est ensuite effectuée, et la valeur avec le total le plus élevé est à nouveau sélectionné. La confiance pour la prédiction finale est la somme des pondérations pour la valeur gagnante divisée par le nombre de modèles inclus dans l'ensemble.

Figure 4-40
Noeud Ensemble avec un champ nominal sélectionné comme cible

Tous les champs catégoriels. Pour les champs Booléen et Nominal, les méthodes suivantes sont prises en charge :

- Voting
- Vote pondéré par la confiance
- La confiance la plus élevée l'emporte

Figure 4-41
Noeud Ensemble avec un champ Booléen sélectionnée comme cible



Champs booléens uniquement. Pour les champs booléens uniquement, un certain nombre de méthodes basées sur la propension sont également disponibles :

- Vote pondéré - propension brute
- Vote pondéré - propension ajustée
- Propension brute moyenne
- Propension ajustée moyenne

Ex æquo du vote. Pour les méthodes de vote, vous pouvez indiquer le mode de résolution des ex æquo.

- **Sélection aléatoire.** Une des valeurs ex æquo est choisie au hasard.
- **Confiance la plus grande.** La valeur ex æquo prédite avec la plus grande confiance gagne. Remarque : ce n'est pas forcément la même que la plus grande confiance de toutes les valeurs prédites.
- **Propension brute ou ajustée (champs booléens uniquement).** La valeur ex æquo prédite avec la plus grande propension absolue, où la propension absolue est calculée avec :

```
abs(0.5 - propensity) *
2
```

Ou, dans le cas d'une propension ajustée :

```
abs(0.5 - adjusted propensity) * 2
```

Noeud Calculer

L'une des fonctionnalités les plus performantes de IBM® SPSS® Modeler est la capacité à modifier les valeurs de données et à calculer de nouveaux champs à partir de données existantes. Au cours des projets de Data mining très longs, il est courant d'effectuer plusieurs calculs tels que l'extraction d'un ID client d'une chaîne des données du log Web ou la création d'une valeur de durée de vie de client basée sur des données démographiques et de transaction. Toutes ces transformations peuvent être effectuées à l'aide des divers noeuds d'opérations sur les champs.

Plusieurs noeuds permettent de calculer de nouveaux champs :

Le noeud Calculer modifie les valeurs de données ou crée des nouveaux champs à partir d'un ou de plusieurs champs existants. Il crée des champs de type formule, booléen, ensemble, nominal, statistiques, comptage et conditionnel. Pour plus d'informations, reportez-vous à la section Noeud Calculer sur p. 166.

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser. Pour plus d'informations, reportez-vous à la section Noeud Recoder sur p. 186.

Le noeud Discrétiser crée automatiquement des champs nominaux (ensemble) sur la base des valeurs d'un ou de plusieurs champs continus (intervalle numérique) existants. Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus comme écarts par rapport à la moyenne. Une fois les intervalles du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure. Pour plus d'informations, reportez-vous à la section Noeud Discrétiser sur p. 191.

Le noeud Binariser calcule plusieurs champs booléens en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux. Pour plus d'informations, reportez-vous à la section Noeud Binariser sur p. 210.

Le noeud Restructurer convertit un champ nominal ou un champ booléen en un groupe de champs renseignés à partir des valeurs d'un autre champ. Par exemple, si l'on considère un champ nommé *type de paiement*, qui comporte les valeurs *crédit*, *liquide* et *débit*, trois champs sont alors créés (*crédit*, *liquide*, *débit*), chacun contenant la valeur du paiement réel effectué. Pour plus d'informations, reportez-vous à la section Noeud Restructurer sur p. 211.

Le noeud Historiser crée des champs contenant des données provenant de champs d'enregistrements antérieurs. Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Avant d'utiliser un noeud Historiser, vous pouvez trier les données à l'aide d'un noeud Trier. Pour plus d'informations, reportez-vous à la section Noeud Historiser sur p. 240.

Utilisation du noeud Calculer

A l'aide du noeud Calculer, vous pouvez créer six types de nouveau champ à partir d'un ou de plusieurs champs existants :

- **Formule.** Le nouveau champ est le résultat d'une expression CLEM arbitraire.

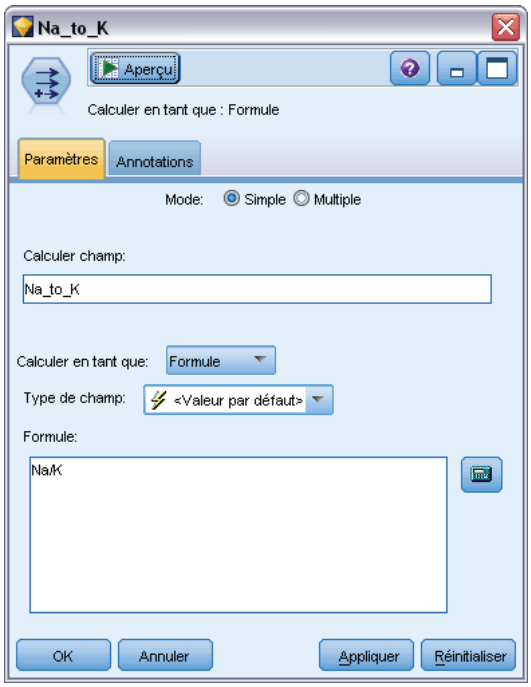
- **Booléen.** Le nouveau champ est un champ booléen, représentant une condition spécifique.
- **Nominal :** Le nouveau champ est un champ nominal. Autrement dit, ses membres constituent un groupe de valeurs spécifiées.
- **Etat.** Le nouveau champ est l'un de deux états. Le passage d'un état à l'autre est déclenché par une condition donnée.
- **Effectifs.** Le nouveau champ est basé sur le nombre de fois qu'une condition est vraie (true).
- **Conditionnel.** Le nouveau champ est la valeur de l'une des deux expressions, selon la valeur d'une condition.

Chacun de ces noeuds contient un ensemble d'options particulières dans la boîte de dialogue du nœud Calculer. Ces options sont traitées dans des rubriques ultérieures.

Paramétrage des options de base du nœud Calculer

Vous trouverez dans la partie supérieure de la boîte de dialogue des noeuds Calculer plusieurs options permettant de sélectionner le type de noeud Calculer dont vous avez besoin.

Figure 4-42
Boîte de dialogue du noeud Calculer



Mode. Sélectionnez le mode Simple ou Multiple, selon que vous souhaitez ou non calculer plusieurs champs. Lorsque le mode Multiple est sélectionné, la boîte de dialogue change : de nouvelles options, adaptées à plusieurs champs de calcul, apparaissent.

Calculer champ. Pour les noeuds Calculer simples, spécifiez le nom du champ que vous voulez calculer et ajouter à chaque enregistrement. Le nom par défaut est Derive *N*, où *N* correspond au nombre de noeuds Calculer créés jusqu'ici dans la session actuelle.

Calculer en tant que. Dans la liste déroulante, sélectionnez le type de nœud Calculer, tel que Formule ou Nominal. Pour chaque type, un nouveau champ est créé en fonction des conditions que vous spécifiez dans la boîte de dialogue correspondante.

La sélection d'une option dans la liste déroulante a pour effet d'ajouter un nouvel ensemble de contrôles dans la boîte de dialogue principale selon les propriétés de chaque type de nœud Calculer.

Type de champ. Sélectionnez le niveau de mesure du nouveau nœud calculé (par exemple, Continu, Catégoriel ou Booléen). Cette option est commune à toutes les formes de noeuds Calculer.

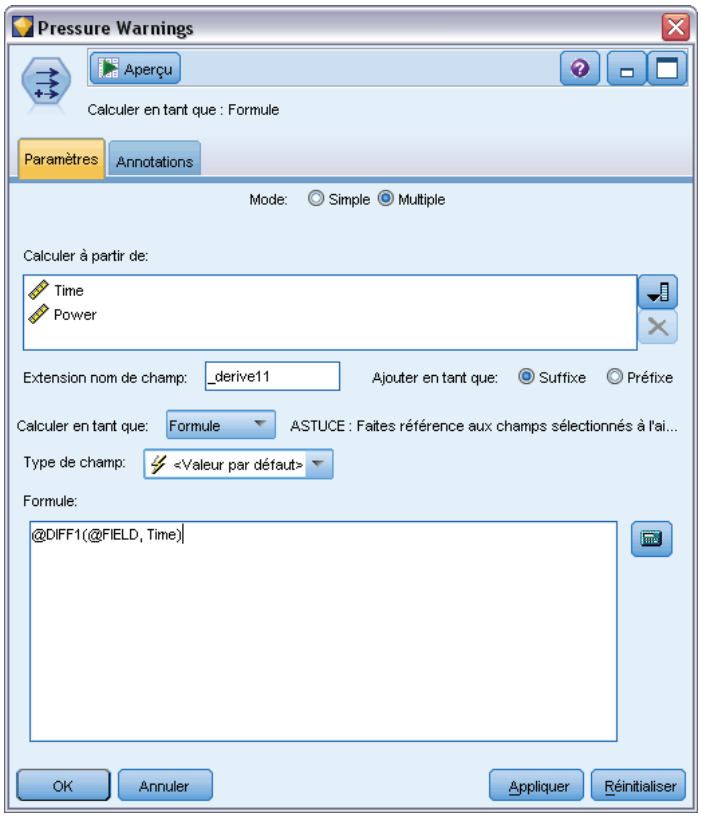
Remarque : le calcul des nouveaux champs demande souvent l'utilisation de fonctions ou expressions mathématiques particulières. Pour créer ces expressions, vous pouvez utiliser le Générateur de formules, disponible à partir de la boîte de dialogue de tous les types de noeud Calculer. Il permet de vérifier les règles et fournit la liste complète des expressions CLEM.

Calcul à partir de plusieurs champs

Le fait de passer en mode Multiple dans le nœud Calculer vous permet d'effectuer un calcul à partir de plusieurs champs reposant sur la même condition et appartenant à un même nœud. Cette fonction vous permet de gagner du temps lorsque vous voulez appliquer des transformations identiques à plusieurs champs de votre ensemble de données. Par exemple, si vous voulez créer un modèle de régression permettant de calculer le salaire actuel en fonction du salaire de départ et de l'expérience professionnelle, il peut s'avérer utile d'appliquer une transformation logarithmique à ces trois variables. Plutôt que d'ajouter un nouveau nœud Calculer pour chaque transformation, vous pouvez appliquer simultanément la même fonction à tous les champs. Il vous suffit de sélectionner tous les champs à partir desquels vous voulez calculer un nouveau champ, puis de saisir la formule de calcul en utilisant la fonction @FIELD dans les parenthèses des champs.

Remarque : la fonction @FIELD est très pratique pour effectuer un calcul à partir de plusieurs champs en même temps. Elle vous permet de faire référence au contenu des champs actuels, sans spécifier leur nom exact. Par exemple, l'expression CLEM utilisée pour appliquer une transformation logarithmique à plusieurs champs est log(@FIELD).

Figure 4-43
Calcul à partir de plusieurs champs



Les options suivantes apparaissent dans la boîte de dialogue lorsque vous sélectionnez le mode Multiple :

Calculer à partir de. Utilisez le sélecteur de champs pour sélectionner les champs à partir desquels calculer de nouveaux champs. Un champ de sortie est généré pour chaque champ sélectionné. *Remarque* : Les champs sélectionnés n'ont pas besoin d'avoir le même type de stockage. Néanmoins, l'opération de calcul échouera si la condition n'est pas valide pour *tous* les champs.

Extension nom de champ. Entrez l'extension à ajouter aux nouveaux noms de champ. Par exemple, vous pouvez ajouter au nom du nouveau champ contenant le logarithme du champ *Salaire actuel* l'extension *log_* et donc l'intituler *log_Salaire actuel*. A l'aide des cases d'option, spécifiez si l'extension doit être ajoutée au nom du champ en tant que préfixe (au début) ou en tant que suffixe (à la fin). Le nom par défaut est Derive *N*, où *N* correspond au nombre de noeuds Calculer créés jusqu'ici dans la session actuelle.

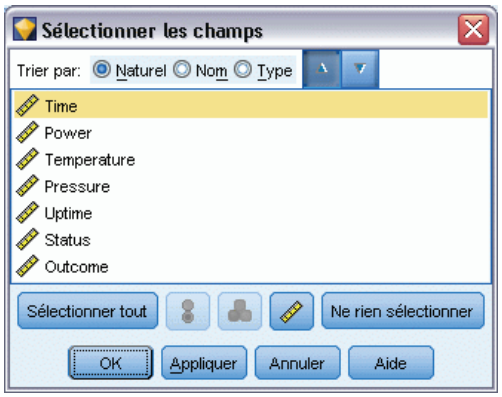
Comme pour les noeuds Calculer en mode Simple, vous devez créer l'expression qui sera utilisée pour calculer le nouveau champ. En fonction du type de l'opération de calcul sélectionnée, vous disposez de plusieurs options pour créer une condition. Ces options sont traitées dans des rubriques ultérieures. Pour créer une expression, vous pouvez simplement la saisir dans les champs de formule ou utiliser le Générateur de formules en cliquant sur le bouton représentant

une calculatrice. N'oubliez pas d'utiliser la fonction @FIELD lorsque les manipulations portent sur plusieurs champs.

Sélection de plusieurs champs

Pour tous les noeuds qui effectuent des opérations sur plusieurs champs d'entrée, tels que les noeuds Calculer (mode Multiple), Agréger, Trier, Courbes et Tracé horaire, vous pouvez facilement sélectionner plusieurs champs à l'aide de la boîte de dialogue Sélection des champs.

Figure 4-44
Sélection de plusieurs champs



Trier par. Vous pouvez trier les champs disponibles à afficher en sélectionnant l'une des options suivantes :

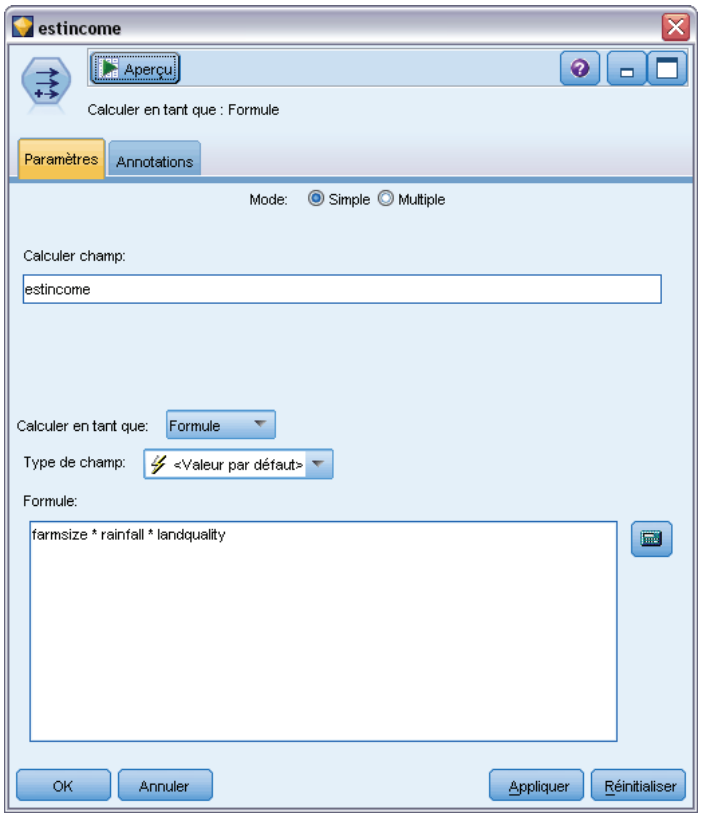
- **Naturel.** Permet d'afficher l'ordre dans lequel les champs ont été transmis via le flux de données dans le nœud actuel.
- **Nom.** Permet de trier les champs à afficher dans l'ordre alphabétique.
- **Type.** Permet de trier les champs en fonction de leur niveau de mesure. Cette option est utile lors de la sélection de champs avec un niveau de mesure particulier.

Sélectionnez les champs un par un dans la liste, ou utilisez la méthode de sélection Maj+clic ou Ctrl+clic pour sélectionner plusieurs champs. Vous pouvez également utiliser les boutons situés en dessous de la liste pour sélectionner des groupes de champs en fonction de leur niveau de mesure, ou pour sélectionner ou désélectionner tous les champs du tableau.

Paramétrage des options du nœud de calcul Formule

Les noeuds de formule Calculer créent un champ pour chaque enregistrement d'un ensemble de données, en fonction des résultats d'une expression CLEM. Notez que l'expression ne peut pas être conditionnelle. Pour calculer des valeurs en fonction d'une expression conditionnelle, vous devez utiliser le type Booléen ou Conditionnel du nœud Calculer.

Figure 4-45
Paramétrage des options d'un nœud de formule Calculer

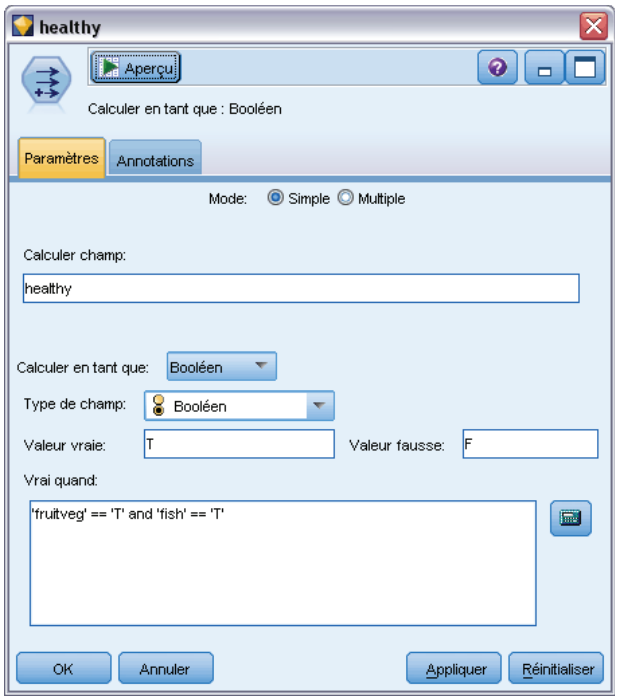


Formule. Indiquez une formule utilisant le langage CLEM pour calculer la valeur du nouveau champ.

Paramétrage des options du nœud de calcul Booléen

Les noeuds booléens Calculer permettent d'indiquer une condition spécifique, telle qu'une tension artérielle élevée ou l'inactivité d'un compte client. Un champ booléen est créé pour chaque enregistrement, et, lorsque la condition true (vrai) est satisfaite, la valeur booléenne correspondante est ajoutée dans le champ.

Figure 4-46
Calcul d'un champ booléen



Valeur vraie (true). Indiquez la valeur à inclure dans le champ booléen pour les enregistrements qui respectent la condition spécifiée plus bas. La valeur par défaut est T.

Valeur fausse (false). Indiquez la valeur à inclure dans le champ booléen pour les enregistrements qui ne respectent *pas* la condition spécifiée plus bas. La valeur par défaut est F.

Vrai quand. Indiquez une condition CLEM pour évaluer certaines valeurs de chaque enregistrement et attribuer à l'enregistrement la valeur true (vrai) ou false (faux) (définie plus haut). Remarque : la valeur true (vrai) est attribuée aux enregistrements dans le cas des valeurs numériques non fausses (false).

Remarque : pour renvoyer une chaîne vide, vous devez saisir des guillemets d'ouverture et de fermeture, sans rien entre les deux (""). Les chaînes vides sont souvent utilisées, par exemple, comme valeur false (faux) afin de permettre aux valeurs true (vrai) de ressortir plus clairement dans un tableau. De la même manière, ayez recours aux guillemets pour utiliser une valeur de chaîne qui serait traitée en tant que nombre autrement.

Exemple

Dans les versions de IBM® SPSS® Modeler antérieures à 12.0, des réponses multiples ont été importées dans un champ unique, avec des valeurs séparées par des virgules.

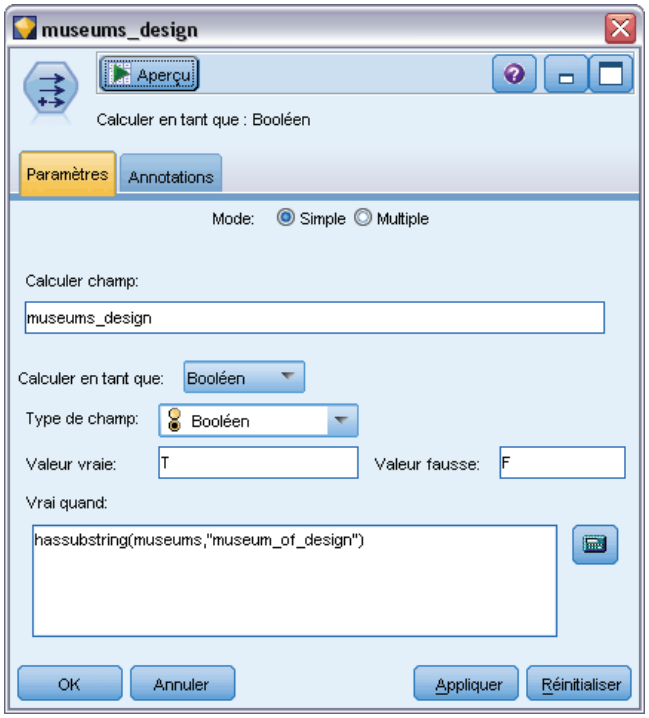
musées
musée_du_design,institut_des_textiles_et_de_la_mode
musée_du_design

musées
musée_archéologique
$null$
musée_des_beaux_arts,musée_des_sciences,autre

Afin de préparer ces données pour l'analyse, vous pouvez utiliser la fonction hassubstring afin de générer un champ booléen distinct pour chaque réponse ; entrez pour cela une expression semblable à ce qui suit :

```
hassubstring(museums,"museum_of_design")
```

Figure 4-47
Calcul d'un champ booléen à l'aide de la fonction hassubstring



Paramétrage des options du nœud de calcul Ensemble

Les noeuds d'ensemble Calculer sont utilisés pour exécuter un ensemble de conditions CLEM afin de déterminer la condition remplie par chaque enregistrement. Chaque fois qu'une condition est remplie pour un enregistrement, une valeur (indiquant l'ensemble de conditions rempli) est ajoutée au nouveau champ calculé.

Figure 4-48
Utilisation d'un noeud d'ensemble Calculer

Valeur par défaut. Indiquez la valeur à utiliser dans le nouveau champ si aucune condition n'est satisfaite.

Définir le champ sur. Indiquez la valeur à entrer dans le nouveau champ lorsqu'une condition particulière est satisfaite. Chaque valeur de la liste est associée à une condition que vous spécifiez dans la colonne adjacente.

Si cette condition est vraie. Indiquez une condition pour chaque membre du champ d'ensemble à répertorier. Utilisez le Générateur de formules pour faire votre choix parmi les fonctions et les champs disponibles. Vous pouvez utiliser les flèches et le bouton Supprimer pour réorganiser ou supprimer des conditions.

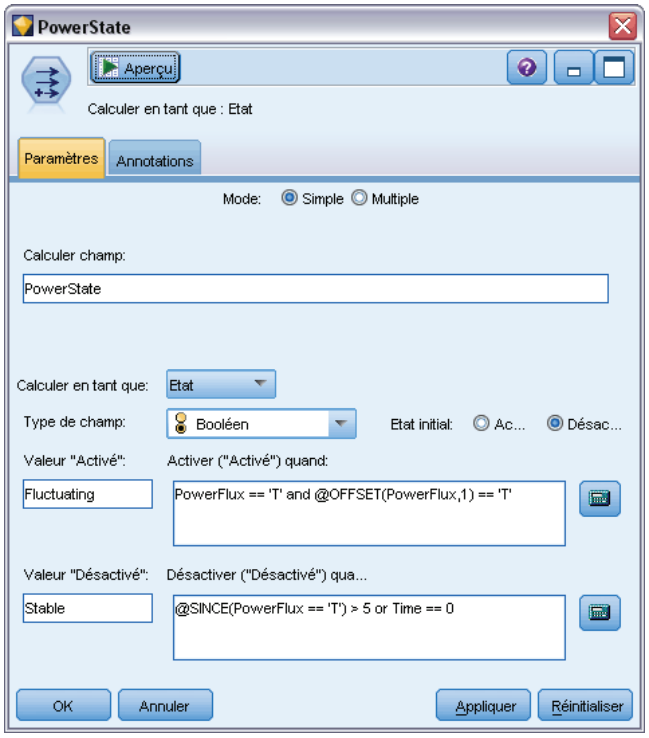
Une condition fonctionne en testant les valeurs d'un champ particulier de l'ensemble de données. Au fur et à mesure que les conditions sont testées, les valeurs spécifiées plus haut sont assignées au nouveau champ pour indiquer les éventuelles conditions satisfaites. Si aucune condition n'est satisfaite, la valeur par défaut est utilisée.

Paramétrage des options du noeud de calcul Etat

Les noeuds d'état Calculer sont semblables aux noeuds booléens Calculer. Un noeud Booléen définit des valeurs si une condition *unique* est satisfaite ou non pour l'enregistrement actuel. Le noeud d'état Calculer, quant à lui, peut modifier les valeurs d'un champ en fonction de sa réponse

à *deux conditions indépendantes*. Autrement dit, la valeur est modifiée (activée ou désactivée) en fonction de la réponse à chaque condition.

Figure 4-49
Utilisation d'un noeud d'état Calculer



Etat initial. Indiquez si vous souhaitez attribuer à chaque enregistrement du nouveau champ la valeur initiale Activé ou Désactivé. Cette valeur peut changer au fur et à mesure que les conditions sont respectées.

Valeur "Activé". Indiquez la valeur du nouveau champ si la condition Activé est vérifiée.

Activer quand. Choisissez la condition CLEM qui détermine le passage à l'état activé lorsque la condition a la valeur true (vrai). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Valeur "Désactivé". Indiquez la valeur du nouveau champ si la condition Désactivé est vérifiée.

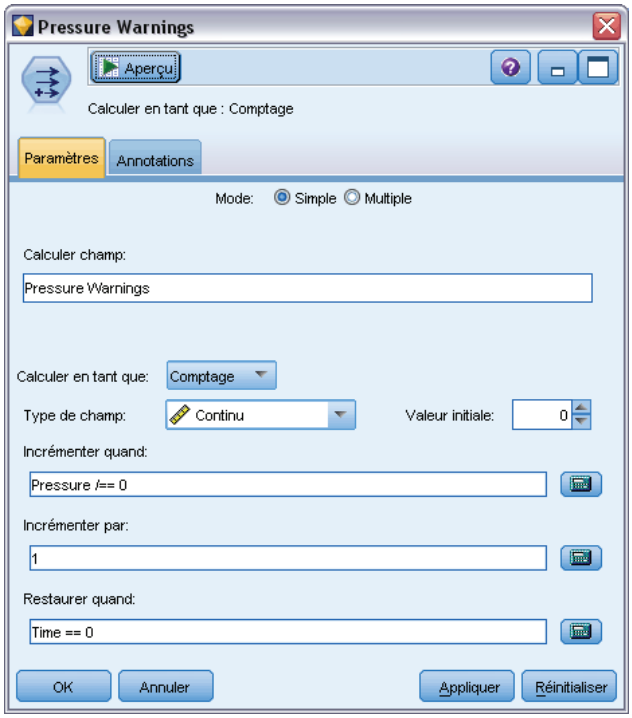
Désactiver quand. Choisissez la condition CLEM qui détermine le passage à l'état désactivé lorsque la condition a la valeur false (faux). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Remarque : pour spécifier une chaîne vide, vous devez saisir des guillemets d'ouverture et de fermeture, sans rien entre les deux (""). De la même manière, ayez recours aux guillemets pour utiliser une valeur de chaîne qui serait traitée en tant que nombre autrement.

Paramétrage des options du noeud de calcul Comptage

Les noeuds de calcul Calculer sont utilisés pour appliquer une série de conditions aux valeurs d'un champ numérique de l'ensemble de données. Au fur et à mesure que les conditions sont respectées, la valeur du champ Comptage calculé augmente en fonction de l'incrément spécifié. Ce type de noeud Calculer est pratique pour les séries temporelles.

Figure 4-50
Options Comptage de la boîte de dialogue du noeud Calculer



Valeur initiale. Définit une valeur utilisée pour le nouveau champ lors de l'exécution. La valeur initiale doit être une constante numérique. Utilisez les flèches pour augmenter ou diminuer la valeur.

Incrémenter quand. Spécifiez la condition CLEM qui, lorsqu'elle est satisfaite, modifie la valeur calculée en se basant sur le chiffre indiqué dans Incrémenter par. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

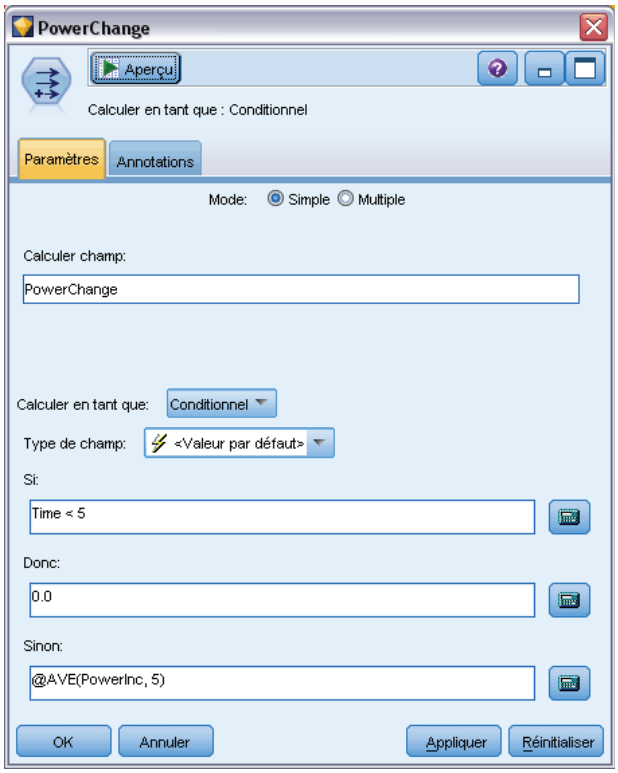
Incrémenter par. Définissez la valeur de l'incrément. Vous pouvez utiliser une constante numérique ou le résultat d'une expression CLEM.

Restaurer quand. Indiquez la condition qui, lorsqu'elle est satisfaite, restaure la valeur calculée sur sa valeur initiale. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Paramétrage des options du noeud de calcul Conditionnel

Les noeuds conditionnels Calculer utilisent la série d'instructions If-Then-Else pour le calcul de la valeur du nouveau champ.

Figure 4-51
Utilisation d'un noeud conditionnel Calculer



Si. Indiquez une condition CLEM qui sera évaluée pour chaque enregistrement lors de l'exécution. Si cette condition a la valeur true (vrai) (ou qu'elle n'est pas false (faux), dans le cas de valeurs numériques), le nouveau champ reçoit la valeur indiquée à côté de l'expression Donc ci-dessous. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

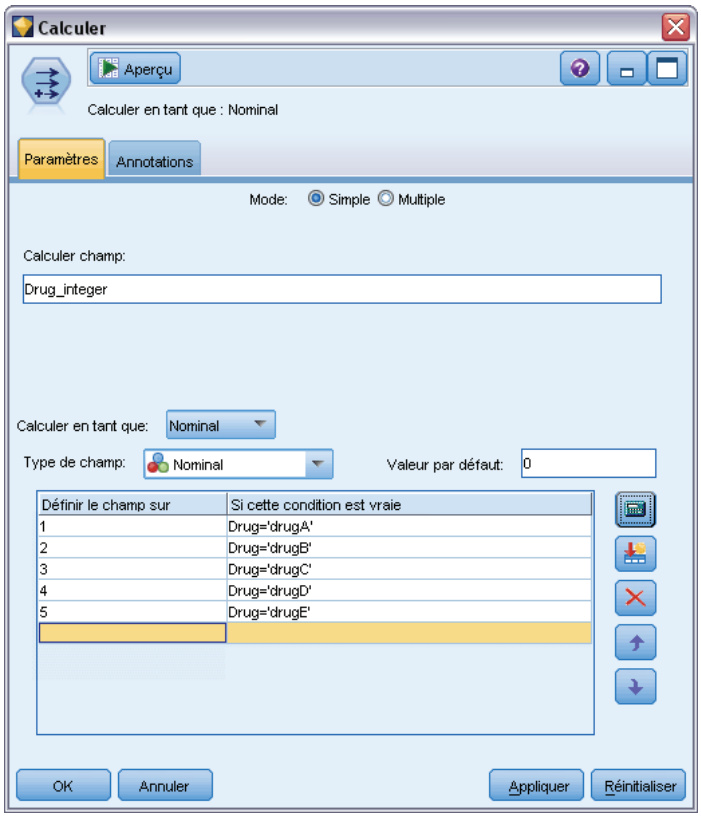
Donc. Indiquez la valeur ou l'expression CLEM à utiliser pour le nouveau champ si l'instruction Si ci-dessus a la valeur true (vrai) (ou non-false (non fausse)). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Sinon. Indiquez la valeur ou l'expression CLEM à utiliser pour le nouveau champ si l'instruction Si ci-dessus a la valeur false (faux). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Recodage des valeurs à l'aide du noeud Calculer

Les noeuds Calculer peuvent également être utilisés pour recoder des valeurs, en convertissant, par exemple, un champ de type chaîne comportant des valeurs catégorielles en champ nominal (ensemble) numérique.

Figure 4-52
Recodage des valeurs de chaînes



- Pour l'option Calculer en tant que, sélectionnez le type de champ (Nominal, Booléen, etc.) approprié.
- Indiquez les conditions de recodage des valeurs. Par exemple, vous pouvez définir la valeur sur 1 si Drug='drugA', 2 si Drug='drugB', etc.

Noeud Remplacer

Les noeuds Remplacer sont utilisés pour remplacer les valeurs de champ et pour modifier le stockage. Vous pouvez décider de remplacer les valeurs reposant sur une condition CLEM spécifiée, telle que @BLANK(FIELD). Vous pouvez également choisir de remplacer tous les blancs ou toutes les valeurs nulles par une valeur précise. Les noeuds Remplacer sont souvent utilisés avec le noeud Typer pour remplacer des valeurs manquantes. Par exemple, vous pouvez remplacer des blancs avec la valeur moyenne d'un champ en spécifiant une expression telle que @GLOBAL_MEAN. Cette expression remplace tous les blancs par la valeur moyenne calculée par le noeud V. globales (Valeurs globales).

Figure 4-53
Boîte de dialogue du noeud Remplacer

Renseigner les champs. A l'aide du sélecteur de champs (bouton situé à droite du champ de texte), sélectionnez les champs de l'ensemble de données dont vous souhaitez analyser et remplacer les valeurs. Le comportement par défaut consiste à remplacer les valeurs en fonction des expressions Condition et Remplacer par spécifiées plus bas. Vous pouvez également choisir une autre méthode de remplacement à l'aide des options Remplacer ci-dessous.

Remarque : lorsque vous sélectionnez plusieurs champs à remplacer par une valeur définie par l'utilisateur, il est important que les champs soient de même type (numériques ou symboliques).

Remplacer. Sélectionnez cette option pour remplacer les valeurs des champs sélectionnés à l'aide de l'une des méthodes suivantes :

- **Basé sur une condition.** Cette option active le champ Condition et le Générateur de formules pour vous permettre de créer une expression utilisée comme condition pour le remplacement par la valeur spécifiée.
- **Toujours.** Remplace toutes les valeurs du champ sélectionné. Par exemple, vous pouvez utiliser cette option pour convertir le stockage du revenu en une chaîne, grâce à l'expression CLEM suivante : (to_string(income)).
- **Valeurs non renseignées.** Remplace toutes les valeurs vides spécifiées par l'utilisateur dans le champ sélectionné. La condition standard @BLANK(@FIELD) est utilisée pour sélectionner les blancs. *Remarque* : vous pouvez définir les blancs via l'onglet Types du noeud source ou via un noeud Typer.

- **Valeurs nulles.** Remplace toutes les valeurs système nulles dans le champ sélectionné. La condition standard @NULL(@FIELD) est utilisée pour sélectionner les valeurs nulles.
- **Valeurs nulles et non renseignées.** Remplace les valeurs non renseignées et les valeurs système nulles dans le champ sélectionné. Cette option est utile lorsque vous ne savez pas avec certitude si les valeurs nulles ont été définies comme valeurs manquantes.

Condition. Cette option est disponible lorsque vous avez sélectionné l'option Basé sur une condition. Cette zone de texte vous permet d'indiquer une expression CLEM pour l'évaluation des champs sélectionnés. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

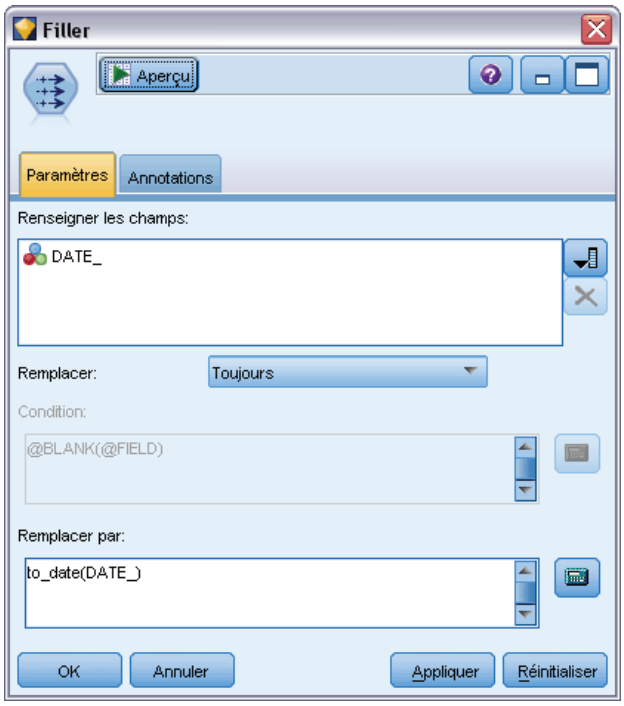
Remplacer par. Indiquez une expression CLEM pour attribuer une nouvelle valeur aux champs sélectionnés. Vous pouvez également remplacer la valeur par une valeur nulle en entrant undef dans la zone de texte. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Remarque : lorsque les champs sélectionnés sont des chaînes, vous devez les remplacer par une valeur de chaîne. Si vous utilisez la valeur par défaut 0 ou une autre valeur numérique en tant que valeur de remplacement pour les champs de type chaîne, une erreur est générée.

Conversion du stockage à l'aide du noeud Remplacer

En utilisant la condition Remplacer d'un noeud Remplacer, vous pouvez facilement convertir le type de stockage d'un ou de plusieurs champs. Par exemple, la fonction de conversion to_integer permet de convertir le type chaîne d'un *revenu* en un type entier, grâce à l'expression CLEM suivante : to_integer(income).

Figure 4-54
Utilisation d'un noeud Remplacer pour convertir le stockage d'un champ



Vous pouvez afficher les fonctions de conversion disponibles et créer automatiquement une expression CLEM en utilisant le Générateur de formules. Dans la liste déroulante Fonctions, sélectionnez Conversion pour afficher la liste des fonctions de conversion de stockage. Les fonctions de conversion suivantes sont disponibles :

- to_integer(ITEM)
- to_real(ITEM)
- to_number(ITEM)
- to_string(ITEM)
- to_time(ITEM)
- to_timestamp(ITEM)
- to_date(ITEM)
- to_datetime(ITEM)

Conversion des valeurs date et heure. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue des options de flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez MOIS AAAA comme format de date par défaut pour le flux.

Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne

contenant des valeurs catégorielles. Pour plus d'informations, reportez-vous à la section Recodage des valeurs à l'aide du noeud Calculer sur p. 177.

Noeud Anonymiser

Le noeud Anonymiser permet de masquer les noms de champ, les valeurs de champ ou les deux types de données lorsque vous travaillez avec des données à inclure dans un modèle situé en aval du noeud. De cette façon, vous pouvez distribuer librement le modèle généré (par exemple à l'assistance technique) sans craindre que des utilisateurs non autorisés aient la possibilité de visualiser des données confidentielles telles que les fichiers du personnel ou les dossiers médicaux de patients.

Selon l'endroit où vous placez le noeud Anonymiser dans le flux, il se peut que vous deviez apporter des modifications à d'autres noeuds. Par exemple, si vous insérez un noeud Anonymiser en amont d'un noeud Sélectionner, les critères de sélection de ce dernier doivent être modifiés s'ils agissent sur des valeurs qui sont désormais anonymisées.

La méthode à utiliser pour l'anonymisation dépend de plusieurs facteurs. Pour les noms de champ, ainsi que pour toutes les valeurs de champ excepté les niveaux de mesure continus, les données sont remplacées par une chaîne du type :

*prefix*_S*n*

où *prefix*_ est une chaîne définie par l'utilisateur ou la chaîne par défautanon_ et *n* est une valeur entière qui commence à 0 et qui est incrémentée pour chaque valeur unique (par exemple, anon_S0, anon_S1, etc.).

Les valeurs de champ du type Continu doivent être transformées car les intervalles numériques se rapportent à des valeurs entières ou réelles plutôt qu'à des chaînes. En tant que telles, elles peuvent être anonymisées uniquement par la transformation de l'intervalle en un intervalle différent. Les données d'origine sont ainsi masquées. La transformation d'une valeur *x* de l'intervalle est exécutée de la façon suivante :

A*(*x* + B)

où :

A est un facteur d'échelle, obligatoirement supérieur à 0.

B est un décalage de translation à ajouter aux valeurs.

Exemple

Soit un champ *AGE* avec le facteur d'échelle *A* défini sur 7 et le décalage de translation *B* défini sur 3, les valeurs relatives à *AGE* sont transformées de la façon suivante :

7*(AGE + 3)

Paramétrage des options du noeud Anonymiser

Ici, vous pouvez choisir les champs qui auront leurs valeurs masquées plus en aval.

Les champs de données doivent être instanciés en amont du noeud Anonymiser pour que les opérations d'anonymisation puissent être exécutées. Pour instancier les données, cliquez sur le bouton Lire les valeurs d'un noeud Typer ou sur l'onglet Types d'un noeud source.

Figure 4-55
Paramétrage des options du noeud Anonymiser

Champ. Répertorie les champs de l'ensemble de données actuel. Si des noms de champ ont déjà été anonymisés, ils apparaissent ici.

Mesure. Niveau de mesure du champ.

Anonymiser des valeurs. Sélectionnez un ou plusieurs champs, cliquez sur cette colonne et choisissez Oui pour anonymiser la valeur de champ à l'aide du préfixe par défaut anon_ ; choisissez Spécifier pour afficher une boîte de dialogue qui permet de saisir votre propre préfixe ou, dans le cas de valeurs de champ de type *Continu*, indiquez si la transformation des valeurs de champ doit utiliser des valeurs aléatoires ou définies par l'utilisateur. Il n'est pas possible de spécifier au cours de la même opération des types de champ *Continu* et non-*Continu* ; vous devez spécifier chaque type séparément.

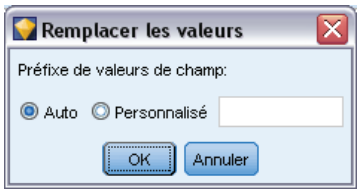
Afficher les champs actuels. Sélectionnez cette option pour afficher les champs des ensembles de données connectés au noeud Anonymiser. Par défaut, cette option est sélectionnée.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher les champs des ensembles de données qui étaient auparavant connectés au noeud. Cette option est utile lorsque vous copiez des noeuds d'un flux à un autre, ou lorsque vous enregistrez ou rechargez des noeuds.

Spécification des modalités d'anonymisation des valeurs de champ

La boîte de dialogue Remplacer les valeurs permet de choisir entre l'utilisation du préfixe par défaut pour les valeurs de champ anonymisées et l'utilisation d'un préfixe personnalisé. Lorsque vous cliquez sur OK dans cette boîte de dialogue, le paramètre de l'option Anonymiser des valeurs de l'onglet Paramètres devient Oui pour le ou les champs sélectionnés.

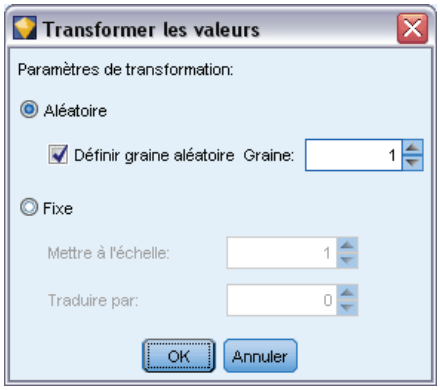
Figure 4-56
Boîte de dialogue Remplacer les valeurs



Préfixe de valeurs de champ. Le préfixe par défaut pour les valeurs de champ anonymisées est anon_ ; sélectionnez Personnalisé et entrez, si vous le souhaitez, votre propre préfixe.

La boîte de dialogue Transformer les valeurs apparaît uniquement pour les champs du type Continu ; elle permet de spécifier si la transformation des valeurs de champ doit utiliser des valeurs aléatoires ou définies par l'utilisateur.

Figure 4-57
Boîte de dialogue Transformer les valeurs



Aléatoire. Sélectionnez cette option afin d'utiliser des valeurs aléatoires pour la transformation. L'option Définir graine aléatoire est sélectionnée par défaut. Spécifiez une valeur dans le champ Graine ou utilisez la valeur par défaut.

Fixe : Sélectionnez cette option afin de définir vos propres valeurs pour la transformation.

- **Mettre à l'échelle.** nombre par lequel les valeurs de champ sont multipliées dans la transformation. La valeur minimale est 1. La valeur maximale est normalement de 10, mais elle peut être diminuée pour éviter tout dépassement.
- **Traduire par.** nombre qui sera ajouté aux valeurs de champ dans la transformation. La valeur minimale est 0. La valeur maximale est normalement de 1000, mais elle peut être diminuée pour éviter tout dépassement.

Anonymisation des valeurs de champ

Les valeurs des champs sélectionnés pour l'anonymisation dans l'onglet Paramètres sont anonymisées :

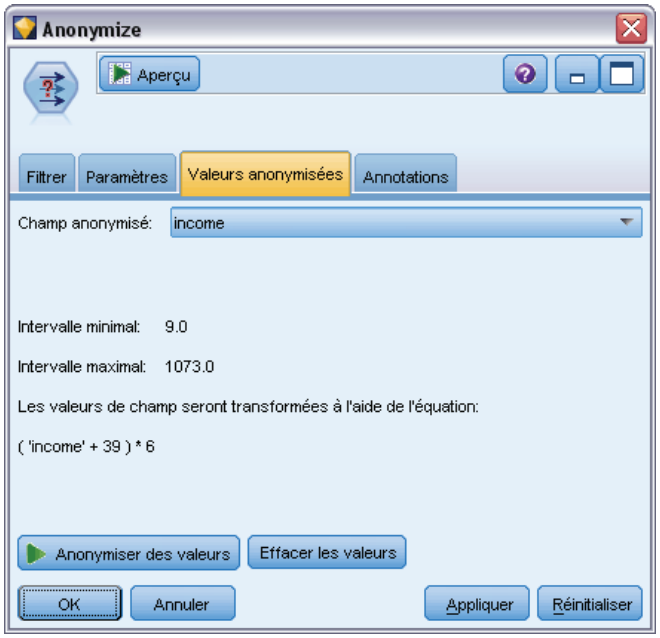
- lorsque vous exécutez le flux contenant le noeud Anonymiser ;
- lorsque vous prévisualisez les valeurs.

Pour prévisualiser les valeurs, cliquez sur le bouton Anonymiser des valeurs dans l'onglet Valeurs anonymisées. Sélectionnez ensuite un nom de champ dans la liste déroulante.

Si le niveau de mesure est de type Continu, les éléments suivants s'affichent :

- valeurs minimale et maximale de l'intervalle d'origine
- équation utilisée pour transformer les valeurs

Figure 4-58
Anonymisation des valeurs de champ



Si le niveau de mesure est différent de la valeur Continue, l'écran affiche la valeur d'origine et la valeur anonymisée pour ce champ.

Figure 4-59
Anonymisation des valeurs de champ

Un affichage sur fond jaune indique que le paramètre du champ sélectionné a changé depuis la dernière anonymisation des valeurs ou que des modifications ont été apportées aux données situées en amont du noeud Anonymiser, de sorte que les valeurs anonymisées ne sont peut-être plus correctes. L'ensemble actuel de valeurs apparaît ; cliquez de nouveau sur le bouton Anonymiser des valeurs pour générer un nouvel ensemble de valeurs conforme au paramètre actuel.

Anonymiser des valeurs. Crée des valeurs anonymisées pour le champ sélectionné et les affiche dans le tableau. Si vous utilisez une graine aléatoire pour un champ de type Continu, le fait de cliquer sur ce bouton à plusieurs reprises crée un ensemble de valeurs différent à chaque fois.

Effacer les valeurs. Efface les valeurs d'origine et les valeurs anonymisées du tableau.

Noeud Recoder

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser. Par exemple, vous pouvez recoder les valeurs du nom *Produit* en trois groupes, comme *Ustensiles de cuisine*, *Salle de bains et linge* et *Appareils ménagers*. Cette opération est généralement exécutée directement à partir d'un noeud Proportion par regroupement des valeurs et génération d'un noeud Recoder. Pour plus d'informations, reportez-vous à la section Utilisation d'un noeud Proportion dans le chapitre 5 sur p. 314.

La recodification peut s'effectuer pour un ou plusieurs champs symboliques. Vous pouvez également décider de remplacer les valeurs d'un champ existant par de nouvelles valeurs ou de générer un nouveau champ.

Avant d'utiliser un noeud Recoder, assurez-vous qu'aucun autre noeud d'opérations sur les champs n'est plus adéquat pour cette tâche :

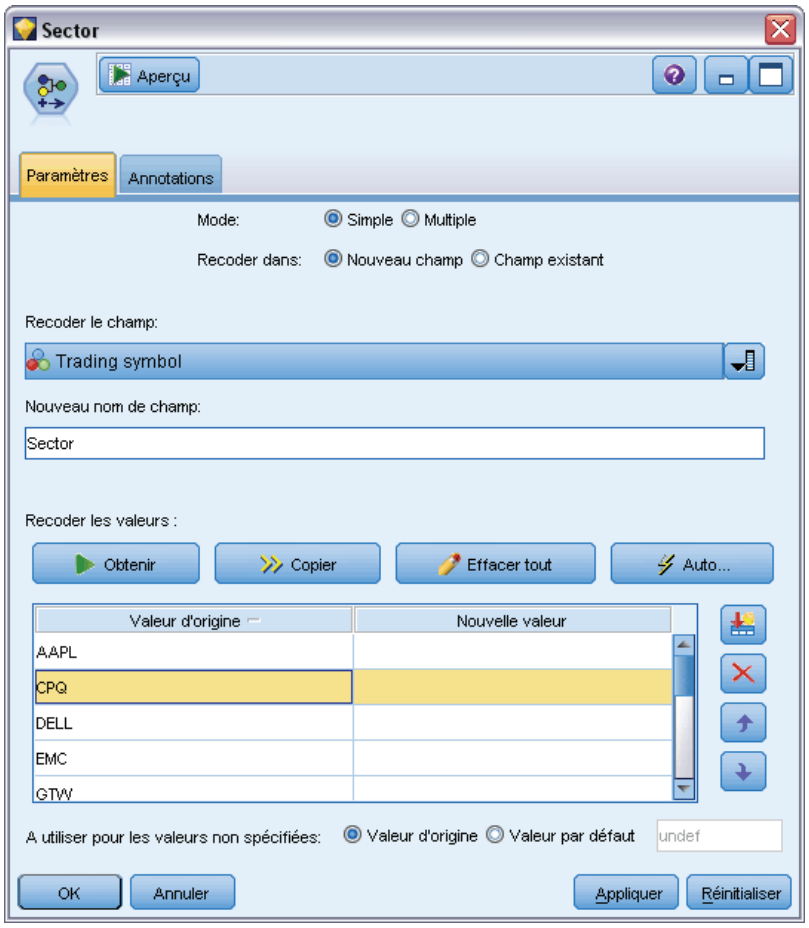
- Pour transformer des intervalles numériques en ensembles à l'aide d'une méthode automatique, telle que celle des rangs ou des centiles, vous devez utiliser un noeud Discrétiser. Pour plus d'informations, reportez-vous à la section Noeud Discrétiser sur p. 191.
- Pour classifier manuellement des intervalles numériques en ensembles, vous devez utiliser un noeud Calculer. Par exemple, si vous souhaitez réduire des valeurs de salaire en catégories d'intervalle salarial, vous devez utiliser un noeud Calculer pour définir chaque catégorie manuellement.
- Pour créer un ou plusieurs champs booléens sur la base des valeurs d'un champ catégoriel, tel que *Type_hypothèque*, vous devez utiliser un noeud Binariser.
- Pour convertir un champ catégoriel en stockage numérique, vous pouvez utiliser un noeud Calculer. Vous pouvez, par exemple, convertir les valeurs *Non* et *Oui* respectivement en valeurs 0 et 1. Pour plus d'informations, reportez-vous à la section Recodage des valeurs à l'aide du noeud Calculer sur p. 177.

Paramétrage des options du noeud Recoder

L'utilisation du noeud Recoder se divise en trois étapes :

- Tout d'abord, choisissez si vous souhaitez recodifier plusieurs champs ou un seul champ.
- Ensuite, choisissez soit de recoder un champ existant, soit de créer un nouveau champ.
- Enfin, utilisez les options dynamiques de la boîte de dialogue du noeud Recoder pour mapper les ensembles comme vous le souhaitez.

Figure 4-60
Boîte de dialogue du noeud Recoder



Mode. Sélectionnez Simple pour recodifier les catégories d'un champ. Sélectionnez Multiple pour activer les options permettant de transformer plusieurs champs simultanément.

Recoder dans. Sélectionnez Nouveau champ pour conserver le champ nominal d'origine et calculer un autre champ contenant les valeurs recodifiées. Sélectionnez Champ existant pour écraser les valeurs du champ d'origine et les remplacer par les nouvelles classifications. Il s'agit avant tout d'une opération de type remplacement.

Lorsque vous avez indiqué le mode et les options de remplacement, vous devez sélectionner le champ de transformation et indiquer les nouvelles valeurs de classification à l'aide des options dynamiques situées dans la moitié inférieure de la boîte de dialogue. Ces options varient en fonction du mode sélectionné plus haut.

Recoder les champs. Utilisez le sélecteur de champs à droite pour sélectionner un (mode Simple) ou plusieurs (mode Multiple) champs catégoriels.

Nouveau nom de champ. Indiquez le nom du nouveau champ nominal contenant les valeurs recodées. Cette option n'est disponible qu'en mode Simple lorsque l'option Nouveau champ plus haut est sélectionnée. Lorsque l'option Champ existant est sélectionnée, le nom du champ

d'origine est conservé. Lorsque vous travaillez en mode Multiple, cette option est remplacée par des contrôles permettant de spécifier une extension ajoutée à chaque nouveau champ. Pour plus d'informations, reportez-vous à la section Recodification de plusieurs champs sur p. 190.

Recoder les valeurs. Ce tableau établit un mappage clair entre les anciennes valeurs d'ensemble et celles que vous indiquez ici.

- **Valeur d'origine.** Cette colonne répertorie les valeurs existantes des champs sélectionnés.
- **Nouvelle valeur.** Utilisez cette colonne pour saisir les nouvelles valeurs de catégorie ou sélectionnez-en une dans la liste déroulante. Lorsque vous générez automatiquement un noeud Recoder avec les valeurs provenant d'un graphique Proportion, ces valeurs sont incluses dans la liste déroulante. Cela vous permet de mapper les valeurs existantes rapidement avec un ensemble de valeurs connu. Par exemple, les organisations de santé regroupent parfois les diagnostics différemment selon le réseau et les paramètres régionaux. Après une fusion ou un rachat, toutes les parties doivent recodifier les données nouvelles ou même existantes de manière homogène. Au lieu d'attribuer un type manuellement à chaque cible à partir d'une longue liste, vous pouvez lire la principale liste des valeurs dans IBM® SPSS® Modeler, exécuter un graphique Proportion pour le champ *Diagnostic*, et générer un nœud Recoder (valeurs) pour ce champ directement à partir du graphique. Ce processus rend toutes les valeurs Diagnostic cible disponibles à partir de la liste déroulante Nouvelles valeurs.

▶ Cliquez sur Obtenir pour lire les valeurs d'origine d'un ou de plusieurs des champs sélectionnés plus haut.

▶ Cliquez sur Copier pour coller les valeurs d'origine dans la colonne *Nouvelle valeur* pour les champs qui n'ont pas encore été mappés. Les valeurs d'origine non mappées sont ajoutées à la liste déroulante.

▶ Cliquez sur Tout effacer pour effacer toutes les spécifications de la colonne *Nouvelle valeur*. *Remarque* : cette option n'efface pas les valeurs de la liste déroulante.

▶ Cliquez sur Auto pour générer automatiquement des entiers consécutifs pour chacune des valeurs d'origine. Seules les valeurs entières peuvent être générées (les valeurs réelles telles que 1,5 ou 2,5 ne peuvent pas l'être).

Figure 4-61
Boîte de dialogue Classification auto

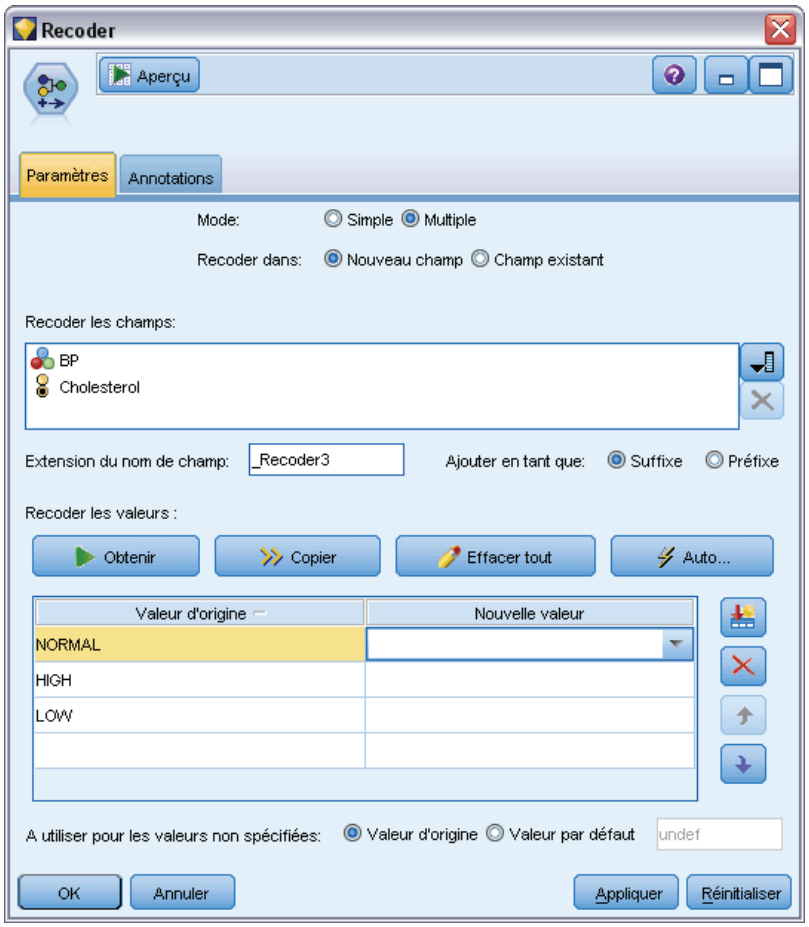
Par exemple, vous pouvez générer automatiquement des numéros d'ID consécutifs pour des noms de produit ou des numéros pour des cours proposés par une université. Cette fonctionnalité correspond à la recodification automatique des ensembles de IBM® SPSS® Statistics.

A utiliser pour les valeurs non spécifiées. Cette option est utilisée pour remplacer les valeurs non spécifiées dans le nouveau champ. Vous pouvez choisir de conserver la valeur d'origine en sélectionnant Valeur d'origine ou d'indiquer une valeur par défaut.

Recodification de plusieurs champs

Pour mapper simultanément les valeurs de catégorie de plusieurs champs, paramétrez le mode sur Multiple. Les nouveaux paramètres décrits ci-dessous sont alors activés dans la boîte de dialogue Recoder.

Figure 4-62
Options de la boîte de dialogue Dynamique pour la recodification de plusieurs champs



Recoder les champs. Utilisez le sélecteur de champs situé à droite pour sélectionner les champs à transformer. Le sélecteur de champs permet de sélectionner tous les champs à la fois ou des champs de même type, par exemple des champs nominaux ou d'intervalle.

Extension nom de champ. Lorsque vous recodez plusieurs champs simultanément, il est plus efficace d'indiquer une extension commune à ajouter à tous les nouveaux champs plutôt qu'un nom différent pour chaque champ. Indiquez une extension telle que `_recode`, et précisez si cette extension doit figurer au début ou à la fin des noms de champ d'origine.

Stockage et niveau de mesure des champs recodifiés

Le noeud Recoder crée toujours un champ de type nominal à partir de l'opération de recodification. Ceci peut entraîner, dans certains cas, la modification du niveau de mesure de champ si vous utilisez le mode de recodification Champ existant.

Le stockage du nouveau champ (mode de *stockage* des données et non *utilisation*) est calculé sur la base des options suivantes de l'onglet Paramètres :

- Si les valeurs non spécifiées sont paramétrées pour utiliser une valeur par défaut, le type de stockage approprié est déterminé par l'examen des nouvelles valeurs et par celui de la valeur par défaut. Par exemple, si toutes les valeurs sont analysées comme des entiers, le champ aura le type de stockage Entier.
- Si les valeurs non spécifiées sont paramétrées pour utiliser les valeurs d'origine, le type de stockage est déterminé par celui du champ d'origine. Si toutes les valeurs sont analysées comme disposant du stockage du champ d'origine, ce stockage est conservé ; sinon, il est déterminé par la recherche du type de stockage le plus adéquat, pour les anciennes comme pour les nouvelles valeurs. Par exemple, la recodification 4 => 0, 5 => 0 de l'ensemble d'entiers { 1, 2, 3, 4, 5 } génère un nouvel ensemble d'entiers { 1, 2, 3, 0 }, tandis que la recodification 4 => "valeurs supérieures à 3", 5 => "valeurs supérieures à 3" génère la chaîne { "1", "2", "3", "valeurs supérieures à 3" }.

Remarque : si le type d'origine n'était pas instancié, le nouveau type ne l'est pas non plus.

Noeud Discrétiser

Le noeud Discrétiser permet de créer automatiquement de nouveaux champs nominaux sur la base des valeurs d'un ou de plusieurs champs continus numériques existants (intervalle numérique). Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus de largeur égale ou comme écarts par rapport à la moyenne. Vous pouvez également sélectionner un champ de superviseur catégoriel afin de conserver la force de l'association d'origine entre deux champs.

La création d'intervalles peut s'avérer utile pour un certain nombre de raisons, notamment :

- **Conditions requises pour l'algorithme.** Certains algorithmes, Naive Bayes ou la régression logistique par exemple, ont besoin d'entrées catégorielles.
- **Performances.** Les algorithmes comme la logistique multinomiale peuvent obtenir de meilleures performances si le nombre de valeurs distinctes des champs d'entrée est réduit. Utilisez par exemple la valeur médiane ou moyenne pour chaque noeud plutôt que la valeur d'origine.
- **Confidentialité des données.** Pour les informations personnelles et confidentielles, par exemple les salaires, vous pouvez indiquer des intervalles plutôt que les chiffres exacts afin de protéger la confidentialité.

Un certain nombre de méthodes de création d'intervalles sont disponibles. Une fois les intervalles du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure.

Avant d'utiliser un noeud Discrétiser, vérifiez si une autre technique est éventuellement plus adéquate pour cette tâche :

- Pour indiquer manuellement les points de césure des catégories, telles que des intervalles salariaux prédéfinis, utilisez un noeud Calculer. Pour plus d'informations, reportez-vous à la section Noeud Calculer sur p. 166.
- Pour créer de nouvelles catégories pour des ensembles existants, utilisez un noeud Recoder. Pour plus d'informations, reportez-vous à la section Noeud Recoder sur p. 186.

Traitement des valeurs manquantes

Le noeud Discrétiser traite les valeurs manquantes de l'une des manières suivantes :

- **Blancs définis par l'utilisateur.** Les valeurs manquantes définies comme des blancs sont incluses dans la transformation. Par exemple, si vous avez indiqué –99 pour indiquer une valeur non renseignée à l'aide du noeud Typer, cette valeur sera incluse dans la création des intervalles. Pour ignorer les blancs au cours de la création des intervalles, utilisez un noeud Remplacer pour remplacer les valeurs non renseignées par la valeur système nulle.
- **Valeurs manquantes système ($null$).** Les valeurs nulles sont ignorées lors de la transformation des intervalles. Elles restent nulles après la transformation.

L'onglet Paramètres propose les options des différentes techniques. L'onglet Affichage affiche les points de césure établis pour les données précédemment passées dans ce noeud.

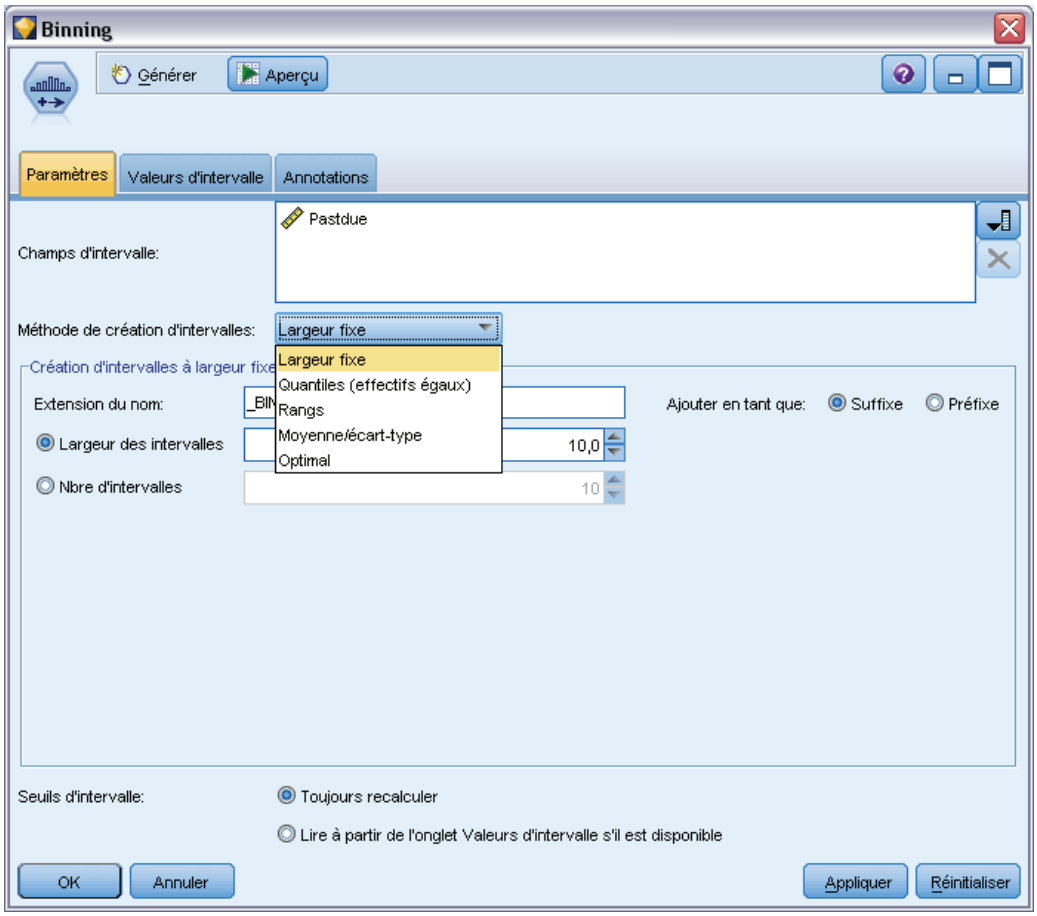
Paramétrage des options du noeud Discrétiser

Le noeud Discrétiser permet de générer automatiquement des intervalles (catégories) à l'aide des techniques suivantes :

- Création d'intervalles à largeur fixe
- Quantiles (effectifs égaux ou somme)
- Moyenne et écart-type
- Rangs
- Optimisé par rapport à un champ de superviseur catégoriel

La partie inférieure de la boîte de dialogue change de manière dynamique en fonction de la méthode de création d'intervalles sélectionnée.

Figure 4-63
Boîte de dialogue du noeud Discrétiser, onglet Paramètres



Champs d'intervalle. Les champs continus (intervalle numérique) en attente de transformation sont affichés ici. Le noeud Discrétiser permet de créer des intervalles pour plusieurs champs simultanément. Ajoutez ou supprimez des champs à l'aide des boutons sur la droite.

Méthode de création d'intervalles. Sélectionnez la méthode utilisée pour déterminer les points de césure des nouveaux intervalles de champ (catégories). Les rubriques suivantes décrivent les options disponibles dans chaque cas.

Seuils des intervalles. Spécifiez comment sont calculés les seuils des intervalles.

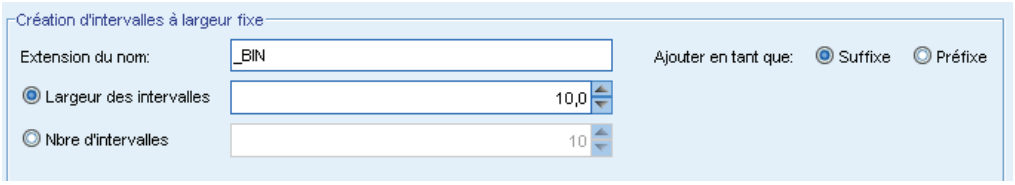
- **Toujours recalculer.** Les points de césure et les affectations d'intervalles sont toujours recalculés lors de l'exécution du nœud.
- **Lire dans l'onglet Valeurs d'intervalle si disponible.** Les points de césure et les attributions d'intervalles sont calculés uniquement en fonction des besoins (par exemple, quand de nouvelles données sont ajoutées).

Les rubriques suivantes présentent les options des méthodes de création d'intervalles disponibles.

Intervalles à largeur fixe

Lorsque vous choisissez la méthode de création d'intervalles Largeur fixe, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Figure 4-64
Boîte de dialogue du noeud Discrétiser (onglet Paramètres) avec options relatives aux intervalles à largeur fixe



Extension du nom. Spécifiez l'extension à utiliser pour les champs générés. L'extension par défaut est *_BIN*. Vous pouvez également indiquer si l'extension doit être ajoutée au début (Préfixe) ou à la fin (Suffixe) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_BIN*.

Largeur des intervalles. Spécifiez la valeur (entier ou réel) utilisée pour le calcul de la "largeur" de l'intervalle. Par exemple, vous pouvez utiliser la valeur par défaut, 10, pour créer les intervalles du champ *Age*. Le champ *Age* couvrant l'intervalle 18–65, les intervalles générés sont les suivants :

Table 4-1
Intervalles du champ Age qui couvre l'intervalle 18–65

Intervalle 1	Intervalle 2	Intervalle 3	Intervalle 4	Intervalle 5	Intervalle 6
>=13 à <23	>=23 à <33	>=33 à <43	>=43 à <53	>=53 à <63	>=63 à <73

Le début des intervalles est calculé de la manière suivante : plus valeur faible analysée moins la moitié de la largeur de l'intervalle indiqué. Par exemple, dans les intervalles ci-dessus, la valeur 13, qui correspond au début des intervalles, a été obtenue grâce au calcul suivant : 18 *[valeur de données la plus faible]* – 5 *[0,5 × (largeur d'intervalle 10)]* = 13.

Nbre d'intervalles. Utilisez cette option pour indiquer un entier déterminant le nombre d'intervalles de largeur fixe (catégories) des nouveaux champs.

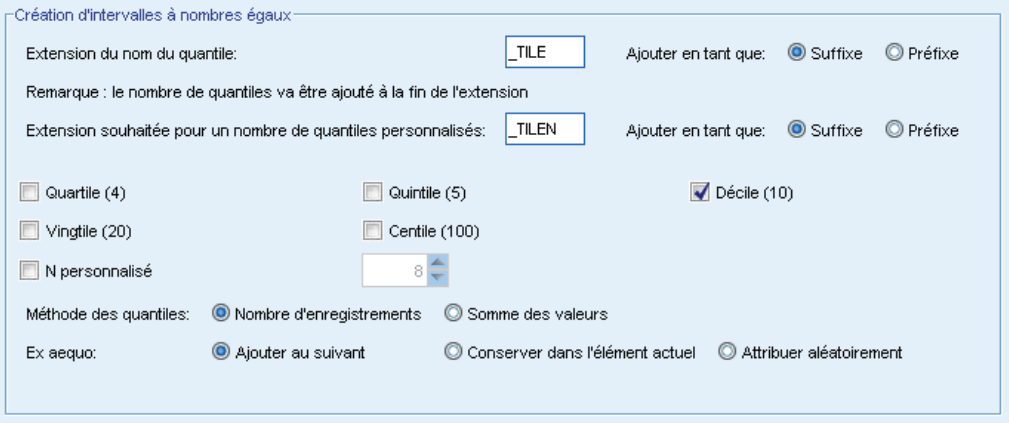
Lorsque vous avez exécuté le noeud Discrétiser dans le cadre d'un flux, vous pouvez afficher les seuils d'intervalle générés en cliquant sur l'onglet Aperçu de la boîte de dialogue du noeud Discrétiser. Pour plus d'informations, reportez-vous à la section Prévisualisation des intervalles générés sur p. 201.

Quantiles (effectifs égaux ou somme)

La méthode de création d'intervalles de type quantile génère des champs nominaux qui peuvent être utilisés pour scinder des enregistrements analysés en groupes de type centiles (ou quartiles, déciles, etc.), de sorte que chaque groupe contienne le même nombre d'enregistrements, ou que la somme des valeurs de chaque groupe soit égale. Les enregistrements sont classés dans l'ordre croissant de la valeur du champ d'intervalle indiqué ; les enregistrements présentant les valeurs les moins élevées pour la variable d'intervalle sélectionnée se voient ainsi attribuer le rang 1, l'ensemble d'enregistrements suivant le rang 2, et ainsi de suite. Les valeurs de seuil

de chaque intervalle sont générées automatiquement en fonction des données et de la méthode des quantiles utilisée.

Figure 4-65
Boîte de dialogue du noeud Discrétiser (onglet Paramètres) avec options relatives aux intervalles à effectifs égaux



Extension du nom du quantile. Spécifiez l'extension utilisée pour les champs générés à l'aide de centiles standard. L'extension par défaut est *_TILE* plus *N*, *N* étant le numéro du quantile. Vous pouvez également indiquer si l'extension doit être ajoutée au début (Préfixe) ou à la fin (Suffixe) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_TILE4*.

Extension personnalisée du quantile. Spécifiez l'extension utilisée pour un intervalle de type quantile personnalisé. La valeur par défaut est *_TILEN*. Dans ce cas, *N* n'est *pas* remplacé par le nombre personnalisé.

Les centiles disponibles sont les suivants :

- **Quartile.** Génère 4 intervalles, chacun contenant 25% des observations.
- **Quintile.** Génère 5 intervalles, chacun contenant 20% des observations.
- **Décile.** Génère 10 intervalles, chacun contenant 10% des observations.
- **Vingtile.** Génère 20 intervalles, chacun contenant 5% des observations.
- **Centile.** Génère 100 intervalles, chacun contenant 1% des observations.
- **N personnalisé.** Sélectionnez cette option pour indiquer le nombre d'intervalles. Par exemple, une valeur de 3 produirait 3 catégories (deux points de césure), chacune contenant 33,3 % des observations.

Si les données contiennent moins de valeurs discrètes que le nombre de quantiles indiqué, tous les quantiles ne sont pas utilisés. La nouvelle proportion peut alors refléter la proportion d'origine des données.

Méthode des quantiles. Indique la méthode utilisée pour affecter des enregistrements à des intervalles.

- **Nombre d'enregistrements.** Cherche à attribuer un nombre égal d'enregistrements à chaque intervalle.
- **Somme des valeurs.** Cherche à attribuer des enregistrements à des intervalles de sorte que la somme des valeurs de chaque intervalle soit égale. Lorsque vous vous intéressez aux efforts de ventes par exemple, cette méthode peut être utilisée pour attribuer des prospects à des groupes de type décile en fonction de la valeur par enregistrement (les prospects qui présentent les valeurs les plus élevées étant placés dans l'intervalle supérieur). Par exemple, une entreprise pharmaceutique peut classer les médecins en groupes de type décile en fonction du nombre d'ordonnances qu'ils rédigent. Alors que chaque décile contient environ le même nombre d'ordonnances, le nombre de personnes à l'origine de ces ordonnances est différent (les personnes qui écrivent le plus d'ordonnances étant regroupées dans le décile 10). Cette approche suppose que toutes les valeurs soient supérieures à zéro ; si tel n'est pas le cas, elle risque de renvoyer des résultats inattendus.

Ex aequo. On parle de condition ex aequo lorsque des valeurs de part et d'autre d'un point de césure sont identiques. Par exemple, si vous utilisez des déciles, et que plus de 10 % des enregistrements présentent la même valeur pour le champ d'intervalle, ces enregistrements ne peuvent pas tous tenir dans le même intervalle sans forcer le seuil d'une façon ou d'une autre. Les valeurs ex aequo peuvent être déplacées vers le haut dans l'intervalle suivant ou conservées dans l'intervalle actuel, à condition qu'elles soient résolues de sorte que tous les enregistrements comportant des valeurs identiques se trouvent dans le même intervalle, et ce, même si cela génère un nombre d'enregistrements par intervalle plus important que prévu. Il est, pour cela, également possible d'ajuster les seuils des intervalles suivants ; les valeurs d'un même ensemble de nombres sont ainsi affectées différemment en fonction de la méthode utilisée pour résoudre les valeurs ex aequo.

- **Ajouter au suivant.** Sélectionnez cette option pour déplacer les valeurs ex aequo vers l'intervalle supérieur suivant.
- **Conserver dans l'élément actuel.** Conserve les valeurs ex aequo dans l'intervalle (inférieur) actuel. Cette méthode peut générer un nombre inférieur d'intervalles.
- **Attribuer de façon aléatoire.** Sélectionnez cette option pour attribuer les valeurs ex æquo de façon aléatoire à un intervalle. Ceci permet de conserver le nombre d'enregistrements dans chaque intervalle de façon égale.

Exemple : Création de quantiles en fonction du nombre d'enregistrements

Le tableau ci-dessous illustre la façon dont les valeurs de champ simplifiées sont classées en quartiles lors de la création de quantiles en fonction du nombre d'enregistrements. Les résultats varient en fonction de l'option de valeurs ex aequo sélectionnée.

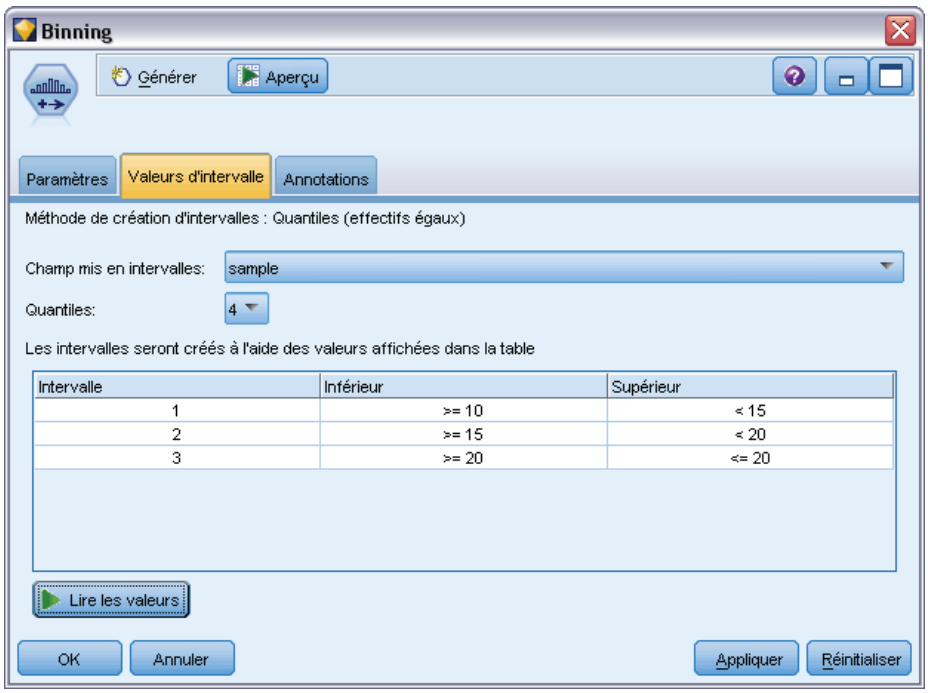
Valeurs	Ajouter au suivant	Conserver dans l'élément actuel
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Le nombre d'éléments par intervalle est calculé de la façon suivante :

total number of value / number of tiles

Dans l'exemple simplifié ci-dessus, le nombre souhaité d'éléments par intervalle est de 1,25 (5 valeurs / 4 quartiles). La valeur 13 (valeur numéro 2) chevauche le seuil de comptage souhaité de 1,25 ; elle est par conséquent traitée différemment selon l'option d'ex aequo sélectionnée. En mode Ajouter au suivant, elle est ajoutée à l'intervalle 2. En mode Conserver dans l'élément actuel, elle reste dans l'intervalle 1, ce qui place l'intervalle des valeurs de l'intervalle 4 en dehors de l'intervalle des valeurs de données existantes. Par conséquent, seuls trois intervalles sont créés et les seuils de chaque intervalle sont ajustés en conséquence.

Figure 4-66
Seuils des intervalles générés

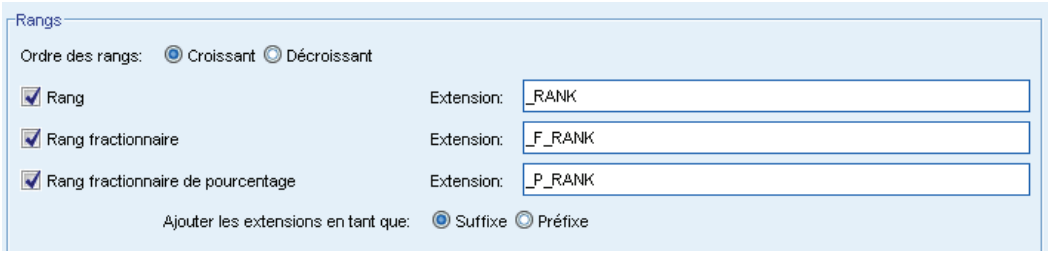


Remarque : L'activation du traitement parallèle peut augmenter la vitesse de création d'intervalles par quantiles.

Classer les observations

Lorsque vous choisissez la méthode de création d'intervalles Rangs, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Figure 4-67
Boîte de dialogue du noeud Discrétiser (onglet Paramètres) avec options relatives aux rangs



Le classement crée de nouveaux champs contenant des rangs, des rangs fractionnaires et des valeurs de centile pour les champs numériques, conformément aux options décrites ci-dessous.

Ordre des rangs. Sélectionnez Croissant (la valeur la plus faible est marquée 1) ou Décroissant (la valeur la plus élevée est marquée 1).

Rang. Sélectionnez cette option pour classer les observations dans l'ordre croissant ou décroissant, comme indiqué ci-avant. L'intervalle des valeurs du nouveau champ sera 1–*N*, *N* étant le nombre de valeurs discrètes présentes dans le champ d'origine. Les valeurs ex aequo reçoivent la moyenne de leur rang.

Rang fractionnaire. Sélectionnez cette option pour classer les observations dans lesquelles la valeur du nouveau champ équivaut au rang divisé par la somme des pondérations des observations non manquantes. Les rangs fractionnaires sont compris dans l'intervalle 0–1.

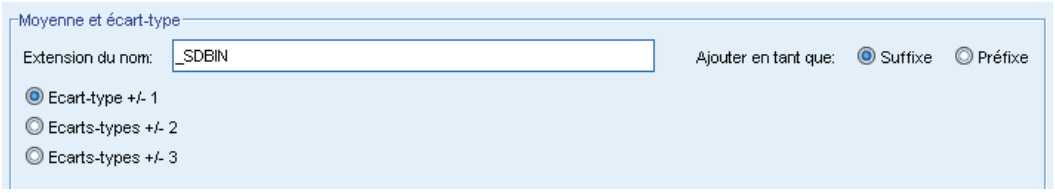
Rang fractionnaire de pourcentage. Chaque rang est divisé par le nombre d'enregistrements avec valeurs valides et multiplié par 100. Les rangs fractionnaires de pourcentage sont compris dans l'intervalle 1–100.

Extension. Pour toutes les options de rang, vous pouvez créer des extensions personnalisées, et indiquer si l'extension doit être ajoutée au début (Préfixe) ou à la fin (Suffixe) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_P_RANK*.

Moyenne/écart-type

Lorsque vous choisissez la méthode de création d'intervalles Moyenne/écart-type, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Figure 4-68
Boîte de dialogue du noeud Discrétiser (onglet Paramètres) avec options relatives à la moyenne/l'écart-type



Cette méthode génère un ou plusieurs nouveaux champs avec catégories en fonction des valeurs de moyenne et d'écart-type de la proportion des champs spécifiés. Sélectionnez le nombre d'écarts à utiliser plus bas.

Extension du nom. Spécifiez l'extension à utiliser pour les champs générés. L'extension par défaut est *_SDBIN*. Vous pouvez également indiquer si l'extension doit être ajoutée au début (Préfixe) ou à la fin (Suffixe) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_SDBIN*.

- **Ecart-type +/– 1.** Sélectionnez cette option pour générer trois intervalles.
- **Ecarts-types +/– 2.** Sélectionnez cette option pour générer cinq intervalles.
- **Ecarts-types +/– 3.** Sélectionnez cette option pour générer sept intervalles.

Par exemple, la sélection de l'option Ecart-type +/–1 génère les trois intervalles calculés ci-dessous :

Intervalle 1	Intervalle 2	Intervalle 3
x < (Mean – Std. Dev)	(Mean – Std. Dev) <= x <= (Mean + Std. Dev)	x > (Mean + Std. Dev)

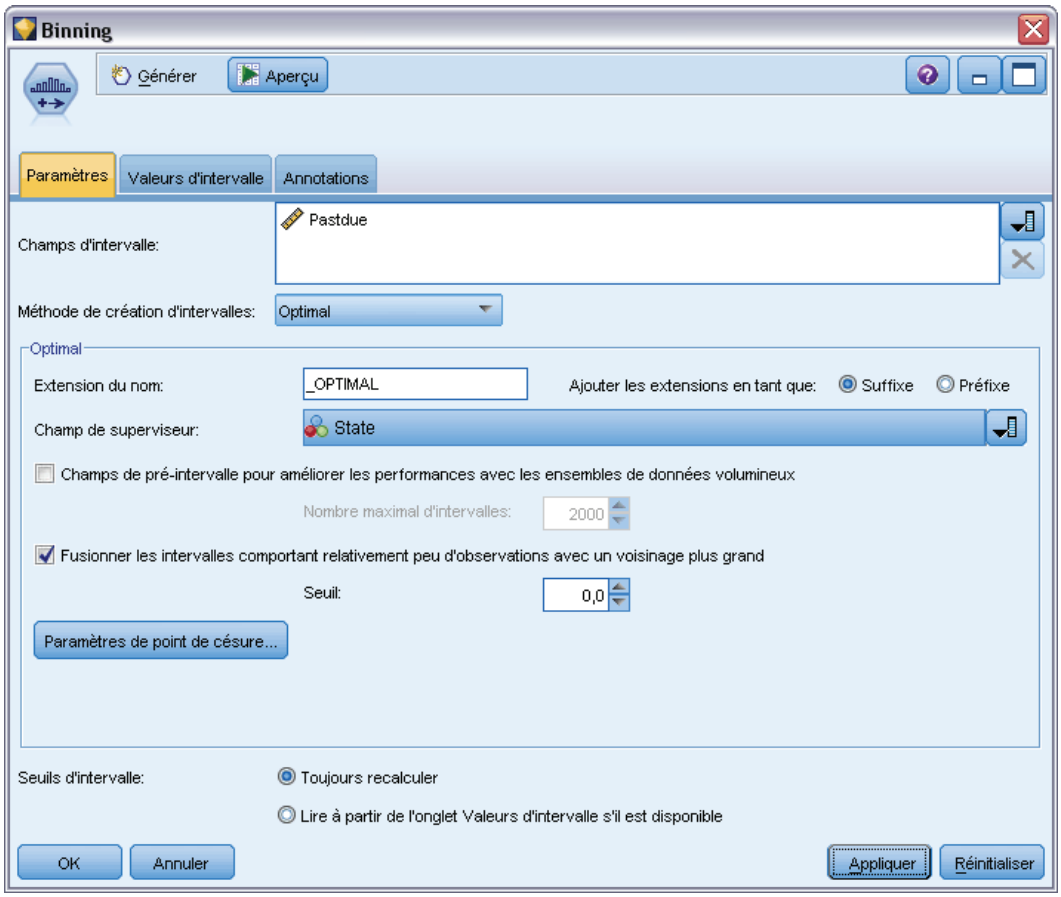
Dans une proportion normale, 68 % des observations sont comprises dans un écart-type par rapport à la moyenne, 95 % dans deux écarts-types et 99 % dans 3 écarts-types. La création de catégories basées sur les écarts-types peut résulter en des intervalles définis en dehors de l'intervalle de données réel et même en dehors de l'intervalle des valeurs de données possibles (par exemple, un intervalle salarial négatif).

Recodage supervisé optimal

Si le champ dans lequel vous souhaitez créer des intervalles est fortement associé à un autre champ catégoriel, vous pouvez sélectionner ce dernier comme champ de superviseur afin de créer les intervalles de façon à préserver la force de l'association d'origine entre les deux champs.

Supposez par exemple que vous ayez utilisé l'analyse des classes pour regrouper les Etats en fonction du taux de prêts immobiliers en souffrance, avec les taux les plus élevés dans la première classe. Dans ce cas, vous pouvez choisir *Pourcentage d'arriéré* et *Pourcentage de forclusion* comme champs d'intervalle, et le champ devant contenir les classes d'appartenance généré par le modèle comme champ de superviseur.

Figure 4-69
Options pour la création d'intervalles optimale ou supervisée



Extension du nom. Indiquez l'extension à utiliser pour les champs générés et déterminez si elle doit être ajoutée au début (Préfixe) ou à la fin (Suffixe) du nom du champ. Vous pouvez par exemple générer deux nouveaux champs nommés *arriéré_OPTIMAL* et *forclusion_OPTIMAL*.

Champ de superviseur. Champ catégoriel utilisé pour construire les intervalles.

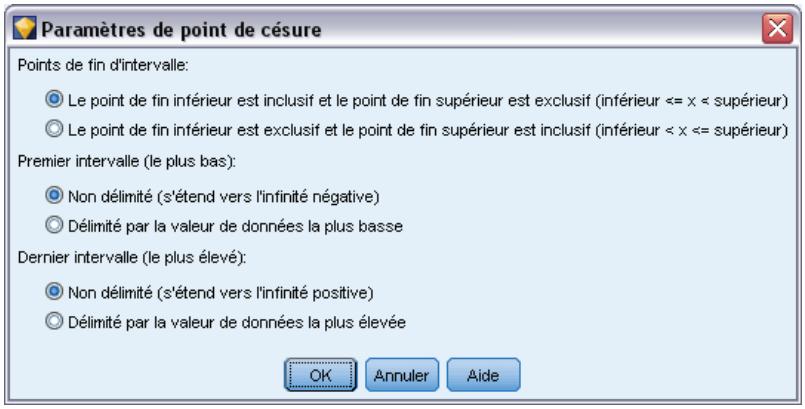
Champs de pré-intervalle pour améliorer les performances avec les ensembles de données volumineux. Indique s'il convient de procéder à un prétraitement pour simplifier la création d'intervalles optimale. Ceci permet de regrouper les valeurs d'échelle en un grand nombre d'intervalles en utilisant une méthode de création d'intervalles simple et non supervisée. Elle représente en outre les valeurs au sein de chaque intervalle par la moyenne et ajuste la pondération d'observation en conséquence avant de passer à la création d'intervalles supervisée. En pratique, cette méthode perd un certain degré de précision mais gagne en vitesse d'exécution. Elle est donc recommandée pour les grands ensembles de données. Vous pouvez également indiquer le nombre maximal d'intervalles avec lesquels doit se terminer toute variable après le prétraitement une fois cette option utilisée.

Fusionner les intervalles comportant relativement peu d'observations avec un voisinage plus grand. Si cette option est activée, indique qu'un intervalle est fusionné si le rapport de sa taille (nombre d'observations) avec celle d'un intervalle voisin est inférieur au seuil spécifié. Notez que des seuils plus élevés sont susceptibles d'entraîner une fusion plus importante.

Paramètres de point de césure

La boîte de dialogue Paramètres de point de césure vous permet de choisir des options avancées pour l'algorithme de création d'intervalles optimale. Ces options indiquent à l'algorithme comment calculer les intervalles à l'aide du champ cible.

Figure 4-70
Paramètres de point de césure pour la création d'intervalles optimale



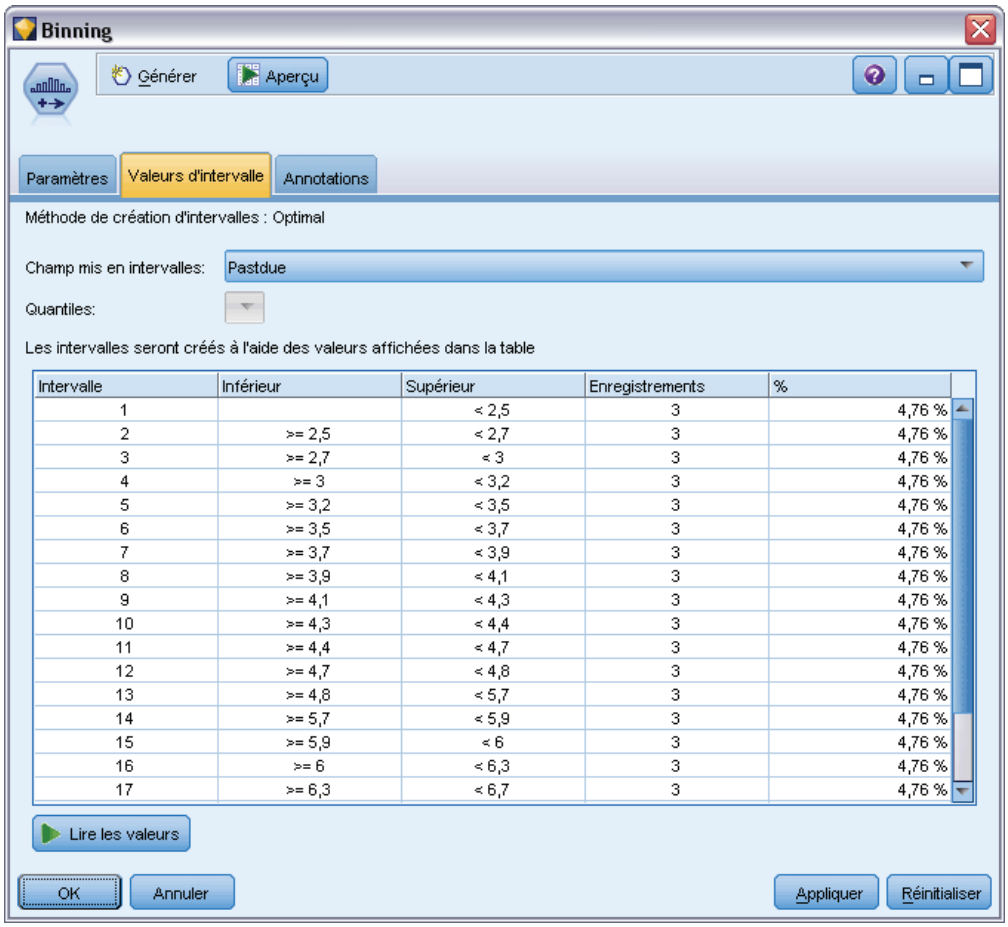
Points de fin d'intervalle. Vous pouvez indiquer si les points de fin inférieur ou supérieur doivent être inclusifs (inférieur <= x) ou exclusifs (inférieur < x).

Premiers et derniers intervalles. Pour le premier et le dernier intervalle, vous pouvez indiquer s'ils doivent être non délimités (tendant vers l'infini positif ou négatif) ou délimités par les points de données inférieur ou supérieur.

Prévisualisation des intervalles générés

L'onglet Valeurs d'intervalle du noeud Discrétiser permet de visualiser les seuils des intervalles générés. Avec le menu Générer, vous pouvez également générer un noeud Calculer qui peut être utilisé pour appliquer ces seuils d'un ensemble de données à l'autre.

Figure 4-71
Boîte de dialogue du noeud Discrétiser, onglet Valeurs d'intervalle



Intervalle	Inférieur	Supérieur	Enregistrements	%
1		< 2,5	3	4,76 %
2	>= 2,5	< 2,7	3	4,76 %
3	>= 2,7	< 3	3	4,76 %
4	>= 3	< 3,2	3	4,76 %
5	>= 3,2	< 3,5	3	4,76 %
6	>= 3,5	< 3,7	3	4,76 %
7	>= 3,7	< 3,9	3	4,76 %
8	>= 3,9	< 4,1	3	4,76 %
9	>= 4,1	< 4,3	3	4,76 %
10	>= 4,3	< 4,4	3	4,76 %
11	>= 4,4	< 4,7	3	4,76 %
12	>= 4,7	< 4,8	3	4,76 %
13	>= 4,8	< 5,7	3	4,76 %
14	>= 5,7	< 5,9	3	4,76 %
15	>= 5,9	< 6	3	4,76 %
16	>= 6	< 6,3	3	4,76 %
17	>= 6,3	< 6,7	3	4,76 %

Champ mis en intervalles. Dans la liste déroulante, sélectionnez le champ à afficher. A des fins de clarté, les noms de champ affichés reprennent le nom du champ d'origine.

Quantiles. Dans la liste déroulante, sélectionnez le quantile, tel que 10 ou 100, à afficher. Cette option est disponible uniquement lorsque les intervalles ont été générés à l'aide de la méthode des quantiles (effectifs égaux ou somme égale).

Seuils des intervalles. Les valeurs de seuil sont affichées ici pour chaque intervalle généré, avec le nombre d'enregistrements qui correspondent à chaque intervalle. Pour la méthode de création d'intervalles optimale uniquement, le nombre d'enregistrements dans chaque intervalle est présenté comme un pourcentage du total. Il est impossible d'appliquer des seuils lorsque la méthode de création d'intervalles par rang est utilisée.

Lire les valeurs. Lit les valeurs mises en intervalles de l'ensemble de données. Notez que les seuils sont également remplacés dès que de nouvelles données passent dans le flux.

Génération d'un noeud Calculer

Vous pouvez utiliser le menu Générer pour créer un noeud Calculer fondé sur les seuils actuels. Cela est utile lors de l'application de seuils d'intervalle établis d'un ensemble de données à un autre. Par ailleurs, si les points de séparation sont connus, l'opération Calculer est plus efficace (c'est-à-dire plus rapide) que l'opération Discrétiser dans le cas des ensembles de données volumineux.

Noeud Analyse RFM

Le noeud Analyse RFM (Récence, Effectif, Monétaire) permet de déterminer de façon quantitative les clients susceptibles d'être les meilleurs par l'étude de leur dernier achat (récence), l'effectif de leurs achats (effectif), et la somme dépensée lors de toutes les transactions (monétaire).

Le raisonnement derrière l'analyse RFM est que les clients qui achètent un produit ou un service une fois sont susceptibles de l'acheter à nouveau. Les données clients catégorisées se divisent en un certain nombre d'intervalles, avec les critères de création d'intervalles ajustés selon les besoins. Dans chacun des intervalles, un score est attribué aux clients. Ces scores sont ensuite combinés pour offrir un score RFM global. Ce score est une représentation de l'appartenance du client aux intervalles créés pour chacun des paramètres RFM. Ces données mises en intervalles peuvent s'avérer suffisantes pour vos besoins, par exemple, en identifiant les clients importants les plus fidèles. Elles peuvent être également transmises dans un flux pour une modélisation et une analyse plus approfondies.

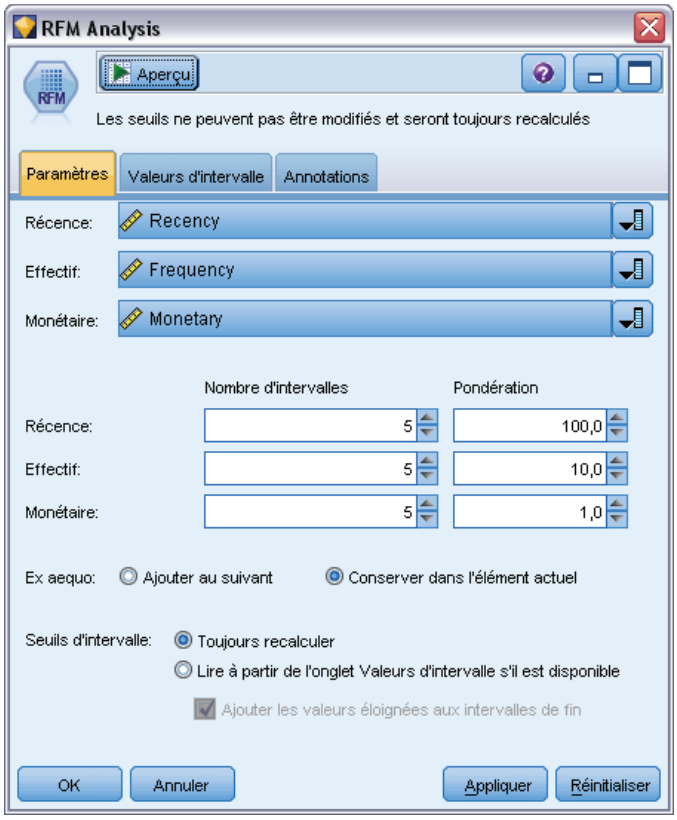
Remarque : bien que la capacité à analyser et à classer les scores RFM est un outil pratique, vous devez cependant garder à l'esprit certains facteurs lors de son utilisation. Il peut être tentant de cibler les clients avec les meilleurs classements. Toutefois, une sur-sollicitation de ces clients peut conduire à un certain ressentiment et une baisse effective de l'activité commerciale continue. Cela vaut également la peine de garder à l'esprit que les clients avec des scores bas ne doivent pas être négligés mais plutôt encouragés pour qu'ils deviennent de meilleurs clients. Inversement, des scores élevés seuls ne reflètent pas forcément une bonne perspective de ventes, selon le marché. Par exemple, un client dans l'intervalle 5 pour la récence, indiquant qu'il a effectué des achats très récemment, peut ne pas être le meilleur client cible pour une personne vendant des produits coûteux plus durables tels que des voitures ou des télévisions.

Remarque : Selon le mode de stockage de vos données, vous devrez peut être faire précéder le noeud Analyse RFM par un noeud Agréger RFM pour transformer les données en un format utilisable. Par exemple, les données d'entrée doivent être au format client avec une seule ligne par client. Si les données des clients sont au format transactionnel, utilisez un nœud Agréger RFM en amont pour calculer les champs Récence, Effectif et Montant. Pour plus d'informations, reportez-vous à la section Noeud Agréger RFM dans le chapitre 3 sur p. 86.

Les noeuds Agréger RFM et Analyse RFM de IBM® SPSS® Modeler sont configurés pour utiliser la création d'intervalles indépendants ; en d'autres termes, ils classent et espacent les données sur chaque mesure de valeur de proximité dans le temps, d'effectif et de valeur monétaire, sans tenir compte de leur valeur ni des deux autres mesures.

Paramètres du noeud Analyse RFM

Figure 4-72
Définition des options de Analyse RFM



Récence. A l'aide du sélecteur de champs (bouton à droite de la zone de texte), sélectionnez le champ Récence. Il peut s'agir d'une date, d'un horodatage ou d'un simple nombre. Remarque : lorsqu'une date ou un horodatage représente la date de la transaction la plus récente, la valeur la plus élevée est considérée comme étant la plus récente. Là où un nombre est indiqué, il représente le temps écoulé depuis la transaction la plus récente et la valeur la plus basse est considérée comme la plus récente.

Remarque : Si le noeud Analyse RFM est précédé dans le flux par un noeud Agréger RFM, les champs Récence, Effectif et Monétaire générés par le noeud Agréger RFM doivent être sélectionnés comme entrées dans le noeud Analyse RFM.

Effectif. A l'aide du sélecteur de champs, sélectionnez le champ Effectif à utiliser.

Monétaire. A l'aide du sélecteur de champs, sélectionnez le champ Monétaire à utiliser.

Nombre d'intervalles. Pour chacun des trois types de sorties, sélectionnez le nombre d'intervalles à créer. La valeur par défaut est 5.

Remarque : Le nombre minimum d'intervalles est 2, et le maximum est 9.

Pondération. Par défaut, la plus haute importance lors du calcul des scores est accordée aux données de récence, suivies de l'effectif, puis du montant. Si besoin est, vous pouvez modifier la pondération affectant un ou plusieurs de ces éléments pour changer celui qui se voit accorder la plus haute importance.

Le score RFM est calculé comme suit : (Score de récence x pondération de récence) + (score d'effectif x pondération d'effectif) + (score du montant x pondération du montant).

Ex aequo. Indiquez la manière dont les scores (ex æquo) identiques doivent être mis en intervalles. Les options sont les suivantes :

- **Ajouter au suivant.** Sélectionnez cette option pour déplacer les valeurs ex aequo vers l'intervalle supérieur suivant.
- **Conserver dans l'élément actuel.** Conserve les valeurs ex aequo dans l'intervalle (inférieur) actuel. Cette méthode peut générer un nombre inférieur d'intervalles. (Il s'agit de la valeur par défaut).

Seuils des intervalles. Indiquez si les scores de RFM et les affectations d'intervalles sont toujours recalculés lors de l'exécution du noeud, ou s'ils sont uniquement calculés selon les besoins (par exemple, lors de l'ajout de données). Si vous sélectionnez l'option Read from Bin Values tab if available (Lire dans l'onglet Valeurs d'intervalle si disponible), vous pouvez éditer les points de césures supérieurs et inférieurs pour les différents intervalles dans l'onglet Valeurs d'intervalle.

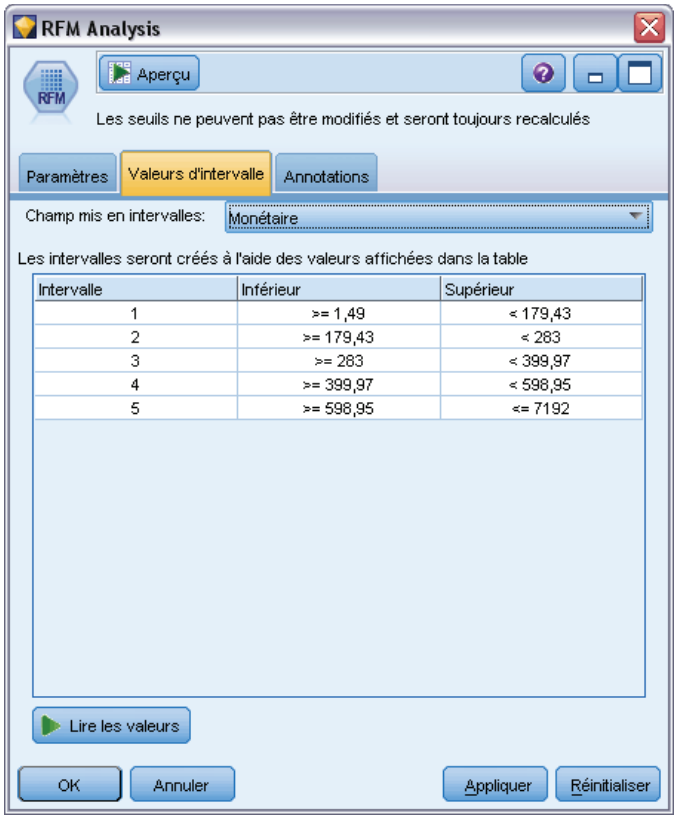
Une fois exécuté, le noeud Analyse RFM met en intervalles les champs Récence, Effectif et Monétaire bruts et ajoute les champs suivants à l'ensemble de données :

- Score de récence. Un classement (valeur d'intervalle) pour la récence
- Score de fréquence. Un classement (valeur d'intervalle) pour l'effectif
- Score monétaire. Un classement (valeur d'intervalle) pour Monétaire
- Score RFM. Le total pondéré des scores de récence, effectif et monétaire.

Ajouter des valeurs éloignées aux intervalles de fin. Si vous cochez cette case, les enregistrements qui figurent au-dessous de l'intervalle inférieur sont ajoutés à l'intervalle inférieur, et ceux au-dessus de l'intervalle supérieur sont ajoutés à l'intervalle le plus grand —sinon, une valeur nulle leur est attribuée. Cette case n'est disponible que si vous sélectionnez Lire dans l'onglet Valeurs d'intervalle si disponible.

Mise en intervalle du noeud Analyse RFM

Figure 4-73
Définition des valeurs de mise en intervalles du noeud Analyse RFM



L'onglet Valeurs d'intervalle permet d'afficher, et dans certains cas, modifier les seuils des intervalles générés.

Remarque : Vous ne pouvez modifier les valeurs dans cet onglet que si vous sélectionnez l'option Lire dans l'onglet Valeurs d'intervalle si disponible dans l'onglet Paramètres.

Champ mis en intervalles. Dans la liste déroulante, sélectionnez un champ pour la séparation en intervalles. Les valeurs disponibles sont celles sélectionnées dans l'onglet Paramètres.

Tableau des valeurs d'intervalle. Les valeurs de seuil de chaque intervalle généré sont affichées ici. Si vous sélectionnez l'option Lire dans l'onglet Valeurs d'intervalle si disponible dans l'onglet Paramètres, vous pouvez modifier les points de césure pour chaque intervalle en double-cliquant sur la cellule pertinente.

Lire les valeurs. Lit les valeurs mises en intervalles à partir de l'ensemble de données et renseigne le tableau de valeurs d'intervalle. Remarque : si vous sélectionnez Toujours recalculer dans l'onglet Paramètres, les seuils d'intervalle seront écrasés lors de l'exécution des nouvelles données via le flux.

Noeud Partitionner

Les noeuds Partitionner sont utilisés pour générer un champ de partition qui sépare les données en sous-ensembles ou en échantillons distincts pour les phases d'apprentissage, de test et de validation de la création de modèles. L'utilisation d'un échantillon pour la génération du modèle et d'un échantillon distinct pour le tester vous permet d'avoir une bonne indication de la manière dont le modèle peut se généraliser à des ensembles de données plus importants, semblables aux données actuelles.

Le noeud Partitionner génère un champ nominal dont le rôle est configuré sur Partitionner. Si vos données comportent déjà un champ adapté, vous pouvez également le désigner en tant que partition à l'aide d'un noeud Typer. Dans ce cas, vous n'avez pas besoin d'un noeud Partitionner distinct. Tout champ nominal instancié comportant deux ou trois valeurs peut être utilisé en tant que partition à l'exception des champs booléens. Pour plus d'informations, reportez-vous à la section Définition du rôle du champ sur p. 150.

Vous pouvez définir plusieurs champs de partition dans un flux, mais vous devrez alors sélectionner un champ de partition unique dans l'onglet Champs de chaque noeud de modélisation utilisant la partition. (Dans le cas d'une seule partition, cette partition est automatiquement utilisée lorsque la fonction de partition est activée.)

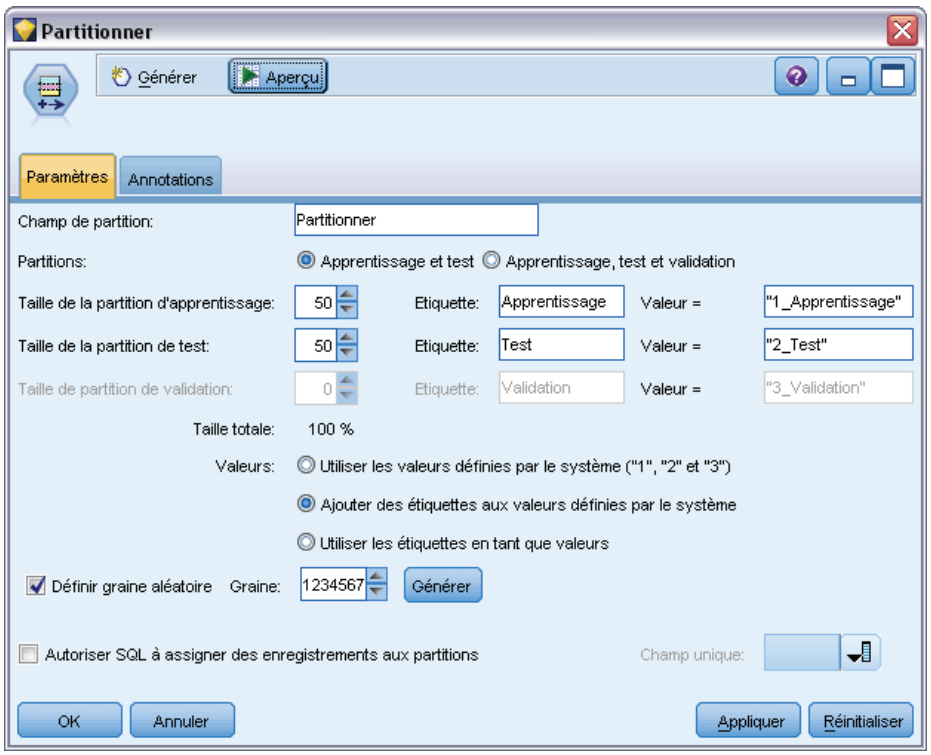
Activation des partitions. Pour utiliser la partition dans une analyse, vous devez l'activer dans l'onglet Options de modèle du noeud de création de modèle ou d'analyse approprié. Désélectionnez cette option pour pouvoir désactiver la partition sans supprimer le champ.

Pour créer un champ de partition en fonction de certains critères, tels qu'un intervalle de date ou un emplacement, vous pouvez également utiliser un noeud Calculer. Pour plus d'informations, reportez-vous à la section Noeud Calculer sur p. 166.

Exemple : Lors de la création d'un flux RFM pour identifier les clients récents qui ont réagi favorablement aux précédentes campagnes de marketing, le service marketing d'une société de ventes utilise un noeud Partitionner pour diviser les données en partitions de formation et de test.

Options du noeud Partitionner

Figure 4-74
Boîte de dialogue du noeud Partitionner, onglet Paramètres



Champ de partition. Indique le nom du champ créé par le noeud.

Partitions. Vous pouvez séparer les données en deux (apprentissage et test) ou trois (apprentissage, test et validation) échantillons.

- **Apprentissage et test.** Partitionne les données en deux échantillons et vous permet de former le modèle avec un échantillon et de le tester avec l'autre.
- **Apprentissage, test et validation.** Partitionne les données en trois échantillons, et vous permet de former le modèle avec un échantillon, de le tester et de l'affiner avec le deuxième et de valider les résultats à l'aide du troisième. La taille de chaque partition s'en trouve réduite, ce qui peut s'avérer très pratique lors de l'utilisation d'un ensemble de données très volumineux.

Taille de la partition. Indique la taille relative de chaque partition. Si la somme des tailles de partition est inférieure à 100 %, les enregistrements non inclus dans une partition sont ignorés. Par exemple, un utilisateur dispose de 10 millions d'enregistrements et de tailles de partition d'apprentissage de 5 % et de test de 10 %. Une fois le noeud exécuté, environ 500 000 enregistrements d'apprentissage et un million d'enregistrements de test doivent exister, le reste ayant été ignoré.

Valeurs : Indique les valeurs utilisées pour représenter chaque échantillon de partition dans les données.

- **Utiliser les valeurs définies par le système ("1", "2" et "3").** Utilise un entier pour représenter chaque partition ; par exemple, tous les enregistrements appartenant à l'échantillon d'apprentissage ont la valeur 1 pour le champ de partition. Cela garantit la portabilité des données entre les différents paramètres régionaux, ainsi que la conservation de l'ordre de tri (de sorte que 1 représente toujours la partition d'apprentissage) si le champ de partition est à nouveau instancié ailleurs (par exemple, lors de la nouvelle lecture des données provenant d'une base de données). Cependant, les valeurs nécessitent une interprétation.
- **Ajouter des étiquettes aux valeurs définies par le système.** Combine l'entier avec une étiquette ; par exemple, les enregistrements de partition d'apprentissage ont la valeur *1_Apprentissage*. Lors de la consultation des données, vous pouvez ainsi déterminer à quoi correspondent les valeurs ; en outre, cela permet de préserver l'ordre de tri. Cependant, les valeurs sont propres à un paramètre régional donné.
- **Utiliser les étiquettes en tant que valeurs.** Utilise l'étiquette sans entier ; par exemple, *Apprentissage*. Cela permet d'indiquer les valeurs en modifiant les étiquettes. Cependant, les données sont alors régionales et la réinstanciation d'une colonne de partition trie les valeurs dans leur ordre de tri naturel, qui ne correspond pas forcément à leur ordre «sémantique».

Définir graine aléatoire. Lors de l'échantillonnage ou du partitionnement d'enregistrements en fonction d'un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée ou cliquez sur le bouton Générer pour générer automatiquement une valeur aléatoire. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Lorsque vous utilisez l'option Définir graine aléatoire avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la graine aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. Pour plus d'informations, reportez-vous à la section Noeud Trier dans le chapitre 3 sur p. 88.

Activez SQL pour affecter des enregistrements à des partitions. (Uniquement pour des bases de données de niveau 1) Cochez cette case pour utiliser les répercussions SQL afin d'affecter des enregistrements à des partitions. À partir de la liste déroulante du champ Unique, sélectionnez un champ ayant des valeurs uniques (tel qu'un champ ID) pour vous assurer que les enregistrements sont affectés de manière aléatoire mais répétitive.

Les niveaux de base de données sont expliqués dans la description du noeud source de la base de données. Pour plus d'informations, reportez-vous à la section Nœud Source de base de données dans le chapitre 2 sur p. 15.

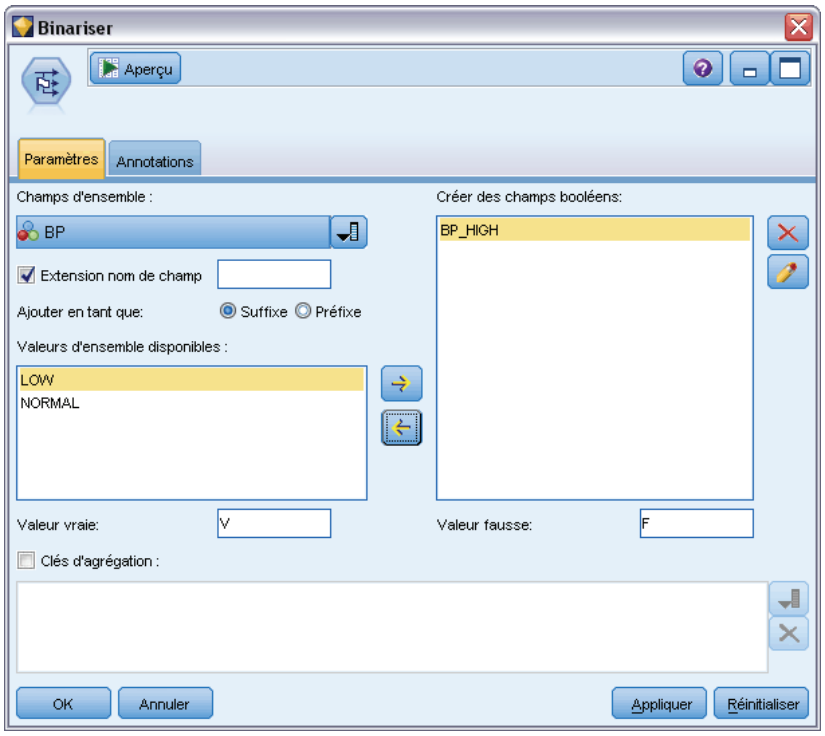
Génération de noeuds Sélectionner

Le menu Générer du noeud Partitionner permet de générer automatiquement un noeud Sélectionner pour chaque partition. Par exemple, vous pouvez sélectionner tous les enregistrements de la partition d'apprentissage afin d'obtenir une évaluation ou une analyse plus poussées avec cette seule partition.

Noeud Binariser

Le noeud Binariser est utilisé pour calculer des champs booléens en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux. Par exemple, votre ensemble de données peut contenir un champ nominal, *TA* (tension artérielle), ainsi que les valeurs *Elevée*, *Normale* et *Faible*. Pour faciliter la manipulation des données, vous pouvez créer un champ booléen pour la tension artérielle élevée ; celui-ci indique alors si le patient a une tension artérielle élevée.

Figure 4-75
Création d'un champ booléen pour une pression artérielle élevée



Paramétrage des options du noeud Binariser

Champs d'ensemble. Répertorie tous les champs de données ayant un niveau de mesure *Nominal* (ensemble). Sélectionnez un champ dans la liste pour afficher les valeurs de l'ensemble. Vous pouvez choisir l'une de ces valeurs pour créer un champ booléen. Pour que vous puissiez voir les champs nominaux disponibles (et leurs valeurs), les données doivent d'abord être entièrement instanciées à l'aide du noeud Typer ou source en amont. Pour plus d'informations, reportez-vous à la section Noeud Typer sur p. 136.

Extension nom de champ. Sélectionnez cette option pour permettre aux commandes de spécifier une extension qui sera ajoutée au nouveau champ booléen en tant que suffixe ou préfixe. Par défaut, les nouveaux noms de champ sont automatiquement créés en combinant le nom de champ d'origine et la valeur du champ afin d'obtenir une étiquette de type *Nomchamp_valeurchamp*.

Valeurs d'ensemble disponibles. Les valeurs de l'ensemble sélectionné plus haut apparaissent ici. Sélectionnez les valeurs pour lesquelles générer des booléens. Par exemple, si les valeurs d'un champ appelé *tension_artérielle* sont *Elevée*, *Moyenne* et *Faible*, vous pouvez sélectionner *Elevée* et l'ajouter à la liste sur la droite. Cette opération entraîne la création d'un champ comportant un booléen pour les enregistrements dotés d'une valeur indiquant une tension artérielle élevée.

Créer des champs booléens. Les nouveaux champs booléens sont répertoriés ici. Vous pouvez spécifier des options déterminant l'attribution du nom aux nouveaux champs à l'aide des contrôles d'extension de nom de champ.

Valeur vraie (true). Indiquez la valeur true (vrai) utilisée par le noeud lors de la définition d'un booléen. Par défaut, cette valeur est T.

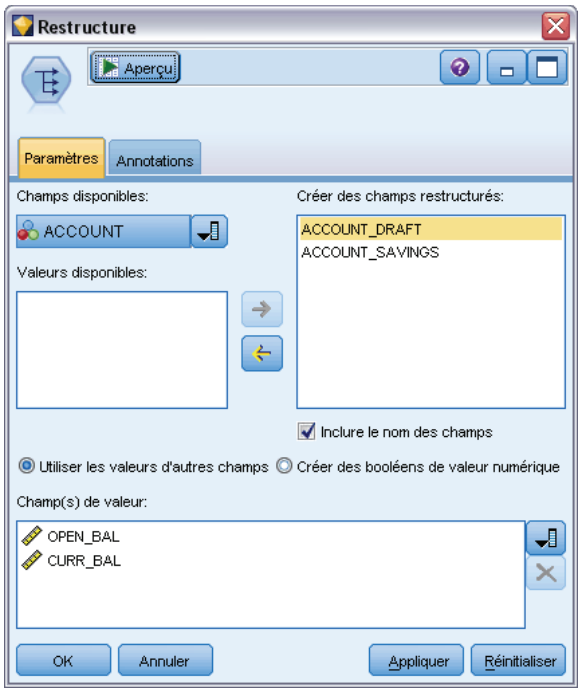
Valeur fausse (false). Indiquez la valeur false (faux) utilisée par le noeud lors de la définition d'un booléen. Par défaut, cette valeur est F.

Clés d'agrégation. Sélectionnez cette option pour grouper les enregistrements sur la base des champs-clés spécifiés plus bas. Lorsque l'option Clés d'agrégation est sélectionnée, tous les champs booléens d'un groupe sont activés si l'*un* des enregistrements a été défini comme true (vrai). Utilisez le sélecteur de champs pour choisir les champs-clés à utiliser pour agréger les enregistrements.

Noeud Restructurer

Le noeud Restructurer peut être utilisé pour générer plusieurs champs en fonction des valeurs d'un champ nominal ou d'un champ booléen. Les champs nouvellement générés peuvent contenir des valeurs issues d'un autre champ ou de champs booléens numériques (0 et 1). La fonctionnalité de ce noeud est semblable à celle du noeud Binariser. Toutefois, il offre une plus grande souplesse d'utilisation. Il permet de créer des champs de tout type (y compris les booléens numériques), à l'aide des valeurs issues d'un autre champ. Vous pouvez ainsi effectuer une agrégation ou d'autres manipulations avec d'autres noeuds situés en aval. (Grâce au noeud Binariser, vous pouvez agréger des champs en une seule étape ; cela peut s'avérer utile lorsque vous créez des champs booléens.)

Figure 4-76
Génération de champs restructurés pour le champ d'ensemble Compte



Par exemple, l'ensemble de données suivant contient un champ nominal *Compte*, et les valeurs *Epargne* et *Courant*. Le solde d'ouverture et le solde actuel sont enregistrés pour chaque compte. Certains clients possèdent plusieurs comptes de chaque type. Supposons que vous souhaitiez savoir si chaque client possède un type de compte particulier et, si tel est le cas, la somme figurant sur chaque type de compte. Utilisez le noeud Restructurer pour générer un champ pour chacune des valeurs du champ *Compte* et sélectionnez *Solde_actuel* comme valeur. Chaque nouveau champ est renseigné par le solde actuel de l'enregistrement concerné.

Table 4-2
Exemple de données avant restructuration

IDclient	Compte	Solde_ ouverture	Solde_actuel
12701	Mode brouillon	1000	1005.32
12702	Epargne	100	144.51
12703	Epargne	300	321.20
12703	Epargne	150	204.51
12703	Mode brouillon	1200	586.32

Table 4-3
Exemple de données après restructuration

IDclient	Compte	Solde_ ouverture	Solde_actuel	Compte_Courant_ Solde_actuel	Compte_Epargne _Solde_actuel
12701	Mode brouillon	1000	1005.32	1005.32	$null$
12702	Epargne	100	144.51	$null$	144.51
12703	Epargne	300	321.20	$null$	321.20
12703	Epargne	150	204.51	$null$	204.51
12703	Mode brouillon	1200	586.32	586.32	$null$

Utilisation du noeud Restructurer avec le noeud Agréger

Dans de nombreux cas, vous pouvez combiner le noeud Restructurer avec le noeud Agréger. Dans l'exemple précédent, un client (doté de l'ID 12703) possède trois comptes. Vous pouvez utiliser un noeud Agréger pour calculer le solde total de chaque type de compte. Le champ-clé est *IDclient* et les champs d'agrégation sont les nouveaux champs restructurés, *Compte_Courant_Solde_actuel* et *Compte_Epargne_Solde_actuel*. Le tableau ci-dessous présente les résultats obtenus.

Table 4-4
Exemple de données après restructuration et agrégation

IDclient	Effectif	Compte_Courant_Solde_ actuel_Somme	Compte_Epargne_Solde_ actuel_Somme
12701	1	1005.32	$null$
12702	1	$null$	144.51
12703	3	586.32	525.71

Paramétrage des options du noeud Restructurer

Champs disponibles. Répertorie tous les champs de données ayant un niveau de mesure *Nominal* (ensemble) ou *Booléen*. Sélectionnez un champ dans la liste pour afficher les valeurs de l'ensemble ou du booléen, puis choisissez les valeurs souhaitées pour créer les champs restructurés. Pour que vous puissiez voir les champs disponibles (et leurs valeurs), les données doivent d'abord être entièrement instanciées à l'aide du noeud Typer ou source en amont. Pour plus d'informations, reportez-vous à la section Noeud Typer sur p. 136.

Valeurs disponibles. Les valeurs de l'ensemble sélectionné plus haut apparaissent ici. Sélectionnez les valeurs pour lesquelles générer des champs restructurés. Par exemple, si les valeurs d'un champ appelé *Tension artérielle* sont *Elevée*, *Moyenne* et *Faible*, vous pouvez sélectionner *Elevée* et l'ajouter à la liste figurant sur la droite. Un champ est ainsi créé et renseigné à partir d'une valeur définie (voir ci-dessous) pour les enregistrements présentant la valeur *Elevée*.

Créer des champs restructurés. Les nouveaux champs restructurés sont répertoriés ici. Par défaut, les nouveaux noms de champ sont automatiquement créés en combinant le nom de champ d'origine et la valeur du champ afin d'obtenir une étiquette de type *Nomchamp_valeurchamp*.

Inclure le nom des champs. Désélectionnez cette option pour ne pas inclure le nom du champ d'origine comme préfixe dans les nouveaux noms de champ.

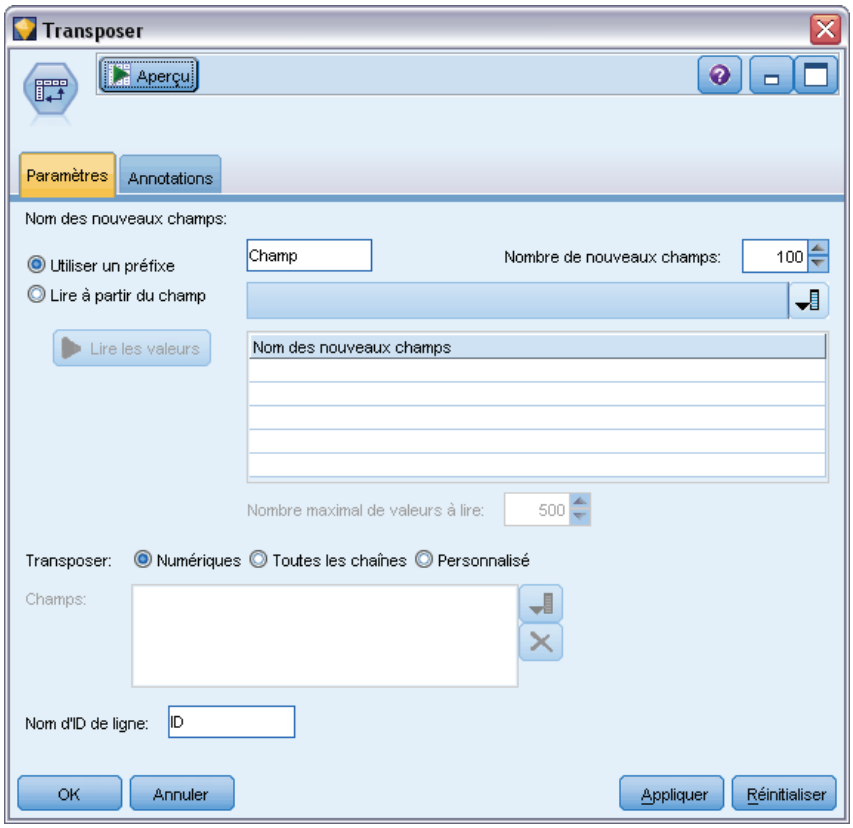
Utiliser les valeurs d'autres champs. Indiquez un ou plusieurs champs dont la valeur sera utilisée pour renseigner les champs restructurés. Utilisez pour cela le sélecteur de champs. Un champ est créé pour chaque champ sélectionné. Le nom du champ de valeur est ajouté au nom du champ restructuré ; par exemple *TA_Elevée_Age* ou *TA_Faible_Age*. Chaque nouveau champ hérite du type du champ de valeur d'origine.

Créer des booléens de valeur numérique. Sélectionnez cette option pour renseigner les nouveaux champs à l'aide de booléens de valeur numérique (0 pour false (faux) et 1 pour true (vrai)), et non à partir d'une valeur d'un autre champ.

Noeud Transposer

Par défaut, les colonnes correspondent aux champs, et les lignes aux enregistrements ou aux observations. Vous pouvez utiliser, si nécessaire, un nœud Transposer pour faire permuter les données des lignes et des colonnes afin que les champs deviennent des enregistrements et que les enregistrements deviennent des champs. Par exemple, si vous disposez de séries temporelles, où chaque série est une ligne et non une colonne, vous pouvez transposer les données avant de procéder à l'analyse.

Figure 4-77
Noeud Transposer, onglet Paramètres



Paramétrage des options du noeud Transposer

Nom des nouveaux champs

Les nouveaux noms de champ peuvent être générés automatiquement à partir d'un préfixe défini ou ils peuvent être lus à partir d'un champ existant dans les données.

Utiliser un préfixe. Cette option génère automatiquement de nouveaux noms de champ à partir du préfixe indiqué (*Champ1*, *Champ2*, etc.). Vous pouvez personnaliser le préfixe comme souhaité. Lorsque vous utilisez cette option, vous devez indiquer le nombre de champs à créer, quel que soit le nombre de lignes présentes dans les données d'origine. Par exemple, si l'option Nombre de nouveaux champs est paramétrée sur 100, toutes les données figurant au-delà des 100 premières lignes sont ignorées. Si les données d'origine contiennent moins de 100 lignes, certains champs prennent la valeur nulle. (Vous pouvez augmenter le nombre de champs selon vos besoins ; il convient toutefois d'éviter d'utiliser ce paramètre pour transposer un million d'enregistrements en un million de champs, ce qui produirait un résultat ingérable.)

Par exemple, supposons que vous disposez de données avec des séries en lignes et un champ distinct (colonne) pour chaque mois. Vous pouvez transposer ces données de telle manière que chaque série apparaisse dans un champ distinct, avec une ligne pour chaque mois.

Figure 4-78
Données d'origine avec séries en lignes

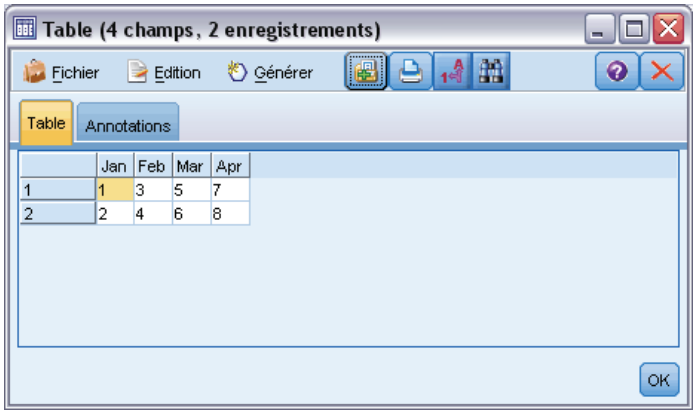
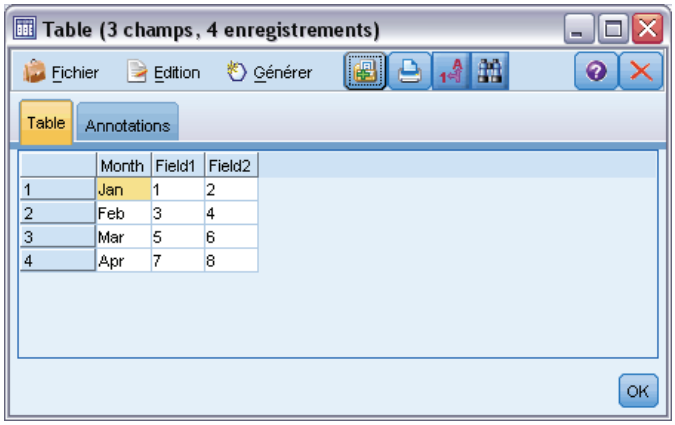


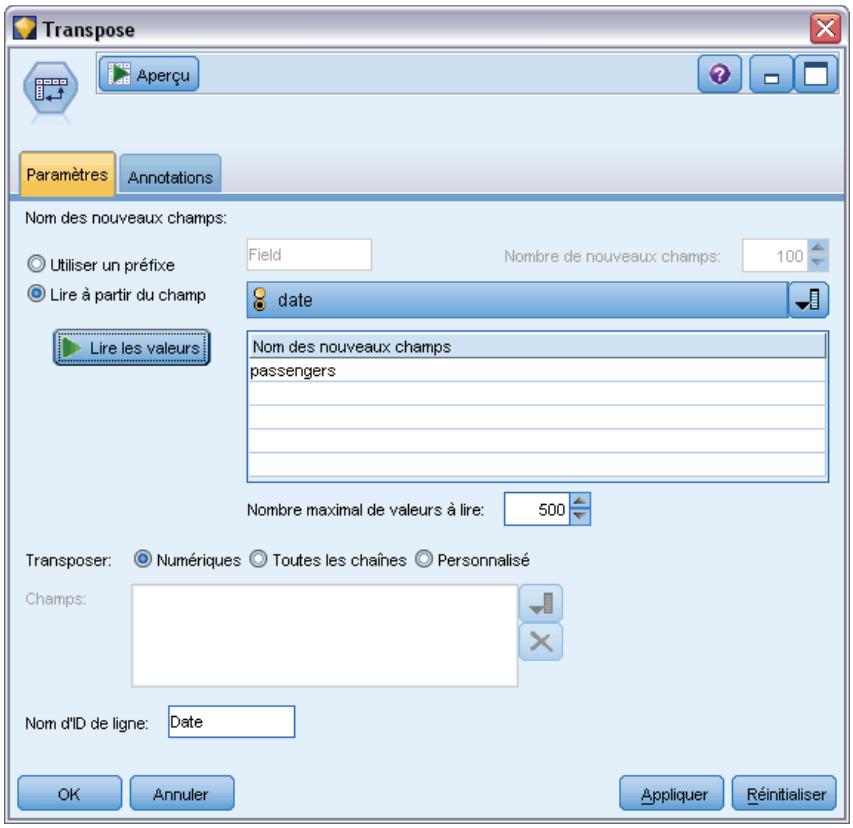
Figure 4-79
Données transposées avec séries en colonnes



Remarque : Pour obtenir les résultats présentés ici, la valeur 100 de l'option Nombre de nouveaux champs a été remplacée par 2 et la valeur Nom d'ID de ligne, ID, a été remplacée par Mois (voir ci-dessous).

Lire à partir du champ. Lit les noms de champ à partir d'un champ existant. Avec cette option, le nombre de nouveaux champs est déterminé par les données, jusqu'à atteindre la limite maximale indiquée. Chaque valeur du champ sélectionné devient un nouveau champ dans les données de sortie. Le champ sélectionné peut présenter n'importe quel type de stockage (entier, chaîne, date, heure, etc.), mais afin d'éviter des noms de fichier en double, chaque valeur du champ sélectionné doit être unique (en d'autres termes, le nombre de valeurs doit correspondre au nombre de lignes). Lorsque des noms de champ en double sont détectés, un avertissement apparaît.

Figure 4-80
Lecture des noms de champ à partir d'un champ existant



- **Lire les valeurs.** Si le champ sélectionné n'a pas été instancié, sélectionnez cette option pour renseigner la liste des nouveaux noms de champ. Si le champ a déjà été instancié, cette étape n'est pas nécessaire.
- **Nombre maximal de valeurs à lire.** Lors de la lecture des noms de champ à partir des données, une limite supérieure est définie afin d'éviter de créer un nombre de champs trop important. (Comme indiqué ci-dessus, la transposition d'un million d'enregistrements en un million de champs produirait un résultat ingérable.)

Par exemple, si la première colonne de données indique le nom de chaque série, vous pouvez utiliser ces valeurs en tant que noms de champ dans les données transposées.

Figure 4-81
Données d'origine avec séries sur une seule ligne

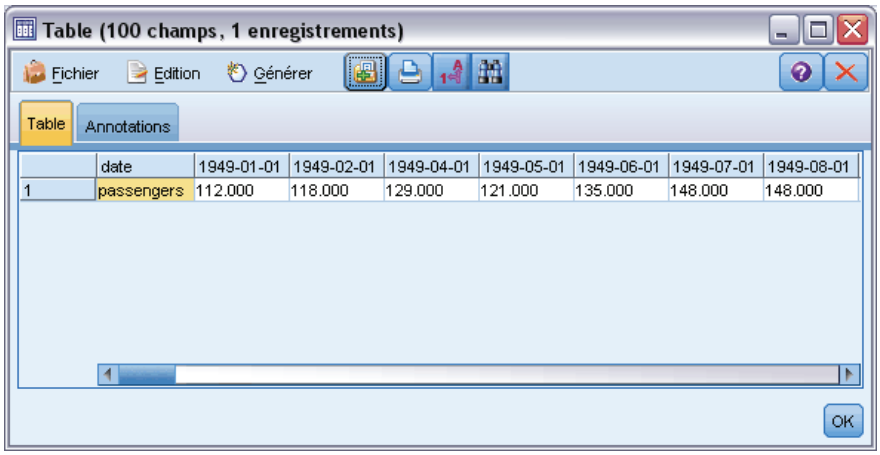
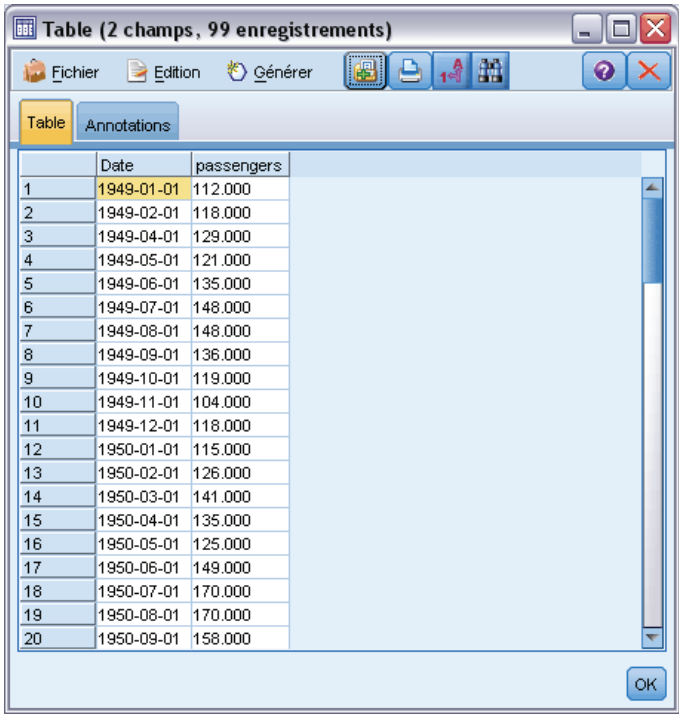


Figure 4-82
Données transposées avec séries en colonnes



Transposer. Par défaut, seuls les champs continus (intervalle numérique) sont transposés (stockage de type entier ou nombre réel). Si nécessaire, vous pouvez sélectionner un sous-ensemble de champs numériques ou transposer des champs de type chaîne. Toutefois, tous les champs transposés doivent comporter le même type de stockage (numérique ou chaîne, mais pas les deux) ; en effet, l'utilisation de champs d'entrée mixtes générerait des valeurs mixtes au sein de chaque colonne de sortie, ce qui irait à l'encontre de la règle selon laquelle toutes les valeurs d'un champ doivent être dotées du même type de stockage. Les autres types de stockage (date, heure, horodatage) ne peuvent pas être transposés.

- **Numériques.** Transpose tous les champs numériques (stockage de type entier ou nombre réel). Le nombre de lignes dans la sortie correspond au nombre de champs numériques dans les données d'origine.
- **Toutes les chaînes.** Transpose tous les champs de type chaîne.
- **Personnalisée.** Permet de sélectionner un sous-ensemble de champs numériques. Le nombre de ligne dans la sortie correspond au nombre de champs sélectionnés. *Remarque* : Cette option est disponible uniquement pour les champs numériques.

Nom d'ID de ligne. Indique le nom du champ d'ID de ligne créé par le noeud. Les valeurs de ce champ sont déterminées en fonction du nom des champs figurant dans les données d'origine.

Astuce : Lorsque vous transposez des séries temporelles de lignes en colonnes et que les données d'origine incluent une ligne, telle que Date, Mois ou Année, qui sert d'étiquetage de période à chaque mesure, veillez à lire ces étiquettes comme des noms de champ dans IBM® SPSS® Modeler (comme le montrent les exemples précédents, qui affichent le mois ou le jour comme noms de champ dans les données d'origine, respectivement) au lieu d'inclure l'étiquette dans la première ligne de données. Ainsi, vous éviterez de mélanger les étiquettes et les valeurs dans chaque colonne (ce qui obligerait les nombres à être lus comme des chaînes car les types de stockage ne peuvent pas être mélangés dans une colonne).

Noeud Intervalles de temps

Le noeud Intervalles de temps vous permet de définir des intervalles et de générer des étiquettes pour les séries temporelles à utiliser dans une modélisation Séries temporelles ou dans un noeud Tracé horaire pour l'estimation ou la prévision. Un ensemble complet d'intervalles de temps, allant des secondes aux années, est pris en charge. Par exemple, si vous disposez d'une série comportant des mesures quotidiennes commençant le 3 janvier 2005, vous pouvez étiqueter les enregistrements à partir de cette date. La deuxième ligne correspond alors au 4 janvier, et ainsi de suite. Vous pouvez également indiquer la périodicité (par exemple, cinq jours par semaine ou huit heures par jour).

Par ailleurs, vous pouvez préciser l'intervalle d'enregistrements à utiliser pour l'estimation. Vous pouvez choisir d'exclure ou non les enregistrements les plus récents de la série et de préciser ou non les ensembles de rétention. Cela vous permet de tester le modèle en retenant les enregistrements les plus récents dans les séries temporelles pour comparer leurs valeurs connues avec les valeurs estimées pour ces périodes.

Vous pouvez également préciser sur combien de périodes futures la prévision doit porter et indiquer les valeurs futures à utiliser pour les prévisions par les noeuds de modélisation Séries temporelles en aval.

Le noeud Intervalles de temps génère un champ *TimeLabel* dont le format correspond à la période et à l'intervalle indiqués, ainsi qu'un champ *TimeIndex* qui attribue un entier unique à chaque enregistrement. Plusieurs champs supplémentaires peuvent également être générés sur la base de la période ou de l'intervalle sélectionné (par exemple, les minutes ou les secondes associées à une mesure).

Vous pouvez étendre ou agréger des valeurs selon vos besoins pour vous assurer que les mesures sont espacées de manière uniforme. Les méthodes de modélisation des séries temporelles nécessitent l'utilisation d'un intervalle uniforme entre chaque mesure, chaque valeur manquante

étant signalée par des lignes vides. Si les données ne répondent pas à ces exigences, le noeud peut les transformer.

Commentaires

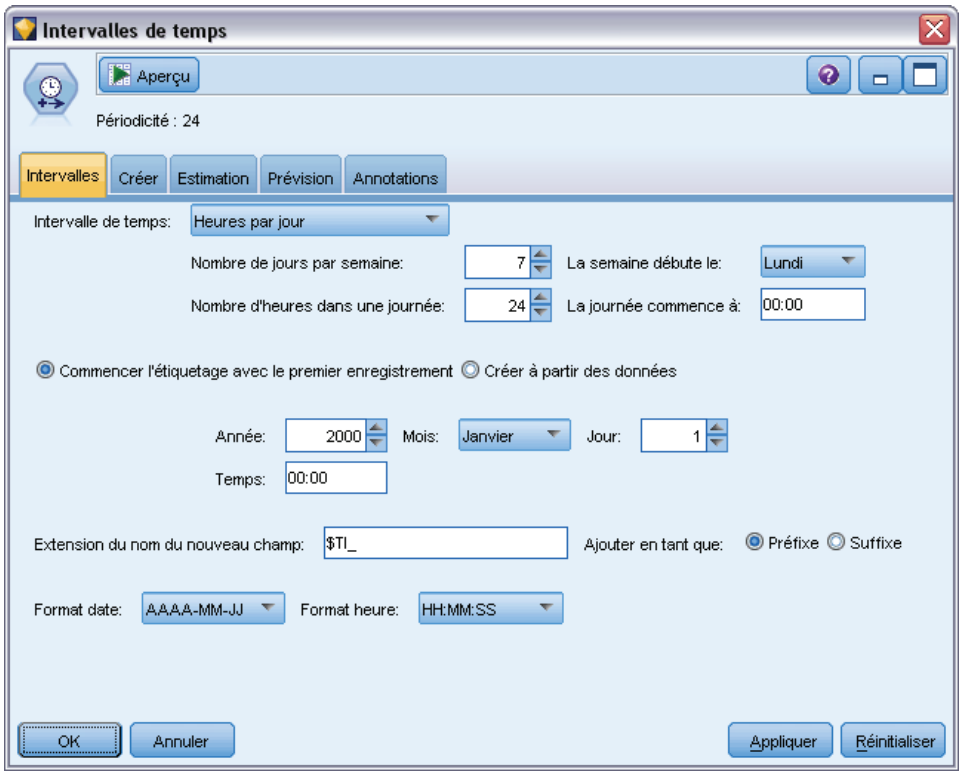
- Il se peut que les intervalles de temps ne correspondent pas au temps réel. Par exemple, une série basée sur une semaine de travail standard de cinq jours traite l'intervalle entre le vendredi et le lundi comme un jour unique.
- Le noeud Intervalle de temps suppose que chaque série se trouve dans un champ ou dans une colonne, avec une ligne par mesure. Si nécessaire, vous pouvez transposer vos données pour respecter cette condition. Pour plus d'informations, reportez-vous à la section Noeud Transposer sur p. 214.
- Pour les séries qui ne sont pas espacées de manière uniforme, vous pouvez indiquer un champ qui identifie la date ou l'heure de chaque mesure. Pour cela, notez que vous devez utiliser comme champ d'entrée un champ Date, Temps ou Horodatage au format adéquat. Si nécessaire, convertissez un champ existant (tel qu'un champ d'étiquette de chaîne) en ce format à l'aide d'un noeud Remplacer. Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer sur p. 180.
- Lorsque vous affichez les détails relatifs aux champs d'index et d'étiquette générés, il peut s'avérer utile d'activer l'affichage des étiquettes de valeur. Par exemple, lorsque vous affichez un tableau comportant des valeurs générées pour des données mensuelles, vous pouvez cliquer sur l'icône des étiquettes de valeur dans la barre d'outils ; les intitulés *Janvier, Février, Mars, etc.* remplacent alors les numéros *1, 2, 3, etc.*.

Figure 4-83
Icône des étiquettes de valeur

Définition d'intervalles de temps

L'onglet Intervalles permet de définir l'intervalle et la périodicité s'appliquant à la création ou à l'étiquetage de la série. Les paramètres disponibles dépendent de l'intervalle sélectionné. Par exemple, si vous sélectionnez Heures par jour, vous pouvez indiquer le nombre de jours par semaine, le jour qui débute la semaine, le nombre d'heures par jour et l'heure de début de chaque jour. Pour plus d'informations, reportez-vous à la section Intervalles pris en charge sur p. 228.

Figure 4-84
Paramètres d'intervalles de temps pour une série horaire



Etiquetage ou création d'une série

Vous pouvez étiqueter les enregistrements de manière consécutive ou créer la série sur la base d'un champ de date, d'horodatage ou d'heure spécifié.

- **Commencer l'étiquetage avec le premier enregistrement.** Indiquez la date de début et/ou l'heure qui servira d'étiquette à des enregistrements successifs. Si vous appliquez une étiquette en fonction de l'heure de la journée, par exemple, vous définissez la date et l'heure auxquelles la série débute, puis un seul enregistrement par heure qui suit. Hormis le fait d'ajouter des étiquettes, cette méthode ne modifie pas les données d'origine. En revanche, elle suppose que les enregistrements sont déjà espacés de manière uniforme, un intervalle constant séparant chaque mesure. Toute mesure manquante doit être indiquée par une ligne vide dans les données.
- **Créer à partir des données.** Pour les séries qui ne sont pas espacées de manière uniforme, vous pouvez indiquer un champ qui identifie la date ou l'heure de chaque mesure. Pour cela, notez que vous devez utiliser comme champ d'entrée un champ Date, Temps ou Horodatage au format adéquat. Par exemple, si vous avez un champ de type chaîne contenant des valeurs telles que *Jan 2000*, *Fév 2000*, etc., vous pouvez le convertir en champ de date à l'aide d'un noeud Remplacer. Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer sur p. 180. L'option Créer à partir des données transforme également les données de manière à ce qu'elles correspondent à l'intervalle spécifié. Pour ce faire, elle étend ou agrège les enregistrements selon les besoins (par exemple,

en cumulant plusieurs semaines en mois ou en remplaçant les enregistrements manquants par des blancs ou des valeurs extrapolées). Vous pouvez utiliser l'onglet Créer afin de spécifier les fonctions utilisées pour étendre ou agréger des enregistrements. Pour plus d'informations, reportez-vous à la section Options de création d'intervalles de temps sur p. 222.

Extension du nom du nouveau champ. Permet de définir un préfixe ou un suffixe appliqué à tous les champs générés par le noeud. Par exemple, si vous utilisez le préfixe *$TI_* par défaut, les champs créés par le noeud sont nommés *$TI_TimeIndex*, *$TI_TimeLabel*, etc.

Format date. Indique le format du champ *TimeLabel* créé par le noeud, en fonction de l'intervalle actuel. Les options disponibles dépendent de la sélection actuelle.

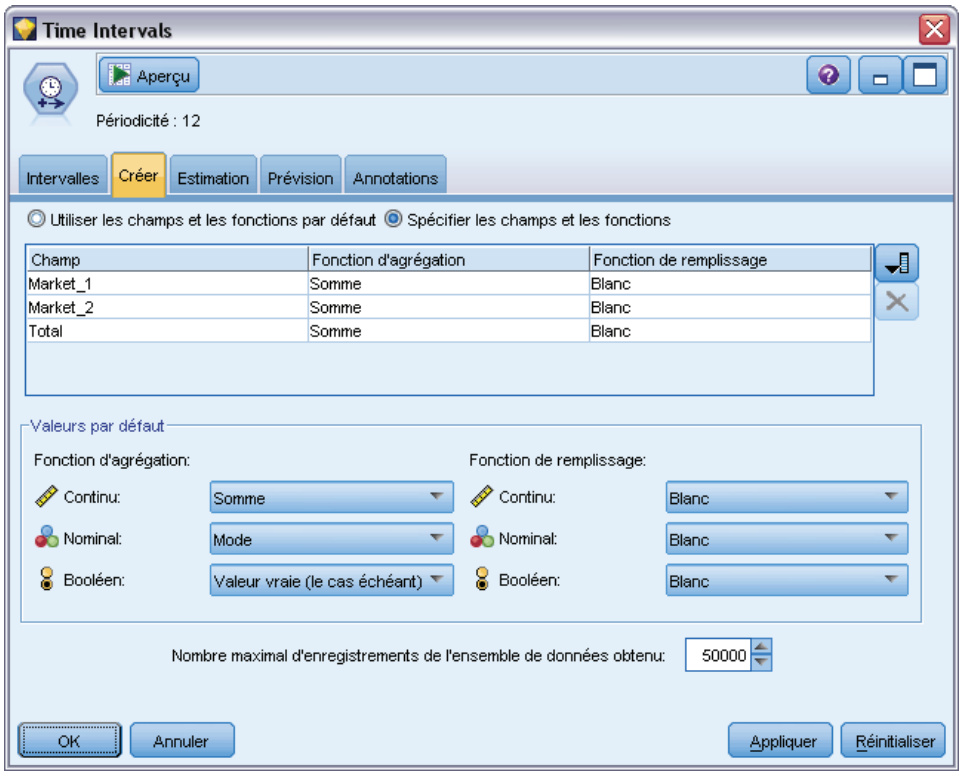
Format heure. Indique le format du champ *TimeLabel* créé par le noeud, en fonction de l'intervalle actuel. Les options disponibles dépendent de la sélection actuelle.

Options de création d'intervalles de temps

L'onglet Créer du noeud Intervalle de temps vous permet de spécifier les options d'agrégation et d'extension des champs afin qu'ils correspondent à l'intervalle indiqué. Ces paramètres s'appliquent uniquement lorsque l'option Créer à partir des données est sélectionnée dans l'onglet Intervalles. Par exemple, en présence d'un mélange de données hebdomadaires et mensuelles, vous pouvez agréger ou cumuler les valeurs hebdomadaires pour obtenir un intervalle mensuel uniforme. En outre, vous pouvez définir un intervalle hebdomadaire et étendre les séries en insérant des valeurs non renseignées pour toutes les semaines manquantes ou en extrapolant des valeurs manquantes à l'aide de la fonction d'extension spécifiée.

Lorsque vous étendez ou agrégez des données, tous les champs d'horodatage ou de date existants sont remplacés par les champs *TimeLabel* et *TimeIndex* générés et sont supprimés de la sortie. Les champs sans type sont également supprimés. Les champs qui concernent une durée sont conservés (un champ qui mesure la durée d'un appel plutôt que l'heure à laquelle l'appel a débuté, par exemple), tant qu'ils sont stockés en interne comme champs temporels et non comme champs d'horodatage. Pour plus d'informations, reportez-vous à la section Définition du stockage et du formatage des champs dans le chapitre 2 sur p. 32. Les autres champs sont agrégés en fonction des options spécifiées dans l'onglet Créer.

Figure 4-85
Noeud Intervalle de temps, onglet Créer



- **Utiliser les champs et les fonctions par défaut.** Indique que tous les champs doivent être agrégés ou étendus selon les besoins, à l'exception des champs de date, des champs d'horodatage et des champs sans type, comme mentionné ci-dessus. La fonction par défaut est appliquée selon le niveau de mesure (par exemple, les champs continus sont agrégés à l'aide de la moyenne, tandis que les champs nominaux utilisent le mode). Vous pouvez modifier les valeurs par défaut d'un ou de plusieurs niveaux de mesure dans la partie inférieure de la boîte de dialogue.
- **Spécifier les champs et les fonctions.** Permet d'indiquer les champs à étendre ou à agréger, ainsi que la fonction utilisée pour chaque champ. Tout champ non sélectionné est supprimé de la sortie. Utilisez les icônes de droite pour ajouter ou supprimer des champs dans le tableau, ou cliquez sur la cellule de la colonne appropriée pour modifier la fonction d'agrégation ou d'extension utilisée pour ce champ et remplacer la valeur par défaut. Les champs sans type sont exclus de la liste et ne peuvent pas être ajoutés au tableau.

Valeurs par défaut. Indique les fonctions d'agrégation et d'extension utilisées par défaut pour différents types de champ. Ces fonctions par défaut sont appliquées lorsque la fonction Utiliser les valeurs par défaut est sélectionnée ; elles sont également appliquées comme valeur par défaut initiale pour tous les nouveaux champs ajoutés au tableau. (La modification des fonctions par défaut n'a aucune incidence sur les paramètres existants dans le tableau mais s'applique à tous les champs ajoutés ultérieurement.)

Fonction d'agrégation. Les fonctions d'agrégation suivantes sont disponibles :

- **Continue.** Les fonctions disponibles pour les champs continus sont Moyenne, Somme, Mode, Min et Max.

- **Nominal :** Les options disponibles sont les suivantes : Mode, Premiers et Derniers. La fonction Premiers recherche la première valeur non nulle (dans un tri par date) dans le groupe d'agrégation, et la fonction Derniers la dernière valeur non nulle dans le groupe.
- **Booléen.** Les options disponibles sont Valeur vraie (le cas échéant), Mode, Premiers et Derniers.

Fonction d'extension. Les fonctions d'extension suivantes sont disponibles :

- **Continue.** Les options disponibles sont Blanc et Moyenne des points les plus récents (c'est-à-dire la moyenne des trois valeurs non nulles les plus récentes avant la période qui sera créée). S'il n'y a pas trois valeurs de ce type, la nouvelle valeur est non renseignée. Les valeurs récentes incluent uniquement les valeurs réelles ; une valeur étendue créée précédemment n'est pas prise en compte dans la recherche d'une valeur non nulle.
- **Nominal.** Blanc et Valeur la plus récente. Par "plus récente" l'on entend la valeur non nulle la plus récente avant la période qui sera créée. Là encore, seules les valeurs réelles sont prises en compte dans la recherche d'une valeur récente.
- **Booléen.** Les options disponibles sont Blanc, Vrai et Faux.

Nombre maximal d'enregistrements de l'ensemble de données obtenu. Indique une limite supérieure relative au nombre d'enregistrements créés (nombre qui deviendrait sinon trop important), particulièrement lorsque l'intervalle de temps est défini en secondes (délibérément ou non). Par exemple, une série comprenant seulement deux valeurs (1 Jan. 2000 et 1 Jan. 2001) génère 31 536 000 enregistrements si elle est exprimée en secondes (60 secondes x 60 minutes x 24 heures x 365 jours). Le système cesse le traitement et affiche un avertissement si le nombre maximal d'enregistrements défini est dépassé.

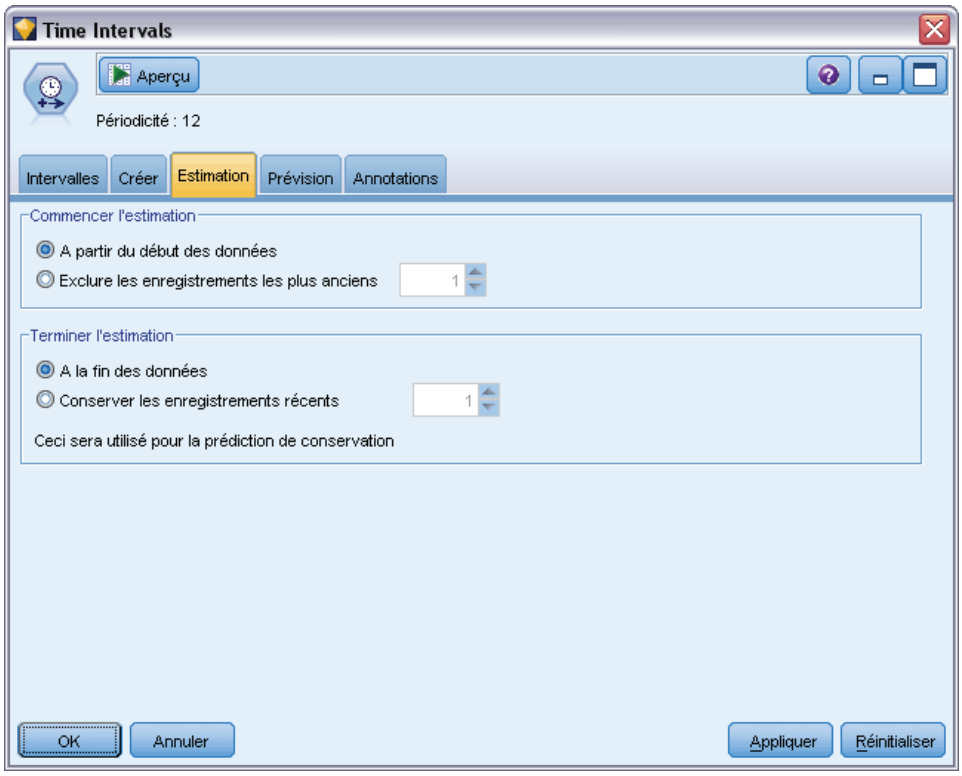
Champ Comptage

Lorsque vous agrégez ou étendez des valeurs, un champ *Comptage* est créé. Il indique le nombre d'enregistrements impliqués dans la détermination du nouvel enregistrement. Ainsi, si quatre valeurs hebdomadaires sont agrégées en un mois unique, le comptage est de 4. Pour un enregistrement étendu, le comptage est de 0. Le nom du champ est composé de l'intitulé *Comptage* auquel est ajouté le préfixe ou le suffixe défini dans l'onglet Intervalles.

Période d'estimation

L'onglet Estimation du noeud Intervalles de temps vous permet d'indiquer l'intervalle d'enregistrements utilisé dans l'estimation du modèle, ainsi que les ensembles de rétention. Ces paramètres peuvent être remplacés dans les noeuds de modélisation en aval si nécessaire. Cependant, il peut être plus pratique de les définir ici que de les définir pour chaque noeud distinct.

Figure 4-86
Noeud Intervalles de temps, onglet Estimation



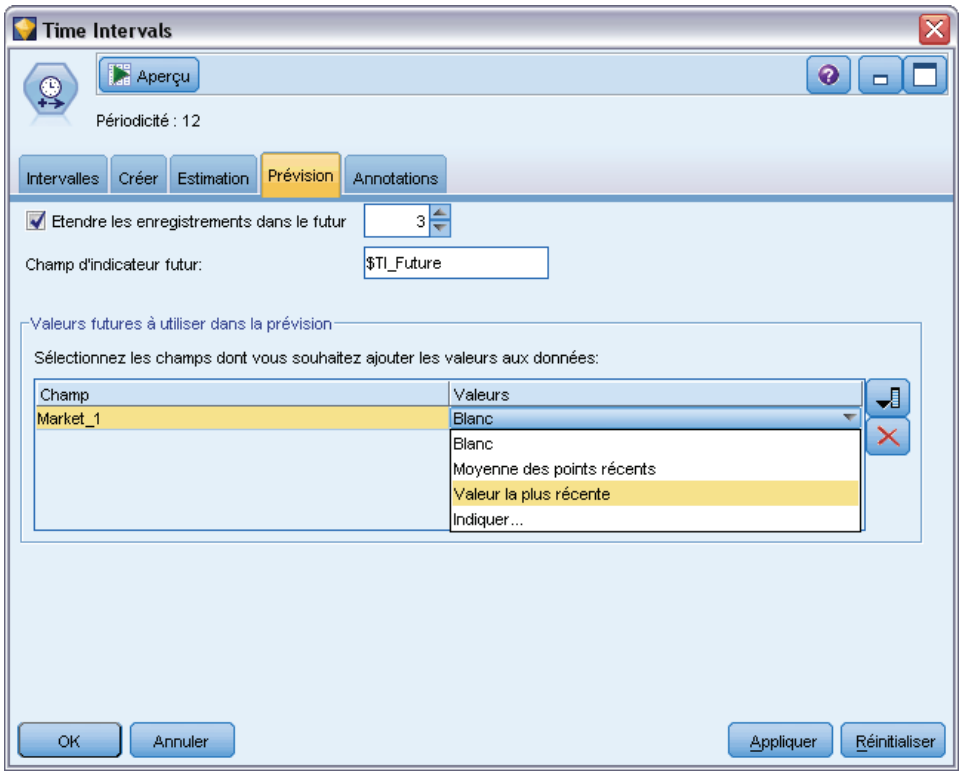
Commencer l'estimation. Vous pouvez commencer la période d'estimation au début des données ou exclure les valeurs anciennes peu utiles pour la prévision. Selon les données, le fait de raccourcir la période d'estimation permettra peut-être d'améliorer les performances (et de réduire le temps nécessaire pour préparer les données) sans perte significative de précision des prévisions.

Terminer l'estimation. Vous pouvez estimer le modèle en utilisant tous les enregistrements, jusqu'à la fin des données, ou «retenir» les enregistrements les plus récents pour évaluer le modèle. Dans le dernier cas, vous "prévoyez" en fait des valeurs déjà connues. Cela vous permet de comparer des valeurs observées et prédictives pour évaluer l'efficacité du modèle.

Prévisions

L'onglet Prévision du noeud Intervalles de temps vous permet d'indiquer le nombre d'enregistrements dont vous souhaitez générer la prévision et de préciser les valeurs futures à utiliser pour la prévision en fonction des noeuds de modélisation Séries temporelles en aval. Ces paramètres peuvent être remplacés dans les noeuds de modélisation en aval si nécessaire. Cependant, il peut être plus pratique de les définir ici que de les définir pour chaque noeud distinct.

Figure 4-87
Noeud Intervalles de temps, onglet Prévision



Etendre les enregistrements dans le futur. Indique le nombre d'enregistrements à prévoir au-delà de la période d'estimation. Notez que ces enregistrements peuvent être ou ne pas être des «prévisions» selon le nombre d'ensembles de rétention indiqué dans l'onglet Estimation.

Champ d'indicateur futur. Etiquette du champ généré qui indique si un enregistrement contient des données de prévision. La valeur par défaut de l'étiquette est *$TI_Future*.

Valeurs futures à utiliser dans la prévision. Pour chaque enregistrement à prédire (ensembles de rétention exclus), si vous utilisez des champs variables indépendantes (dont le rôle défini comme *Entrée*), vous devez indiquer les valeurs estimées pour la période de prévision pour chaque variable indépendante. Vous pouvez indiquer les valeurs manuellement ou les choisir dans une liste.

- **Champ.** Cliquez sur le bouton de sélection de champ et choisissez les champs à utiliser en tant que variables indépendantes. Notez que les champs sélectionnés ici ne sont pas nécessairement utilisés dans la modélisation. Pour qu'un champ soit réellement utilisé en tant que variable indépendante, vous devez le sélectionner dans un noeud de modélisation en aval. Cette boîte de dialogue est cependant pratique pour définir les valeurs futures, pour qu'elles puissent être partagées par plusieurs noeuds de modélisation en aval, sans les définir individuellement dans chaque noeud. Par ailleurs, la liste des champs disponibles peut être soumise à des contraintes dépendantes des sélections effectuées dans l'onglet Créer. Par exemple, si l'option Spécifier les champs et les fonctions est sélectionnée dans l'onglet Créer, les champs non agrégés ou étendus sont supprimés du flux et ne peuvent pas être utilisés dans la modélisation.

Remarque : Si des valeurs futures sont indiquées pour un champ qui n'est plus disponible dans le flux (car il a été supprimé ou parce que de nouvelles sélections ont été faites dans l'onglet Créer), le champ apparaît en rouge dans l'onglet Prévision.

- **Valeurs :** Pour chaque champ, vous pouvez faire un choix dans une liste de fonctions ou cliquer sur Spécifier pour entrer des valeurs manuellement ou faire un choix dans la liste des valeurs prédéfinies. Si les champs variables indépendantes correspondent à des éléments que vous contrôlez ou qui peuvent être déterminés par avance pour d'autres raisons, saisissez ces valeurs manuellement. Par exemple, si vous prévoyez les bénéfices du mois prochain d'un hôtel sur la base du nombre de réservations, vous pouvez indiquer le nombre de réservations réel pour cette période. A l'inverse, si un champ variable indépendante est en rapport avec un élément hors de votre contrôle, tel que le cours d'une valeur mobilière, vous pouvez utiliser une fonction telle que la valeur la plus récente ou la moyenne des points récents.

  Les fonctions disponibles dépendent du niveau de mesure du champ.

Le niveau de mesure	Fonctions
Champ continu ou nominal	Blanc Moyenne des points récents Valeur la plus récente Spécifier
Champ booléen	Blanc Valeur la plus récente Vrai Faux Spécifier

Moyenne des points récents— : calcule la valeur future à partir de la moyenne des trois derniers points de données.

Valeur la plus récente— : définit la valeur future en fonction du point de données le plus récent.

Vrai/Faux— : définit la valeur future d'un champ booléen comme Vrai ou Faux, selon le cas.

Spécifier— : ouvre une boîte de dialogue qui permet de définir manuellement les valeurs futures ou de les choisir dans une liste prédéfinie.

Figure 4-88
Spécification des valeurs futures des variables indépendantes

Valeurs futures

Vous pouvez spécifier les valeurs futures à utiliser pour la prévision par les noeuds de modélisation Séries temporelles en aval. Ces paramètres peuvent être remplacés dans les noeuds de modélisation en aval si nécessaire. Cependant, il peut être plus pratique de les définir ici que de les définir pour chaque noeud distinct.

Vous pouvez entrer ces valeurs manuellement ou cliquer sur le bouton de sélection, dans la partie droite de la boîte de dialogue, pour faire votre choix dans la liste des valeurs définies pour le champ en cours.

Le nombre de valeurs futures que vous pouvez définir correspond au nombre d'enregistrements dont vous étendez la série temporelle dans l'avenir.

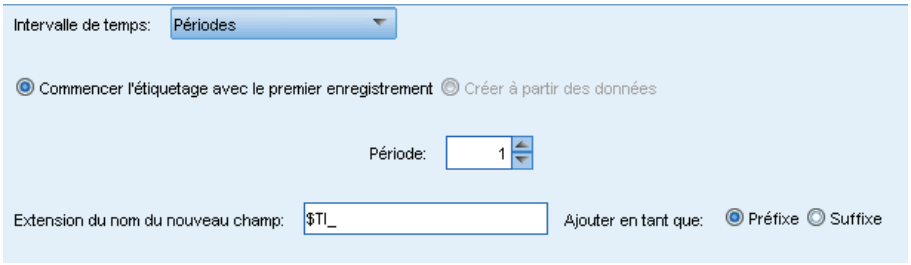
Intervalles pris en charge

Le noeud Intervalles de temps prend en charge l'intervalle total des intervalles, des secondes aux années, ainsi que les périodes cycliques (par exemple, saisonnières) et non cycliques. Vous définissez l'intervalle dans le champ Intervalles de temps, dans l'onglet Intervalles.

Périodes

Sélectionnez Périodes pour étiqueter une série existante non cyclique qui ne correspond à aucun autre intervalle spécifié. La série doit déjà se trouver dans le bon ordre et comporter un intervalle uniforme entre chaque mesure. L'option Créer à partir des données n'est pas disponible lorsque cet intervalle est sélectionné.

Figure 4-89
Paramètres d'intervalles de temps pour des périodes non cycliques



Echantillon de résultat

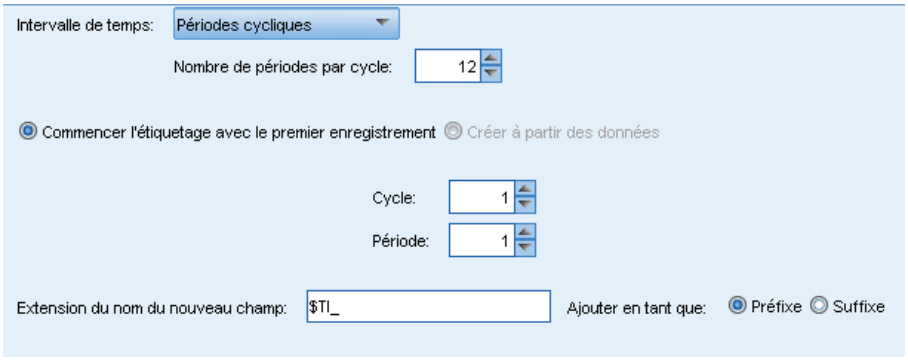
Les enregistrements sont étiquetés de manière incrémentielle sur la base de la valeur de départ indiquée (*Période 1*, *Période 2*, etc.). Les nouveaux champs sont créés comme suit :

$TI_TimeIndex (Entier)	$TI_TimeLabel (Chaîne)	$TI_Period (Entier)
1	Période 1	1
2	Période 2	2
3	Période 3	3
4	Période 4	4
5	Période 5	5

Périodes cycliques

Sélectionnez Périodes cycliques pour étiqueter une série existante comportant un cycle à répétition qui ne correspond pas à l'un des intervalles standard. Par exemple, vous pouvez utiliser cette option si votre exercice comptable ne comprend que 10 mois. La série doit déjà se trouver dans le bon ordre et comporter un intervalle uniforme entre chaque mesure. (L'option Créer à partir des données n'est pas disponible lorsque cet intervalle est sélectionné.)

Figure 4-90
Paramètres d'intervalles de temps pour des périodes cycliques



Echantillon de résultat

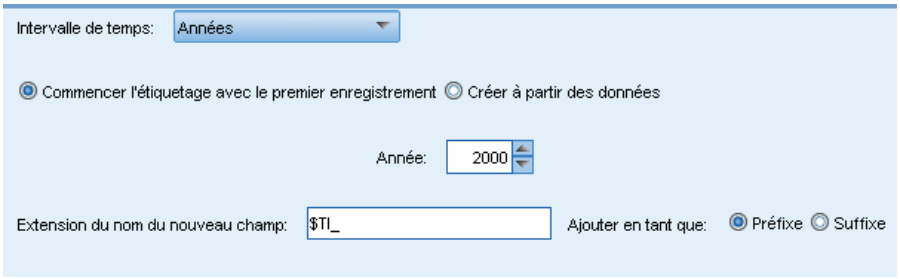
Les enregistrements sont étiquetés de manière incrémentielle sur la base du cycle et de la période de départ spécifiés (*Cycle 1*, *Période 1*, *Cycle 1*, *Période 2*, etc.). Par exemple, si le nombre de périodes par cycle est paramétré sur 3, les nouveaux champs sont créés comme suit :

$TI_TimeIndex (Entier)	**$TI_TimeLabel (Chaîne)**	**$TI_Cycle (Entier)**	**$TI_Period (Entier)**
1	Cycle 1, Période 1	1	1
2	Cycle 1, Période 2	1	2
3	Cycle 1, Période 3	1	3
4	Cycle 2, Période 1	2	1
5	Cycle 2, Période 2	2	2

Années

Pour les intervalles de type années, vous pouvez spécifier l'année de départ à partir de laquelle étiqueter des enregistrements consécutifs. Vous pouvez également sélectionner Créer à partir des données afin d'indiquer un champ d'horodatage ou de date qui identifie l'année de chaque enregistrement.

Figure 4-91
Paramètres d'intervalles de temps pour une série annuelle



Echantillon de résultat

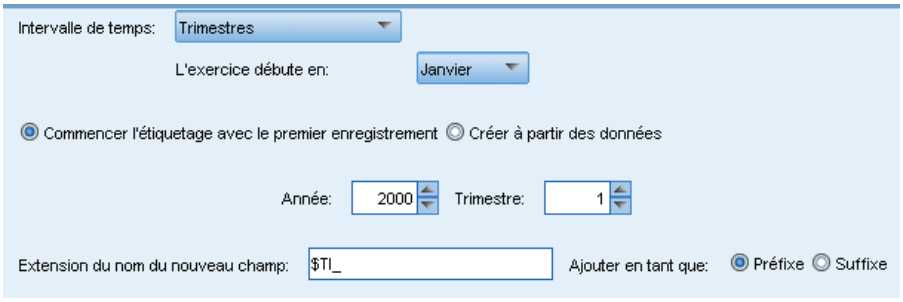
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	$TI-TimeLabel (Chaîne)	$TI-Year (Entier)
1	2000	2000
2	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

Trimestres

Pour une série trimestrielle, vous pouvez spécifier le mois de début de l'exercice comptable. Vous pouvez également définir le trimestre et l'année de départ (par exemple, T1 2000) à partir desquels étiqueter des enregistrements consécutifs, ou bien encore sélectionner Créer à partir des données afin de choisir un champ d'horodatage ou de date qui identifie le trimestre et l'année de chaque enregistrement.

Figure 4-92
Paramètres d'intervalles de temps pour une série trimestrielle



Echantillon de résultat

Pour un exercice comptable commençant en janvier, les nouveaux champs sont créés et renseignés comme suit :

$TI-TimeIndex (Entier)	$TI-TimeLabel (Chaîne)	$TI-Year (Entier)	$TI-Quarter (Entier avec étiquettes)
1	T1 2000	2000	1 (T1)
2	T2 2000	2000	2 (T2)
3	T3 2000	2000	3 (T3)
4	T4 2000	2000	4 (T4)
5	T1 2001	2001	1 (T1)

Si l'année débute à partir d'un autre mois que le mois de janvier, les nouveaux champs se présentent comme suit (si l'on considère un exercice comptable commençant en juillet, par exemple). Pour afficher les étiquettes qui identifient les mois de chaque trimestre, activez l'affichage des étiquettes de valeur en cliquant sur l'icône de la barre d'outils.

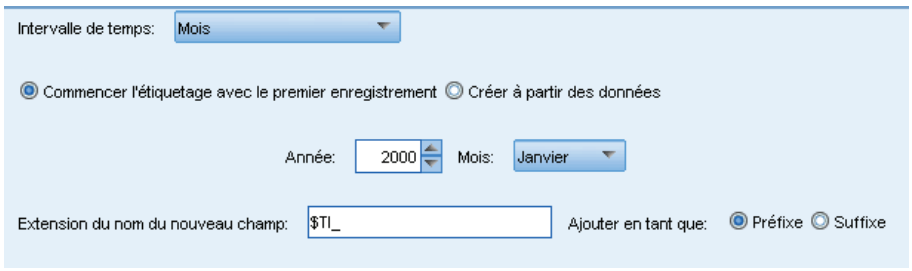
Figure 4-93
Icône des étiquettes de valeur

$TI-TimeIndex (Entier)	$TI-TimeLabel (Chaîne)	$TI-Year (Entier)	$TI-Quarter (Entier avec étiquettes)
1	T1 2000/2001	1	1 (T1 Juil-Sep)
2	T2 2000/2001	1	2 (T2 Oct-Déc)
3	T3 2000/2001	1	3 (T3 Jan-Mar)
4	T4 2000/2001	1	4 (T4 Avr-Juin)
5	T1 2001/2002	2	1 (T1 Juil-Sep)

Mois

Vous pouvez définir l'année et le mois de départ à partir desquels étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage ou de date qui indique le mois de chaque enregistrement.

Figure 4-94
Paramètres d'intervalles de temps pour une série mensuelle



Echantillon de résultat

Les nouveaux champs sont créés comme suit :

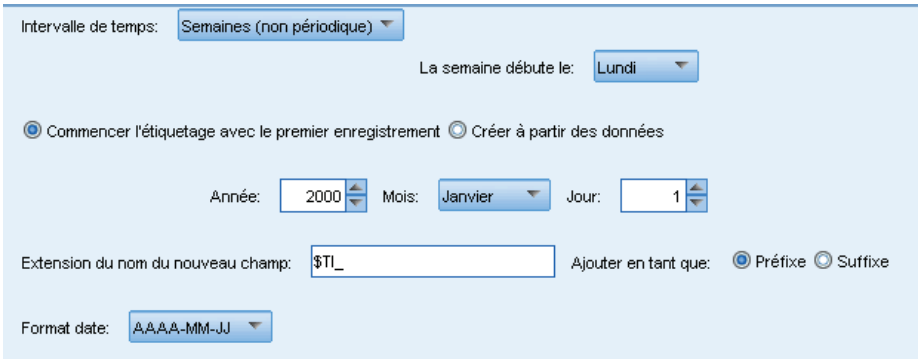
$TI-TimeIndex (Entier)	$TI-TimeLabel (Date)	$TI-Year (Entier)	$TI-Months (Entier avec étiquettes)
1	Jan 2000	2000	1 (janvier)
2	Fév 2000	2000	2 (février)
3	Mar 2000	2000	3 (mars)
4	Avr 2000	2000	4 (avril)
5	Mai 2000	2000	5 (mai)

Semaines (hors période)

Pour une série hebdomadaire, vous pouvez sélectionner le jour de la semaine où le cycle commence.

Notez que les semaines peuvent uniquement être hors période car les mois, les trimestres et même les années ne comportent pas nécessairement le même nombre de semaines. Toutefois, les données horodatées peuvent être facilement agrégées ou étendues au niveau hebdomadaire pour les modèles hors période.

Figure 4-95
Paramètres d'intervalles de temps pour une série hebdomadaire



Echantillon de résultat

Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	$TI-TimeLabel (Date)	$TI-Week (Entier)
1	1999-12-27	1
2	2000-01-03	2
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

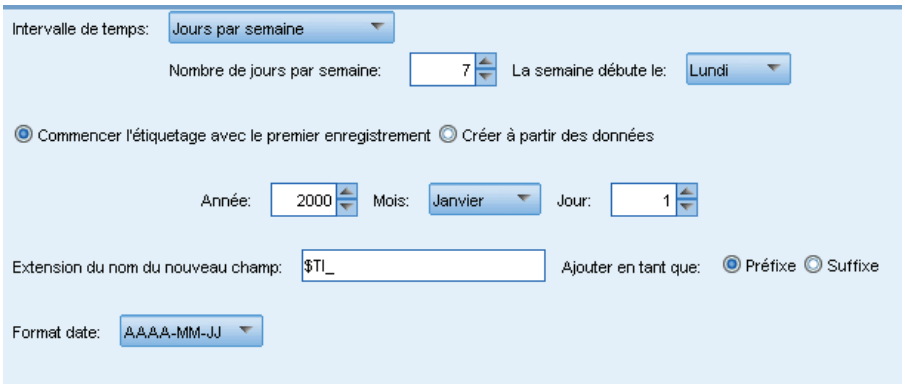
Le champ *$TI-TimeLabel* pour une semaine affiche le premier jour de cette semaine. Dans le tableau précédent, l'utilisateur commence l'étiquetage au 1er janvier 2000. Cependant, la semaine commence le lundi et le 1er janvier 2000 est un samedi. Par conséquent, la semaine qui inclut le 1er janvier commence le 27 décembre 1999 et devient l'étiquette du premier point.

Le format de date détermine les chaînes produites pour le champ *$TI-TimeLabel*.

Jours par semaine

Pour les mesures quotidiennes incluses dans un cycle hebdomadaire, vous pouvez définir le nombre de jours par semaine, ainsi que le jour de début de chaque semaine. Vous pouvez spécifier une date de départ à partir de laquelle étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage ou de date qui indique la date de chaque enregistrement.

Figure 4-96
Paramètres d'intervalles de temps pour une série quotidienne



Echantillon de résultat

Les nouveaux champs sont créés comme suit :

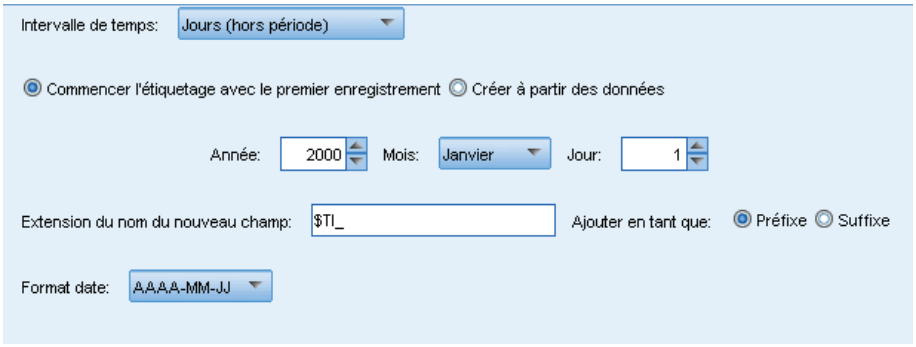
$TI-TimeIndex (Entier)	$TI-TimeLabel (Date)	$TI-Week (Entier)	$TI-Day (Entier avec étiquettes)
1	5 jan 2005	1	3 (mercredi)
2	6 jan 2005	1	4 (jeudi)
3	7 jan 2005	1	5 (vendredi)
4	10 jan 2005	2	1 (lundi)
5	11 jan 2005	2	2 (mardi)

Remarque : La semaine commence toujours à 1 pour la première période et les cycles ne suivent pas le calendrier. Ainsi, la semaine 52 est suivie de la semaine 53, puis 54, et ainsi de suite. La semaine ne correspond pas à une semaine réelle de l'année ; elle reflète simplement le nombre d'incréments hebdomadaires de la série.

Jours (hors période)

Choisissez l'option de jours non périodiques si vous disposez de mesures quotidiennes qui ne s'intègrent pas dans un cycle hebdomadaire standard. Vous pouvez spécifier une date de départ à partir de laquelle étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage ou de date qui indique la date de chaque enregistrement.

Figure 4-97
Paramètres d'intervalles de temps pour une série quotidienne (non périodique)



Echantillon de résultat

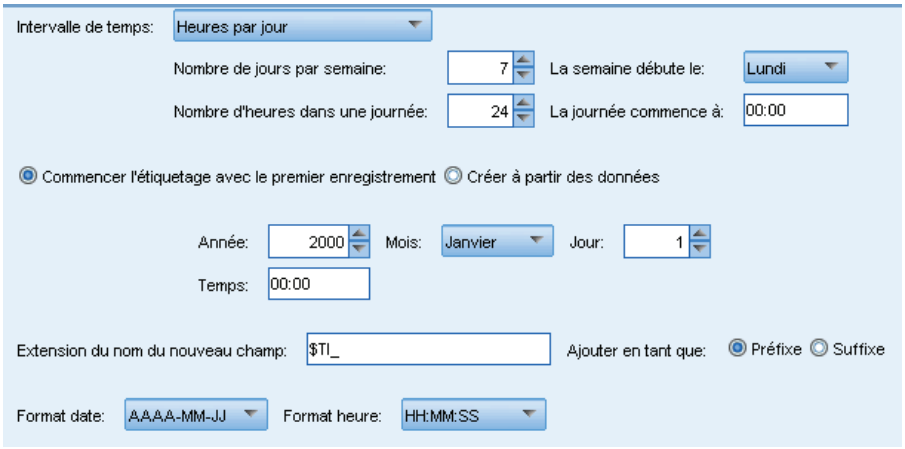
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	**$TI-TimeLabel (Date)**
1	5 jan 2005
2	6 jan 2005
3	7 jan 2005
4	8 jan 2005
5	9 jan 2005

Heures par jour

Pour des mesures horaires qui s'intègrent dans un cycle quotidien, vous pouvez définir le nombre de jours par semaine, le nombre d'heures par jour (une journée de travail de huit heures, par exemple), le jour de début de la semaine et l'heure de début de chaque jour. Les heures peuvent être indiquées à la minute près sur une base de 24 heures (par exemple, 14:05)

Figure 4-98
Paramètres d'intervalles de temps pour une série horaire



Vous pouvez spécifier la date et l'heure de départ à partir desquelles étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage qui identifie la date et l'heure de chaque enregistrement.

Echantillon de résultat

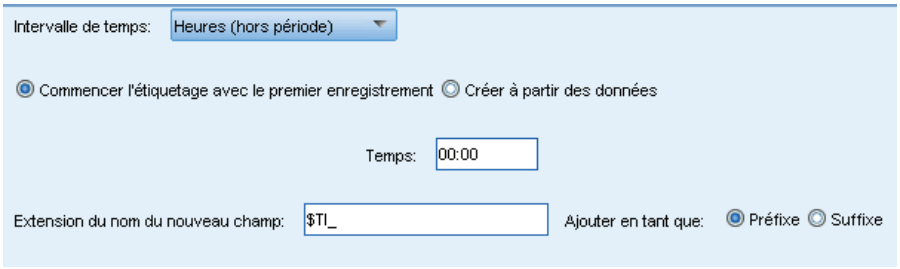
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	**$TI-TimeLabel (Horodatage)**	**$TI-Day (Entier avec étiquettes)**	**$TI-Hour (Entier avec étiquettes)**
1	5 jan 2005 08:00:00	3 (mercredi)	8 (8:00)
2	5 jan 2005 09:00:00	3 (mercredi)	9 (9:00)
3	5 jan 2005 10:00:00	3 (mercredi)	10 (10:00)
4	5 jan 2005 11:00:00	3 (mercredi)	11 (11:00)
5	5 jan 2005 12:00:00	3 (mercredi)	12 (12:00)

Heures (hors période)

Choisissez cette option si vous disposez de mesures horaires qui ne s'intègrent pas dans un cycle quotidien standard. Vous pouvez spécifier une date de départ à partir de laquelle étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ temporel ou un champ d'horodatage qui indique l'heure de chaque enregistrement.

Figure 4-99
Paramètres d'intervalles de temps pour des données annuelles



Les heures sont indiquées sur une base de 24 heures (13:00, par exemple), et ne décrivent pas une boucle (l'heure 25 vient après l'heure 24).

Echantillon de résultat

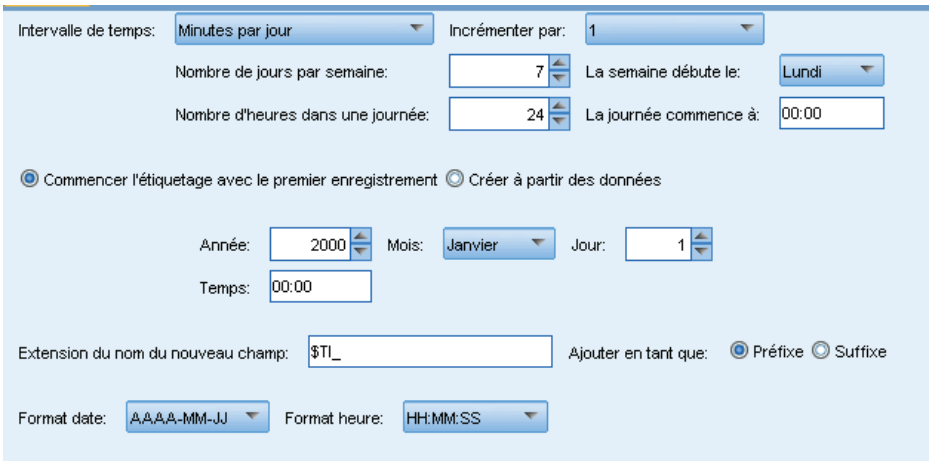
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	**$TI-TimeLabel (Chaîne)**	**$TI-Hour (Entier avec étiquettes)**
1	8:00	8 (8:00)
2	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

Minutes par jour

Pour des mesures réalisées à la minute près et incluses dans un cycle quotidien, vous pouvez définir le nombre de jours par semaine, le jour de début de la semaine, le nombre d'heures par jour et l'heure de début de la journée. Les heures sont indiquées sur une base de 24 heures, à la minute et à la seconde près, séparées par le signe deux-points (par exemple, 14:05:17). Vous pouvez également définir le nombre de minutes de l'incrément (chaque minute, toutes les deux minutes, etc., où l'incrément doit être une valeur divisible par 60).

Figure 4-100
Paramètres d'intervalles de temps pour les minutes par jour



Vous pouvez spécifier la date et l'heure de départ à partir desquelles étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage qui identifie la date et l'heure de chaque enregistrement.

Echantillon de résultat

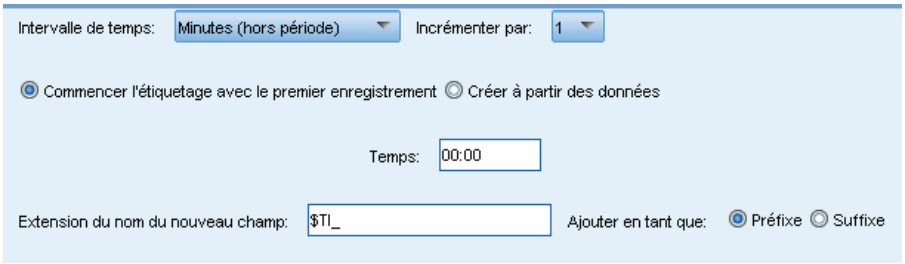
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	$TI-TimeLabel (Horodatage)	$TI-Minute
1	2005-01-05 08:00:00	0
2	2005-01-05 08:01:00	1
3	2005-01-05 08:02:00	2
4	2005-01-05 08:03:00	3
5	2005-01-05 08:04:00	4

Minutes (hors période)

Choisissez cette option si vous disposez de mesures définies à la minute près qui ne s'intègrent pas dans un cycle quotidien standard. Vous pouvez également spécifier le nombre de minutes de l'incrément (chaque minute, toutes les deux minutes, etc., où la valeur définie doit être un nombre divisible par 60).

Figure 4-101
Paramètres d'intervalles de temps pour les minutes (non périodiques)



Vous pouvez spécifier une date de départ à partir de laquelle étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ temporel ou un champ d'horodatage qui indique l'heure de chaque enregistrement.

Echantillon de résultat

Les nouveaux champs sont créés comme suit :

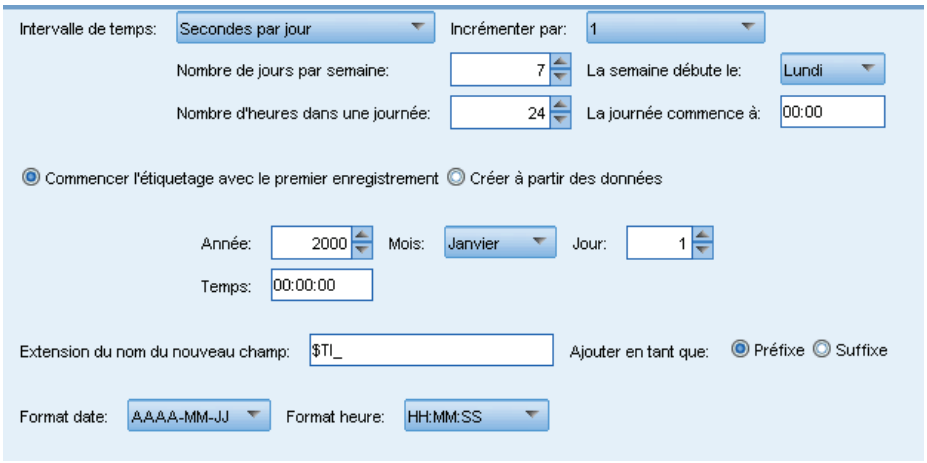
$TI-TimeIndex (Entier)	$TI-TimeLabel (Chaîne)	$TI-Minute
1	8:00	0
2	8:01	1
3	8:02	2
4	8:03	3
5	8:04	4

- La chaîne *TimeLabel* est créée à partir des heures et des minutes séparées par un deux-points. Les heures ne décrivent pas une boucle (l'heure 25 vient après l'heure 24).
- Les minutes sont incrémentées en fonction de la valeur spécifiée dans la boîte de dialogue. Par exemple, si l'incrément est de 2, la valeur *TimeLabel* est 8:00, 8:02, etc., et les minutes sont 0, 2, etc.

Secondes par jour

Pour les intervalles en secondes inclus dans un cycle quotidien, vous pouvez définir le nombre de jours par semaine, le jour de début de la semaine, le nombre d'heures par jour et l'heure de début de la journée. Les heures sont indiquées sur une base de 24 heures, à la minute et à la seconde près, séparées par le signe deux-points (par exemple, 14:05:17). Vous pouvez également spécifier le nombre de secondes de l'incrément (chaque seconde, toutes les deux secondes, etc., où la valeur définie doit être un nombre divisible par 60).

Figure 4-102
Paramètres d'intervalles de temps pour les secondes par jour



Vous pouvez spécifier la date et l'heure de départ à partir desquelles étiqueter des enregistrements consécutifs, ou sélectionner Créer à partir des données pour choisir un champ d'horodatage qui identifie la date et l'heure de chaque enregistrement.

Echantillon de résultat

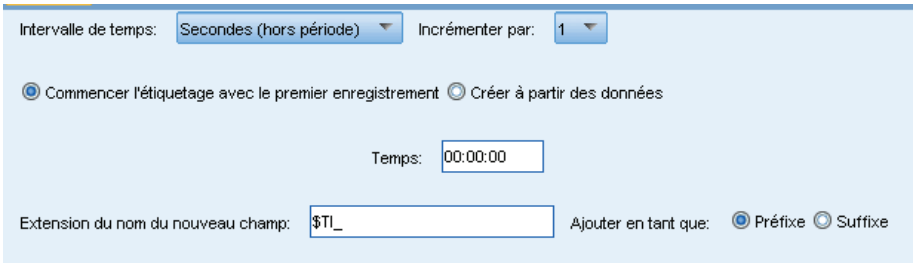
Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	$TI-TimeLabel (Horodatage)	$TI-Minute	$TI-Second
1	2005-01-05 08:00:00	0	0
2	2005-01-05 08:00:01	0	1
3	2005-01-05 08:00:02	0	2
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

Secondes (hors période)

Choisissez cette option si vous disposez de mesures définies à la seconde près qui ne s'intègrent pas dans un cycle quotidien régulier. Vous pouvez également spécifier le nombre de secondes de l'incrément (chaque seconde, toutes les deux secondes, etc., où la valeur définie doit être un nombre divisible par 60).

Figure 4-103
Paramètres d'intervalles de temps pour les secondes (non périodiques)



Spécifiez l'heure de départ de l'étiquetage des enregistrements consécutifs ou sélectionnez Créer à partir des données pour choisir un champ temporel ou un champ d'horodatage qui indique l'heure de chaque enregistrement.

Echantillon de résultat

Les nouveaux champs sont créés comme suit :

$TI-TimeIndex (Entier)	**$TI-TimeLabel (Chaîne)**	**$TI-Minute**	**$TI-Second**
1	8:00:00	0	0
2	8:00:01	0	1
3	8:00:02	0	2
4	8:00:03	0	3
5	8:00:04	0	4

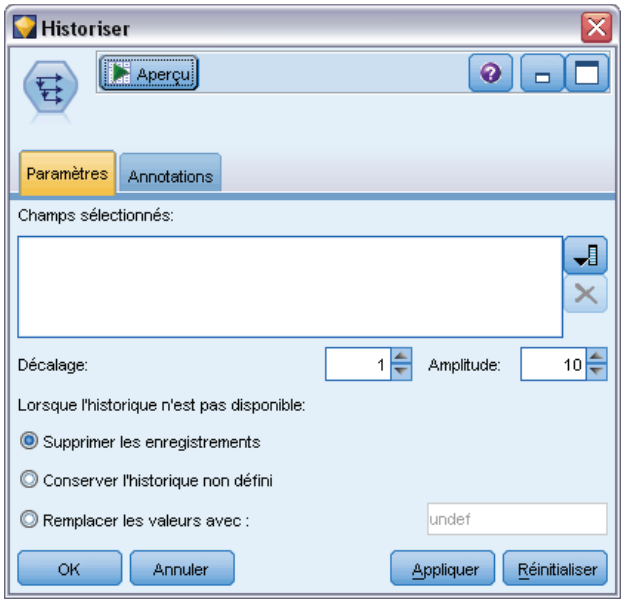
- La chaîne *TimeLabel* est créée à partir des heures, des minutes et des secondes séparées par un deux-points. Les heures ne décrivent pas une boucle (l'heure 25 vient après l'heure 24).
- Les secondes sont incrémentées en fonction du nombre défini comme incrément. Si l'incrément est de 2, la valeur *TimeLabel* est 8:00:00, 8:00:02, etc., et les secondes sont 0, 2, etc.

Noeud Historiser

Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Ils servent à créer des champs contenant des données provenant de champs d'enregistrements antérieurs. Lorsque vous utilisez un noeud Historiser, si vous souhaitez obtenir des données prétriées selon un champ particulier, vous pouvez utiliser le noeud Trier.

Paramétrage des options du noeud Historiser

Figure 4-104
Boîte de dialogue du noeud Historiser



Champs sélectionnés. A l'aide du sélecteur de champs (bouton à droite de la zone de texte), sélectionnez les champs pour lesquels vous souhaitez obtenir un historique. Chaque champ sélectionné est utilisé pour créer des champs pour tous les enregistrements de l'ensemble de données.

Décalage. Indiquez le dernier enregistrement avant l'enregistrement actuel à partir duquel extraire les valeurs de champ historiques. Par exemple, si le décalage est défini sur 3, au fur et à mesure que chaque enregistrement passe dans le noeud, les valeurs de champ du troisième enregistrement précédent sont incluses dans l'enregistrement actuel. Utilisez les paramètres d'amplitude pour indiquer jusqu'à quel enregistrement portera l'extraction. Utilisez les flèches pour rectifier la valeur de décalage.

Amplitude. Indiquez le nombre d'enregistrements précédents desquels extraire des valeurs. Par exemple, si le décalage est défini sur 3 et l'amplitude sur 5, chaque enregistrement qui passe dans le noeud se verra ajouter cinq champs pour chacun des champs spécifiés dans la liste Champs sélectionnés. Autrement dit, lorsque le noeud traite l'enregistrement 10, des champs provenant des enregistrements 7 à 3 sont ajoutés. Utilisez les flèches pour rectifier la valeur d'amplitude.

Lorsque l'historique n'est pas disponible. Sélectionnez l'une des options suivantes pour traiter les enregistrements qui n'ont pas de valeurs historiques. Il s'agit généralement des premiers enregistrements de l'ensemble de données, pour lesquels aucun enregistrement précédent ne peut être utilisé en tant qu'historique.

- **Supprimer les enregistrements.** Sélectionnez cette option pour supprimer les enregistrements dans lesquels aucune valeur d'historique n'est disponible pour le champ sélectionné.

- **Conserver l'historique non défini.** Sélectionnez cette option pour conserver les enregistrements dans lesquels aucune valeur d'historique n'est disponible. Une valeur non définie apparaît dans le champ d'historique en tant que ($null$).
- **Remplacer les valeurs avec.** Indiquez la valeur ou la chaîne à utiliser pour les enregistrements dans lesquels aucune valeur d'historique n'est disponible. La valeur de remplacement par défaut est *undef*, la valeur système nulle. Les valeurs nulles sont indiquées par la chaîne $null$.

Lorsque vous sélectionnez une valeur de remplacement, gardez à l'esprit les règles suivantes pour que l'exécution se déroule correctement :

- Les champs sélectionnés doivent être du même type de stockage.
- Si tous les champs sélectionnés présentent un stockage numérique, la valeur de remplacement doit être analysée en tant qu'entier.
- Si tous les champs sélectionnés présentent un stockage réel, la valeur de remplacement doit être analysée en tant que nombre réel.
- Si tous les champs sélectionnés présentent un stockage symbolique, la valeur de remplacement doit être analysée en tant que chaîne.
- Si tous les champs sélectionnés présentent un stockage date/heure, la valeur de remplacement doit être analysée en tant que champ date/heure.

Si l'une des conditions ci-dessus n'est pas remplie, une erreur se produit lors de l'exécution du noeud Historiser.

Noeud Re-trier

Le noeud Re-trier permet de définir l'ordre naturel utilisé pour afficher les champs situés en aval. Cet ordre a une incidence sur l'affichage des champs en différents endroits : tableaux, listes et sélecteur de champs. Cette opération est utile, par exemple, lorsque vous utilisez des ensembles de données volumineux pour rendre plus visibles les champs intéressants.

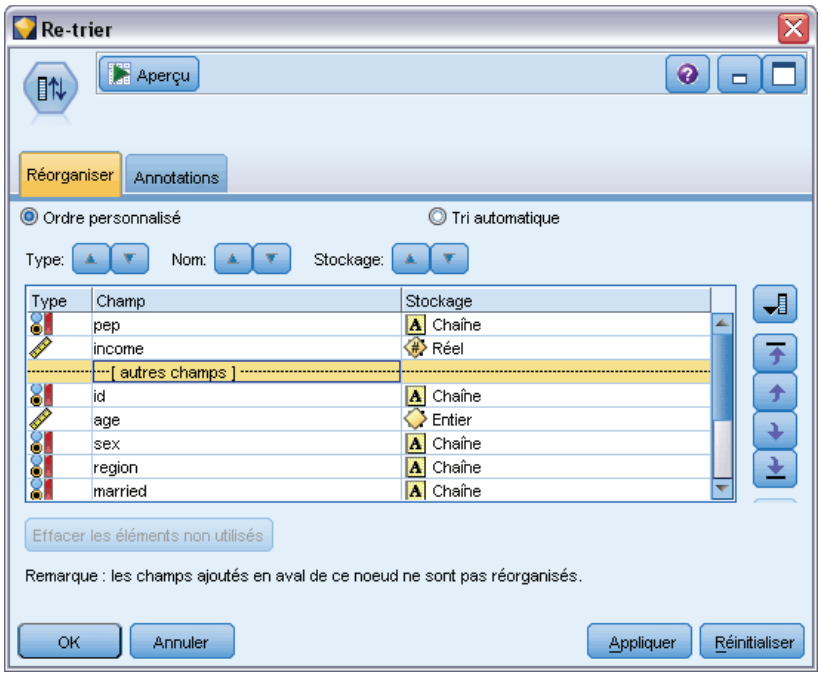
Paramétrage des options du noeud Re-trier

Il existe deux méthodes de réorganisation des champs : ordre personnalisé et tri automatique.

Ordre personnalisé

Sélectionnez Ordre personnalisé pour activer une table de noms et de types de champ dans laquelle vous pouvez afficher tous les champs et utiliser les flèches pour créer un ordre personnalisé.

Figure 4-105
Réorganisation en vue d'afficher les champs intéressants en premier



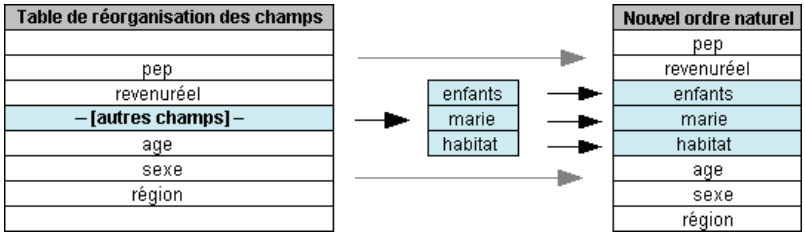
Pour réorganiser les champs :

- Sélectionnez un champ dans le tableau. Utilisez la méthode Ctrl+clic pour sélectionner plusieurs champs.
- Utilisez les boutons représentant une simple flèche pour déplacer les champs d'un rang vers le haut ou vers le bas.
- Utilisez les boutons représentant une flèche et une ligne pour placer les champs tout en bas ou tout en haut de la liste.
- Spécifiez l'ordre des champs qui ne sont pas inclus ici en déplaçant la ligne séparatrice, indiquée par [autres champs], vers le haut ou vers le bas.

Autres champs. L'objectif de la ligne séparatrice [autres champs] est de diviser la table en deux parties.

- Les champs qui apparaissent au-dessus de la ligne séparatrice sont ordonnés (tels qu'ils apparaissent dans la table) avant tous les ordres naturels utilisés pour afficher les champs en aval de ce noeud.
- Les champs qui apparaissent au-dessous de la ligne séparatrice sont ordonnés (tels qu'ils apparaissent dans la table) après tous les ordres naturels utilisés pour afficher les champs en aval de ce noeud.

Figure 4-106
Diagramme illustrant la manière dont les "autres champs" sont incorporés dans le nouvel ordre des champs



- Tous les champs qui n'apparaissent pas dans la table de réorganisation des champs figurent entre les champs "supérieurs" et "inférieurs" à l'emplacement de la ligne séparatrice.

Voici d'autres options de tri personnalisé :

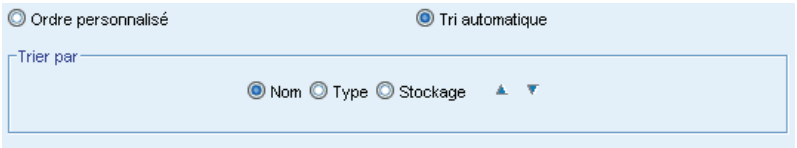
- Triez les champs dans l'ordre croissant ou décroissant en cliquant sur les flèches situées au-dessus de chaque en-tête de colonne (Type, Nom et Stockage). Lorsque vous effectuez un tri par colonne, les champs qui n'y sont pas mentionnés (ceux indiqués par la ligne [autres champs]) sont triés en dernier dans leur ordre naturel.
- Cliquez sur Effacer les éléments non utilisés pour supprimer du noeud Re-trier tous les champs inutilisés. Les champs inutilisés sont affichés en rouge dans le tableau. Cette couleur indique que le champ a été supprimé dans des opérations en amont.
- Indiquez l'ordre de tous les nouveaux champs (les nouveaux champs ou les champs non spécifiés sont identifiés par une icône représentant un éclair). Lorsque vous cliquez sur OK ou sur Appliquer, l'icône disparaît.

Remarque : si des champs sont ajoutés en amont après l'application d'un tri personnalisé, ces nouveaux champs sont ajoutés à la fin de la liste personnalisée.

Tri automatique

Sélectionnez Tri automatique pour indiquer le paramètre de tri. La boîte de dialogue change de manière dynamique pour proposer les options de tri automatique.

Figure 4-107
Réorganisation de tous les champs à l'aide des options de tri automatique



Trier par. Sélectionnez l'un des trois modes de tri des champs du noeud Réorganiser. Les flèches indiquent si l'ordre est croissant ou décroissant. Sélectionnez une option pour apporter une modification.

- Nom
- Type
- Stockage

Les champs ajoutés en amont du noeud Re-trier après l'application d'un tri automatique sont automatiquement placés à l'endroit qui convient, en fonction du type de tri sélectionné.

Chapitre
5

Noeuds Graphiques

Fonctions communes des noeuds Graphiques

Au cours de plusieurs étapes du processus de Data mining, des graphiques et des diagrammes sont utilisés pour explorer les données introduites dans IBM® SPSS® Modeler. Par exemple, vous pouvez connecter un noeud Nuage ou Proportion à une source de données pour obtenir un aperçu des types de données et des proportions. Vous pouvez ensuite effectuer des manipulations de champ et d'enregistrement afin de préparer les données pour des opérations de modélisation en aval. Les graphiques permettent également de vérifier les proportions et les relations entre des champs nouvellement calculés.

La palette Graphiques contient les noeuds suivants :

Le noeud Représentation graphique propose différents types de graphiques dans un noeud unique. Ce noeud permet de choisir les champs de données que vous souhaitez explorer puis de sélectionner un graphique parmi ceux disponibles pour les données sélectionnées. Le noeud filtre automatiquement tous les types de graphiques ne fonctionnant pas avec les sélections de champs. Pour plus d'informations, reportez-vous à la section Noeud Représentation Graphique sur p. 254.

Le noeud Nuage montre les relations existant entre les champs numériques. Vous pouvez créer un graphique Nuage à l'aide de points (diagramme de dispersion) ou de lignes. Pour plus d'informations, reportez-vous à la section Noeud Nuage sur p. 303.

Le noeud Proportion fournit l'occurrence des valeurs symboliques (catégorielles), comme un type de prêt hypothécaire ou le sexe d'un individu. Ce noeud est souvent utilisé pour montrer les déséquilibres des données, déséquilibres que vous pouvez rectifier à l'aide d'un noeud Equilibrer avant la création d'un modèle. Pour plus d'informations, reportez-vous à la section Noeud Proportion sur p. 311.

Le noeud Histogramme montre l'occurrence des valeurs des champs numériques. Il est souvent utilisé pour explorer les données avant toute création de modèles ou manipulation. Semblable au noeud Proportion, le noeud Histogramme sert souvent à montrer les déséquilibres des données. Pour plus d'informations, reportez-vous à la section Onglet Nuage d'histogramme sur p. 317.

Le noeud Résumé fournit la proportion de valeurs d'un champ numérique par rapport aux valeurs d'un autre champ. (Il génère des graphiques semblables aux histogrammes.) Il est utile pour illustrer une variable ou un champ dont les valeurs changent avec le temps. Grâce à la représentation graphique en 3D, vous pouvez en outre inclure un axe symbolique affichant les proportions par catégorie. Pour plus d'informations, reportez-vous à la section Onglet nuage de Résumé sur p. 321.

© Copyright IBM Corporation 1994, 2012.

Le noeud Courbes génère un graphique qui affiche plusieurs champs *Y* pour un seul champ *X*. Les champs *Y* sont représentés par des lignes colorées. Chacun équivaut à un noeud Nuage dont le style est défini sur Ligne et le mode X sur Trier. Les graphiques Courbes sont utiles lorsque vous souhaitez étudier la fluctuation de plusieurs variables au fil du temps. Pour plus d'informations, reportez-vous à la section Noeud Courbes sur p. 325.

Le noeud Relations illustre la force de la relation existant entre les valeurs de plusieurs champs symboliques (catégoriels). Le graphique utilise des lignes d'épaisseur différente pour représenter les forces de connexion. Par exemple, vous pouvez utiliser un noeud Relations pour explorer la relation avec l'achat d'un ensemble d'articles sur un site de commerce électronique. Pour plus d'informations, reportez-vous à la section Noeud Relations sur p. 330.

Le noeud Tracé horaire affiche un ou plusieurs ensembles de données temporelles. En règle générale, vous utilisez un noeud Intervalles de temps, en premier lieu, pour créer un champ *TimeLabel* qui servira d'étiquette à l'axe *x*. Pour plus d'informations, reportez-vous à la section Noeud Tracé horaire sur p. 341.

Le noeud Evaluation permet d'évaluer et de comparer des modèles prédictifs. Le graphique d'évaluation montre l'aptitude des modèles à prédire des résultats spécifiques. Il trie les enregistrements en fonction de la valeur prédite et de la confiance dans cette prévision. Il scinde les enregistrements en groupes de taille égale (**quantiles**), puis reporte la valeur du critère traité pour chaque quantile, du plus élevé au plus faible. Les divers modèles apparaissent sous forme de lignes dans le graphique. Pour plus d'informations, reportez-vous à la section Noeud Evaluation sur p. 346.

Une fois que vous avez ajouté un noeud Graphiques à un flux, vous pouvez double-cliquer sur le noeud pour ouvrir une boîte de dialogue qui permet de définir des options. La plupart des graphiques contiennent un certain nombre d'options spécifiques figurant sur un ou plusieurs onglets. Les onglets comportent également des options communes à tous les graphiques. Les sections suivantes contiennent des informations supplémentaires sur ces options communes.

Une fois que vous avez configuré les options d'un noeud Graphiques, vous pouvez exécuter ce dernier dans la boîte de dialogue ou au sein d'un flux. Dans la fenêtre du graphique créé, vous pouvez générer des noeuds Calculer (Binariser) et Sélectionner en fonction d'une sélection ou d'une zone de données, ce qui entraîne la définition de sous-ensembles de données. Par exemple, vous pouvez utiliser la puissance de cette fonction pour identifier et exclure les valeurs éloignées.

Apparences, superpositions, panneaux et animation

Superpositions et apparences

Les apparences (et les superpositions) ajoutent des dimensions à une visualisation. L'effet d'une apparence (regroupement, juxtaposition ou empilement) dépend du type de visualisation, du type de champ (variable), ainsi que du type de l'élément graphique et des statistiques. Il est par exemple possible d'utiliser un champ catégoriel de couleur pour grouper des points dans un diagramme de dispersion ou pour créer les piles d'un graphique à barres superposées. Un intervalle numérique continu de couleur peut également permettre d'indiquer les valeurs d'intervalle pour chaque point d'un diagramme de dispersion.

Vous devez faire des essais avec les différentes apparences et superpositions pour trouver la solution qui répond le mieux à vos besoins. Les descriptions suivantes peuvent vous aider à faire le bon choix.

Remarque : Certaines apparences ou superpositions ne conviennent pas à certains types de visualisation.

- **Couleur.** Lorsque la couleur est définie par un champ catégoriel, elle fractionne la visualisation en fonction des catégories individuelles, une couleur pour chaque catégorie. Lorsque la couleur est un intervalle numérique continu, elle varie en fonction de la valeur du champ d'intervalle. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement ou d'une observation et qu'un champ d'intervalle est utilisé pour la couleur, la couleur varie en fonction de la *moyenne* du champ d'intervalle.
- **Forme.** La forme est définie par un champ catégoriel qui fractionne la visualisation en éléments de différentes formes, une pour chaque catégorie.
- **Transparence.** Lorsque la transparence est définie par un champ catégoriel, elle fractionne la visualisation en fonction des catégories individuelles, un degré de transparence pour chaque catégorie. Lorsque la transparence est un intervalle numérique continu, elle varie en fonction de la valeur du champ d'intervalle. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement ou d'une observation et qu'un champ d'intervalle est utilisé pour la transparence, la couleur varie en fonction de la *moyenne* du champ d'intervalle. A la valeur la plus élevée, les éléments graphiques sont complètement transparents. A la valeur la plus basse, ils sont complètement opaques.
- **Etiquette de données.** Les étiquettes de données sont définies par tout type de champ dont les valeurs sont utilisées pour créer des étiquettes qui sont attachées aux éléments graphiques.
- **Taille.** Lorsque la taille est définie par un champ catégoriel, elle fractionne la visualisation en fonction des catégories individuelles, une taille pour chaque catégorie. Lorsque la taille est un intervalle numérique continu, elle varie en fonction de la valeur du champ d'intervalle. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement ou d'une observation et qu'un champ d'intervalle est utilisé pour la taille, la taille varie en fonction de la *moyenne* du champ d'intervalle.

Figure 5-1
Graphique avec une superposition de couleurs

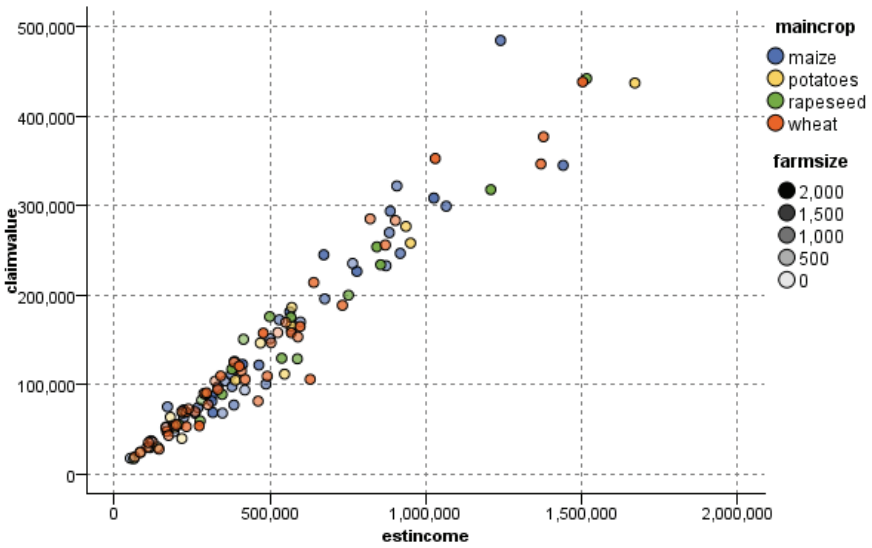
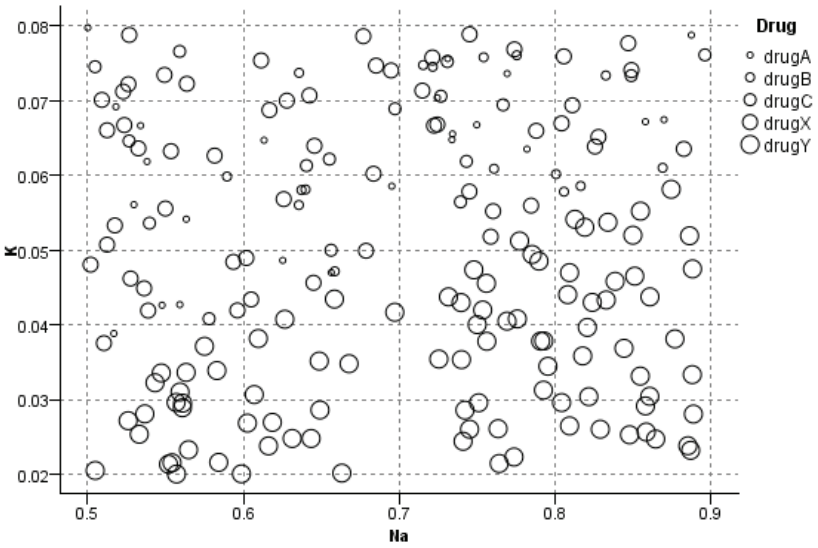


Figure 5-2
Graphique avec une superposition de tailles

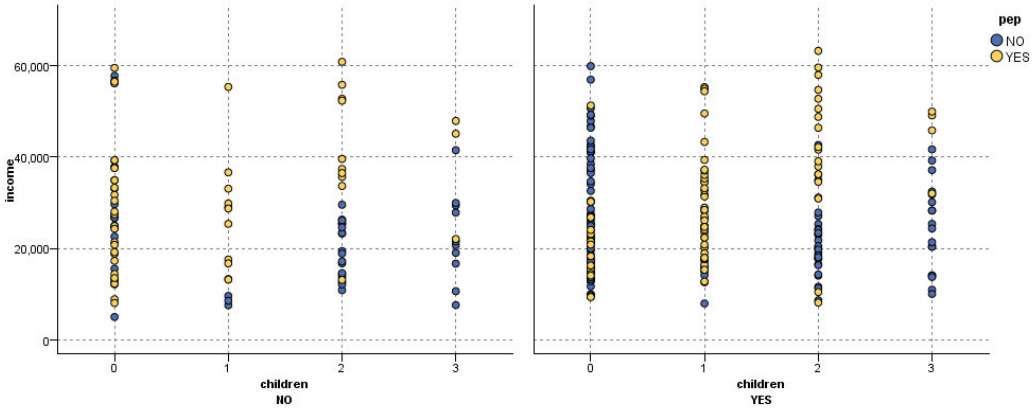


Panneaux et animation

Division en panels. La Panélisation, également appelée Création de facettes, crée un tableau de graphiques. Un graphique est généré pour chaque catégorie dans les champs de panélisation, mais tous les panneaux apparaissent simultanément. La panélisation permet de vérifier si la

visualisation est soumise aux conditions des champs de panélisation. Vous pouvez par exemple panéliser un histogramme par sexe pour déterminer si les distributions des fréquences sont égales parmi les hommes et les femmes. Autrement dit, vous pouvez vérifier si le salaire est soumis aux différences de sexe. Sélectionnez un champ catégoriel pour la création de panneaux.

Figure 5-3
Graphique avec panneaux indiquant le statut familial (OUI/NON)



Animation. L'animation ressemble à la panélisation du fait que plusieurs graphiques sont créés à partir des valeurs du champ d'animation, mais ces graphiques ne sont pas représentés ensemble. Vous utilisez alors les commandes en mode d'interaction pour animer la sortie et parcourir une séquence de graphiques individuels. En outre, contrairement à la panélisation, l'animation ne nécessite pas de champ catégoriel. Vous pouvez indiquer un champ continu dont les valeurs sont fractionnées en intervalles automatiquement. Vous pouvez faire varier la taille de l'intervalle à l'aide des commandes d'animation en mode d'interaction. Certaines visualisations n'offrent pas l'option d'animation.

Figure 5-4
Graphique Nuage animé utilisant une variable dotée de trois catégories, curseur à un niveau de tension artérielle faible

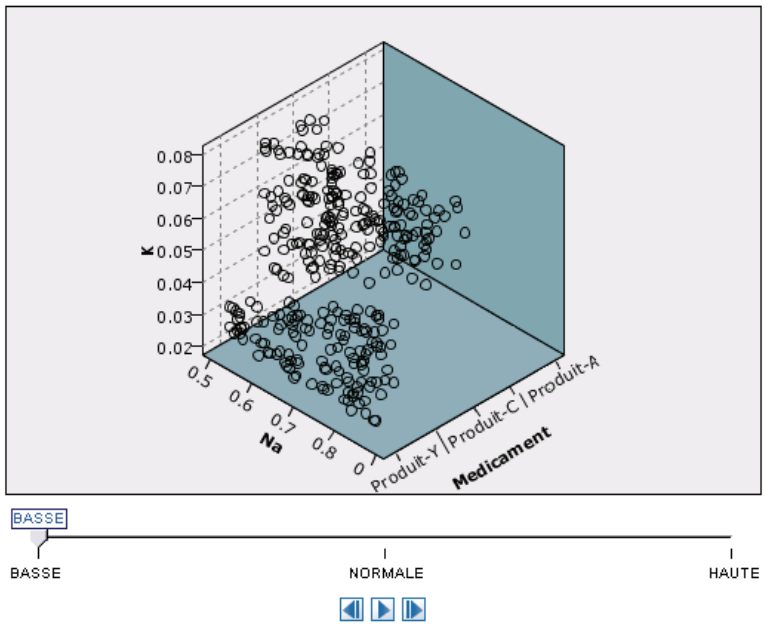


Figure 5-5
Graphique Nuage animé utilisant une variable dotée de trois catégories, curseur à un niveau de tension artérielle normale

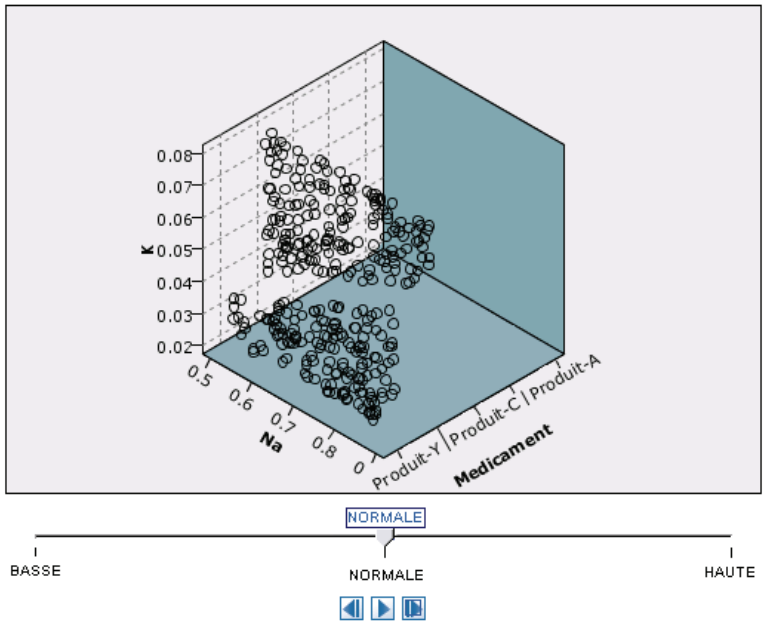
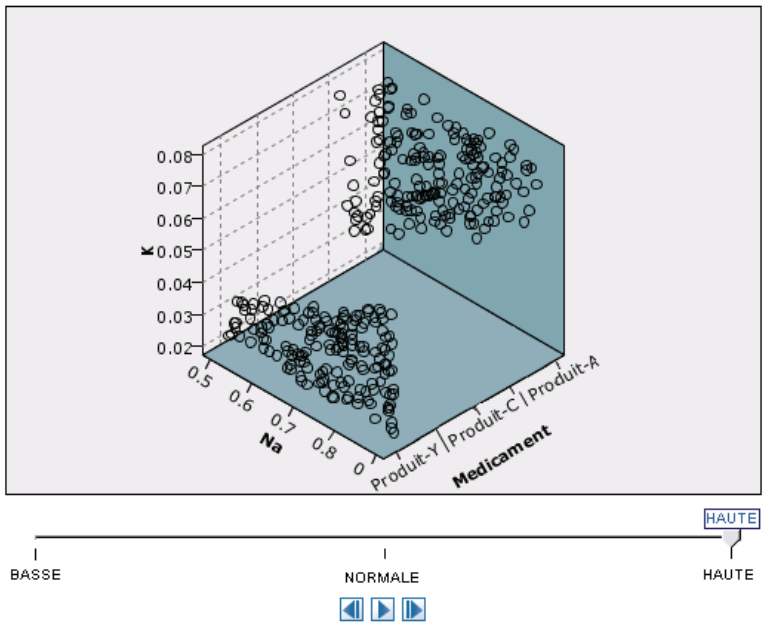


Figure 5-6
Graphique Nuage animé utilisant une variable dotée de trois catégories, curseur à un niveau de tension artérielle élevée



Utilisation de l'onglet Sortie

Vous pouvez définir les options suivantes pour tous les types de graphiques : elles concernent les noms de fichier et l'affichage des graphiques générés.

Remarque : Les graphiques de noeud Proportion ont des fonctions supplémentaires.

Nom de sortie. Spécifie le nom du graphique généré lorsque le noeud est exécuté. L'option Automatique sélectionne un nom en fonction du nœud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir Personnalisé pour indiquer un autre nom.

Sortie à l'écran. Sélectionnez cette option pour générer et afficher le graphique dans une nouvelle fenêtre.

Sortie dans le fichier. Sélectionnez cette option pour enregistrer la sortie sous forme de fichier.

- **Graphique de sortie.** Sélectionnez cette option pour produire la sortie sous forme de graphique. Disponible uniquement dans les noeuds Proportion.
- **Table de sortie.** Sélectionnez cette option pour produire la sortie sous forme de tableau. Disponible uniquement dans les noeuds Proportion.
- **Nom du fichier.**Indiquez le nom de fichier du graphique ou du tableau généré. Utilisez le bouton ... pour indiquer l'emplacement d'un fichier spécifique.
- **Type de fichier.** Spécifiez le type de fichier dans la liste déroulante. Pour tous les noeuds Graphiques, à l'exception du noeud Proportion avec une option Tableau de sortie, les types de fichiers graphiques disponibles sont les suivants.

  - Bitmap (*.bmp*)

- PNG (*.png*)
- Objet de sortie (*.cou*)
- JPEG (*.jpg*)
- HTML (*.html*)
- Document ViZml (*.xml*) à utiliser dans d'autres applications IBM® SPSS® Statistics.

Pour l'option Tableau de sortie dans le noeud Proportion, les types de fichiers disponibles sont les suivants.

- Données délimitées par des tabulations (*.tab*)
- Données délimitées par une virgule (*.csv*)
- HTML (*.html*)
- Objet de sortie (*.cou*)

Paginer la sortie. Lorsque vous enregistrez la sortie au format HTML, cette option est activée pour vous permettre de contrôler la taille de chaque page HTML. (S'applique uniquement au noeud Proportion.)

Lignes par page. Lorsque vous choisissez l'option Paginer la sortie, cette option est activée pour vous permettre de déterminer la longueur de chaque page HTML. Le paramètre par défaut est de 400. (S'applique uniquement au noeud Proportion.)

Utilisation de l'onglet Annotations

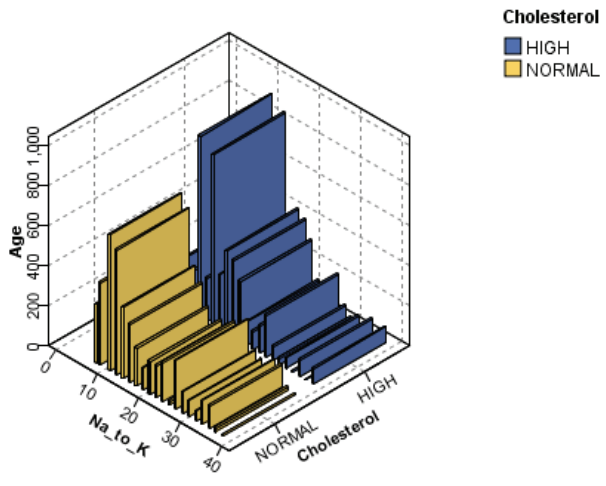
Utilisé pour tous les noeuds, cet onglet propose des options permettant de renommer les noeuds, de créer des info-bulles personnalisées et de stocker de longues annotations.

Graphiques en 3D

Les graphiques Nuage et Résumé de IBM® SPSS® Modeler permettent d'afficher des informations sur un troisième axe. Vous disposez ainsi d'une flexibilité accrue lorsque vous visualisez vos données pour sélectionner des sous-ensembles ou calculez de nouveaux champs en vue d'une modélisation.

Une fois que vous avez créé un graphique en 3D, vous pouvez cliquer dessus et faire glisser votre souris pour le faire tourner et le voir sous n'importe quel angle.

Figure 5-7
Graphique Résumé avec axes x, y et z



Il existe deux méthodes pour créer des graphiques en 3D dans SPSS Modeler : représenter des informations sur un troisième axe (véritables graphiques en 3D) ou afficher le graphique avec un effet 3D. Ces deux méthodes sont disponibles pour les graphiques Nuage et Résumé.

Pour représenter des informations sur un troisième axe :

- Dans la boîte de dialogue du noeud Graphiques, cliquez sur l'onglet Nuage.
- Cliquez sur le bouton 3D afin d'activer les options de l'axe *z*.
- Utilisez le sélecteur de champs pour sélectionner le champ de l'axe *z*. Dans certains cas, seuls les champs symboliques sont autorisés. Le sélecteur de champs affiche les champs appropriés.

Pour ajouter un effet 3D à un graphique :

- Une fois le graphique créé, cliquez sur l'onglet Graphiques dans la fenêtre de sortie.
- Cliquez sur le bouton 3D pour convertir la vue en un graphique en trois dimensions.

Noeud Représentation Graphique

Le noeud Représentation Graphique vous permet de choisir parmi de nombreuses sorties graphiques différentes (diagrammes en barres, graphiques sectoriels, histogrammes, diagrammes de dispersion, cartes thermiques, etc.) dans un seul noeud. Dans le premier onglet, vous commencez par choisir les champs de données que vous souhaitez explorer, puis le noeud vous présente une sélection de types de graphiques fonctionnant pour vos données. Le noeud filtre automatiquement tous les types de graphiques ne fonctionnant pas avec les sélections de champs. Vous pouvez définir des options de graphiques détaillées ou plus avancées dans l'onglet Détaillées.

Remarque : vous devez connecter le noeud Représentation Graphique à un flux avec des données afin d'éditer le noeud ou de sélectionner des types de graphiques.

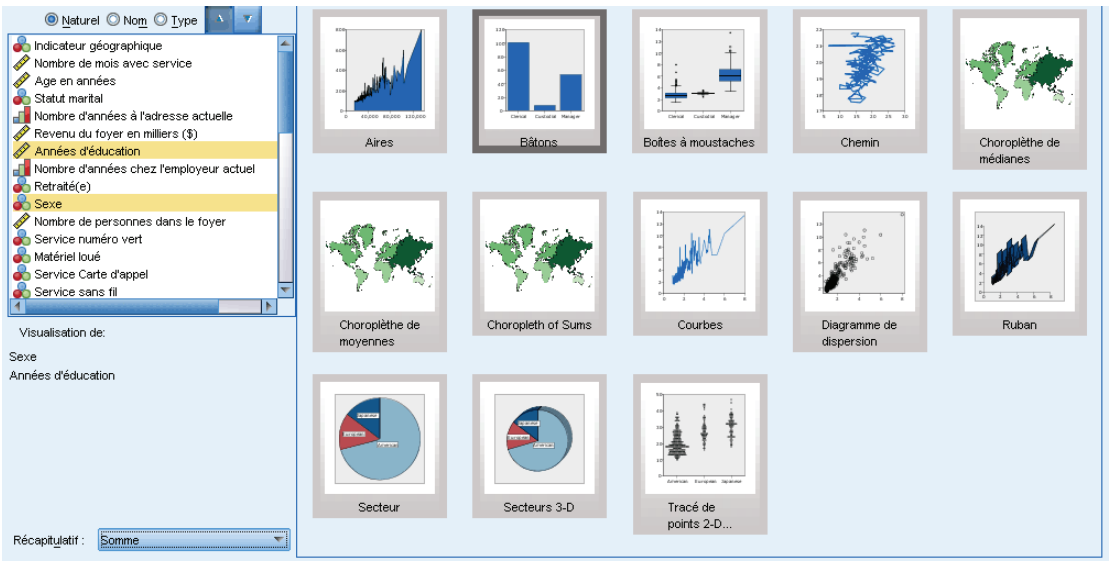
Deux boutons vous permettent de contrôler les modèles de visualisation (et les feuilles de style et les cartes) disponibles :

Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, reportez-vous à la section Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est noté à droite du bouton. Pour plus d'informations, reportez-vous à la section Définition de l'emplacement des modèles, des feuilles de style et des cartes sur p. 290.

Représentation graphique Onglet Base

Figure 5-8
Onglet Base



Si vous n'êtes pas sûr du type de visualisation qui représenterait le mieux vos données, utilisez l'onglet Base. Lorsque vous sélectionnez vos données, un sous-ensemble de types de visualisation appropriés aux données vous est proposé. Pour plus d'exemples, reportez-vous à Représentation graphique - Exemples sur p. 271.

- Sélectionnez un ou plusieurs champs (variables) dans la liste. Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs.

Remarque : le niveau de mesure d'un champ détermine les types de visualisation disponibles. le niveau de mesure peut être modifié en cliquant avec le bouton droit sur le champ dans la liste et en choisissant une option. Pour plus d'informations sur les types de niveau de mesure disponibles, reportez-vous à la rubrique Types de champ (variable) sur p. 258.

- Sélectionnez un type de visualisation. Pour obtenir des descriptions des types disponibles, reportez-vous à la section Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.

- Pour certaines visualisations, vous pouvez choisir des statistiques récapitulatives. Différents sous-ensembles de statistiques sont disponibles selon si la statistique est basée sur un comptage ou calculée à partir d'un champ continu. Les statistiques disponibles dépendent du modèle lui-même. Une liste de toutes les statistiques disponibles est décrite à l'étape suivante.

- Si vous voulez définir d'autres options, telles que des apparences et des champs de panneaux en option, cliquez sur Détaillé. Pour plus d'informations, reportez-vous à la section Onglet détaillé de la représentation graphique sur p. 260.

Statistiques récapitulatives calculées à partir d'un champ continu

- **Moyenne.** Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.
- **Median.** Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.
- **Mode.** Valeur qui revient le plus fréquemment. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.
- **Minimum.** Valeur la plus petite d'une variable numérique.
- **Maximum.** Plus grande valeur d'une variable numérique.
- **Intervalle**. La différence entre les valeurs minimale et maximale.
- **Intervalle de milieu**. Le milieu de l'intervalle est la valeur pour laquelle la distance du minimum est égale à la distance du maximum.
- **Sum.** Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- **Somme cumulée**. Somme cumulée des valeurs. Chaque élément graphique affiche la somme correspondant à un sous-groupe à laquelle la somme totale des groupes précédents est ajoutée.
- **Somme de pourcentage**. Le pourcentage de chaque sous-groupe basé sur une valeur de champ comparée à la somme de tous les groupes.
- **Somme de pourcentage cumulé**. Le pourcentage cumulatif de chaque sous-groupe basé sur une valeur de champ comparée à la somme de tous les groupes. Chaque élément graphique affiche le pourcentage correspondant à un sous-groupe auquel le pourcentage total des groupes précédents est ajouté.
- **Variance.** Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.
- **Ecart-type.** Mesure de dispersion par rapport à la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si la moyenne d'âge est de 45

avec un écart-type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.

- **Erreur standard.** Mesure du degré de variation de la valeur d'une statistique test, d'un échantillon à l'autre. Il s'agit de l'écart-type de la distribution de l'échantillon pour une statistique. Par exemple, l'erreur standard de la moyenne est l'écart-type des moyennes d'échantillon.
- **Kurtosis.** Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.
- **Skewness.** Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Les statistiques suivantes de zone peuvent résulter en plusieurs éléments graphiques par sous-groupe. Lorsque des éléments graphiques tels les intervalles, les aires ou les bords sont utilisés, une statistique de zone résulte en un élément graphique affichant l'intervalle. Tous les autres éléments graphiques résultent en deux éléments distincts, l'un affichant le début de l'intervalle et l'autre affichant la fin de celui-ci.

- **Région : Intervalle**. L'intervalle de valeurs entre les valeurs minimale et maximale.
- **Région : Intervalle de confiance de moyenne de 95 %**. Intervalle de valeurs ayant 95% de chances de contenir la valeur moyenne de la population.
- **Région : Intervalle de confiance de 95 % d'un individu**. Intervalle de valeurs ayant 95% de chances de contenir la valeur prédite d'une observation individuelle.
- **Région : Ecart-type de 1 inférieur/supérieur à la moyenne**. Intervalle de valeurs entre un **écart-type** de 1 au-dessus et au-dessous de la **moyenne**.
- **Région : Erreur standard de 1 inférieure/supérieure à la moyenne**. Intervalle de valeurs entre une **erreur standard** de 1 au-dessus et au-dessous de la **moyenne**.

Statistiques récapitulatives basées sur un comptage

- **Comptage**. Nombre d'observations / de lignes
- **Nombre cumulé**. Nombre d'observations / de lignes cumulé. Chaque élément graphique affiche l'effectif correspondant à un sous-groupe auquel l'effectif total des groupes précédents est ajouté.

- **Pourcentage de compte**. Le pourcentage des observations / lignes dans chaque sous-groupe comparé au nombre total d'observations / de lignes.
- **Pourcentage de compte cumulé**. Le pourcentage cumulé des observations / lignes dans chaque sous-groupe comparé au nombre total d'observations / de lignes. Chaque élément graphique affiche le pourcentage correspondant à un sous-groupe auquel le pourcentage total des groupes précédents est ajouté.

Types de champ (variable)

Des icônes apparaissent en regard des champs dans la liste des champs pour indiquer le type de champ et le type des données. Les icônes permettent également d'identifier les ensembles de réponses multiples.

Niveau de mesure	**Le type de données**			
	Numérique	Chaîne	Date	Heure
Continu		n/a		
Vecteur ordonné				
Set				

Ensemble à réponses multiples, catégories multiples	
Ensemble de réponses multiples, dichotomies multiples	

Niveau de mesure

Le niveau de mesure d'un champ est important lors de la création d'une visualisation. Vous trouverez ci-dessous une description des niveaux de mesure. Le niveau de mesure peut être temporairement modifié en cliquant avec le bouton droit sur un champ dans la liste de champs et en choisissant une option. Dans la plupart des cas, vous devrez considérer uniquement les deux classifications de champs les plus larges, catégorielle et continue :

Qualitatives. Données possédant un nombre limité de valeurs ou de catégories distinctes (par exemple, le sexe ou la religion). Les champs catégoriels peuvent être des chaînes (alphanumériques) ou des champs numériques qui utilisent des codes numériques pour représenter les catégories (par exemple 0 = *homme* et 1 = *femme*). Elles sont également appelées données qualitatives. Les ensembles, les ensembles ordonnés et les booléens sont des champs catégoriels.

-
-
-

Continue. Données mesurées sur une échelle d'intervalle ou une échelle de rapport, où les valeurs des données indiquent à la fois l'ordre des valeurs et la distance entre ces valeurs. Par exemple, un salaire de 72 195 dollars est supérieur à un salaire de 52 398 dollars, et la distance entre les deux valeurs est 19 797 dollars. Egalement appelées données quantitatives ou d'échelle ou données d'intervalle numérique.

Les champs catégoriels définissent des catégories dans la visualisation, en général pour séparer ou grouper des éléments graphiques. Les champs continus sont souvent regroupés dans des catégories de champs catégoriels. Par exemple, une visualisation par défaut de revenus pour des catégories par sexe affichera le revenu moyen des hommes et celui des femmes. Les valeurs brutes des champs continus peuvent aussi être représentées graphiquement comme dans un diagramme de dispersion. Par exemple, un diagramme de dispersion peut afficher le revenu actuel et le revenu de départ pour chaque observation. Un champ catégoriel peut être utilisé pour regrouper les observations par sexe.

Types de données

Le niveau de mesure n'est pas la seule propriété à déterminer le type d'un champ. Un champ est également stocké en tant que type de données spécifique. Les types de données spécifiques sont les chaînes (données non-numériques comme les lettres), les valeurs numériques (nombres réels) et les dates. A la différence du niveau de mesure, le type de données d'un champ ne peut pas être modifié de manière temporaire. Il vous faut changer le mode de stockage des données dans l'ensemble de données d'origine.

Vecteurs de multiréponses

Certains fichiers de données prennent en charge un type spécial de champ appelé **ensemble de réponses multiples**. Les ensembles à réponses multiples ne sont pas vraiment des « champs » au sens habituel du terme. Les ensembles à réponses multiples utilisent des champs multiples pour enregistrer les réponses aux questions lorsque la personne interrogée peut donner plus d'une réponse. Les ensembles à réponses multiples sont traités comme des champs catégoriels ; et on peut faire avec les ensembles à réponses multiples pratiquement tout ce qu'on peut faire avec les champs catégoriels.

Les ensembles de réponses multiples peuvent être des ensembles de dichotomies multiples ou des ensembles de catégories multiples.

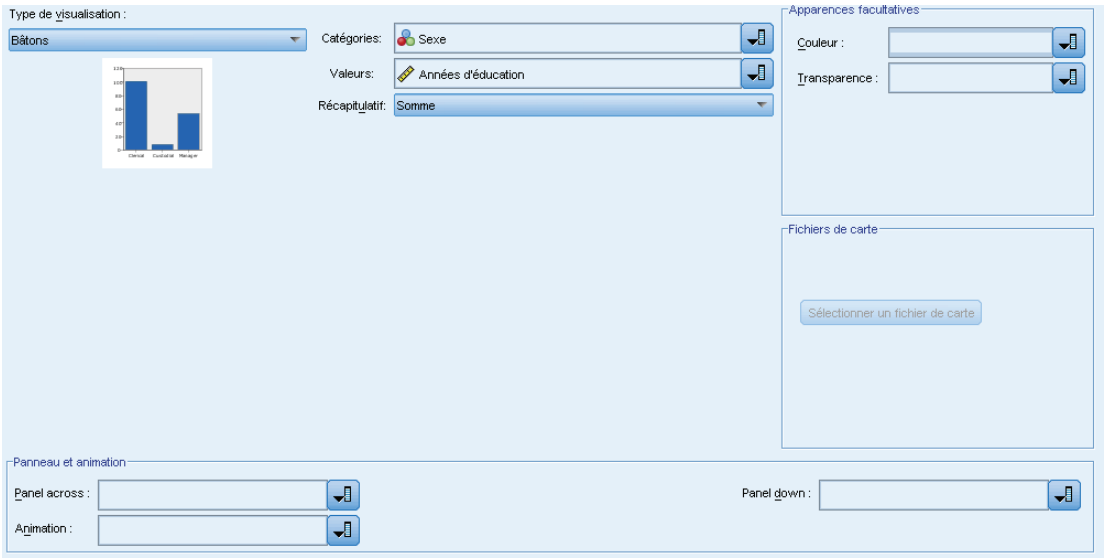
Vecteurs de dichotomies multiples. Un ensemble de dichotomies multiples consiste généralement en plusieurs champs dichotomiques : des champs pouvant prendre deux valeurs uniquement comme oui/non, présent/absent, coché/décoché. Les champs peuvent ne pas être strictement dichotomiques, cependant tous les champs de l'ensemble sont codés de la même manière.

Par exemple, dans une enquête qui fournit cinq réponses possibles à la question « A laquelle des sources suivantes vous fiez-vous pour vos informations ? » La personne interrogée peut indiquer plusieurs choix en cochant une case en regard de chaque choix. Les cinq réponses deviennent cinq champs dans le fichier de données, codées 0 pour *Non* (non cochée) et 1 pour*Oui* (cochée).

Ensemble de catégories multiples. Un ensemble de catégories multiples consiste en plusieurs champs, tous codés de la même façon, avec le plus souvent plusieurs catégories de réponses possibles. Par exemple, une question dans une enquête vous demande « Nommez trois nationalités qui décrivent le mieux votre origine ethnique » . Il existe des centaines de réponses possibles à cette question, mais à des fins de codage la liste est limitée aux 40 nationalités les plus courantes, toutes les autres possibilités étant regroupées sous l'étiquette « autre » . Dans le fichier de données, les trois réponses deviennent trois champs, contenant chacun 41 catégories (40 codées et une catégorie « autre »).

Onglet détaillé de la représentation graphique

Figure 5-9
Onglet Détaillé



Utilisez l'onglet Détaillé lorsque vous savez quel type de visualisation vous voulez créer ou lorsque vous voulez ajouter des apparences, des panneaux et/ou des animations en option à une visualisation. Pour plus d'exemples, reportez-vous à Représentation graphique - Exemples sur p. 271.

- Si vous avez sélectionné un type de visualisation sur l'onglet Base, il sera affiché. Sinon, choisissez-en un dans la liste déroulante. Pour plus d'informations sur les types de visualisation, reportez-vous à la rubrique Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.
- Juste à droite de l'image miniature de visualisation se trouvent les commandes permettant de définir les champs (variables) requis pour le type de visualisation. Vous devez spécifier l'ensemble de ces champs.
- Pour certaines visualisation, vous pouvez sélectionner des statistiques récapitulatives. Dans certains cas (avec les graphiques à barres par exemple), vous pouvez utiliser l'une des ces options récapitulatives pour l'apparence transparence. Pour obtenir des descriptions des statistiques récapitulatives, reportez-vous à la section Représentation graphique Onglet Base sur p. 255.

- Vous pouvez sélectionner une ou plusieurs des apparences en option. Celles-ci peuvent ajouter des dimensions en vous permettant d'inclure d'autres champs dans la visualisation. Vous pouvez par exemple utiliser un champ pour faire varier la taille des points dans un diagramme de dispersion. Pour plus d'informations sur les apparences en option, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247. Veuillez noter que l'apparence transparence n'est pas prise en charge via la génération de scripts.
- Si vous créez une visualisation de carte, le groupe Fichiers cartes affiche le ou les fichiers cartes qui seront utilisés. Si un fichier carte par défaut existe, ce fichier aparaît. Pour modifier le fichier carte, cliquez sur Sélectionner un fichier carte pour afficher la boîte de dialogue Sélectionner les cartes. Vous pouvez également spécifier le fichier carte par défaut dans cette boîte de dialogue. Pour plus d'informations, reportez-vous à la section Sélection des fichiers cartes pour les visualisations de carte sur p. 261.
- Vous pouvez sélectionner une ou plusieurs des options de création de panneaux ou d'animation. Pour plus d'informations sur les options de création de panneaux et d'animation, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Sélection des fichiers cartes pour les visualisations de carte

Si vous sélectionnez un modèle de visualisation de carte, il vous faut un fichier carte qui définit les informations géographiques permettant de dessiner la carte. S'il n'existe pas de fichier carte par défaut, celui-ci sera utilisé pour la visualisation de carte. Pour choisir un autre fichier carte, cliquez sur Sélectionner un fichier carte dans l'onglet Détaillé pour afficher la boîte de dialogue Sélectionner les cartes.

La boîte de dialogue Sélectionner les cartes vous permet de choisir un fichier carte principal et un fichier carte de référence. Les fichiers cartes définissent les informations géographiques permettant de dessiner la carte. Votre application est installée avec un ensemble de fichiers cartes standard. S'il y a d'autres fichiers de forme ESRI que vous souhaitez utiliser, vous devez d'abord convertir ces fichiers en fichiers SMZ. Pour plus d'informations, reportez-vous à la section Conversion et distribution des fichiers de formes Carte sur p. 293. Après avoir converti la carte, cliquez sur Gérer... dans la boîte de dialogue Sélecteur de modèles pour importer la carte dans le système de gestion afin qu'il soit disponible dans la boîte de dialogue Sélectionner les cartes.

Les points suivants sont à prendre en compte lors de la spécification des fichiers cartes :

- Tous les modèles de cartes nécessitent au moins un fichier carte.
- Le fichier carte relie généralement un attribut-clé de carte à la clé de données.
- Si le modèle ne nécessite pas de clé de carte qui relie à une clé de données, il nécessite un fichier carte de référence et des champs spécifiant les coordonnées (comme la longitude et la latitude) pour dessiner les éléments sur la carte de référence.
- Les modèles de carte en superposition nécessitent deux cartes : un fichier carte principal et un fichier carte de référence. La carte de référence est dessinée en premier pour qu'elle se trouve derrière le fichier carte principal.

Pour des informations sur la terminologie des cartes comme les attributs et les fonctions, consultez Concepts principaux des cartes sur p. 294.

Fichier carte. Vous pouvez sélectionner n'importe quel fichier carte qui se trouve dans le système de gestion. Il contient des fichiers cartes préinstallés et les fichiers cartes que vous avez importés. Pour plus d'informations sur la gestion des fichiers cartes, consultez Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Clé de carte. Spécifiez l'attribut à utiliser comme clé qui relie le fichier carte à la clé de données.

Enregistrez le fichier carte et les paramètres par défaut. Sélectionnez cette case si vous souhaitez utiliser le fichier carte sélectionné comme fichier par défaut. Si vous avez spécifié un fichier carte par défaut, vous n'avez pas besoin de spécifier un fichier carte à chaque fois que vous créez une visualisation de carte.

Clé de données. Cette commande indique la même valeur que celle qui apparaît dans l'onglet Détaillé du Sélecteur de modèles. Elle est fournie ici dans un but pratique au cas où vous auriez besoin de modifier la clé en raison du fichier carte spécifique que vous auriez choisi.

Afficher toutes les fonctions des cartes dans la visualisation. Lorsque cette option est sélectionnée, toutes les fonctions de la carte apparaissent dans la visualisation, même s'il n'existe pas de valeur de clé de données correspondante. Si vous souhaitez uniquement voir les fonctions pour lesquelles vous avez des données , désélectionnez cette option. Les fonctions identifiées par les clés de carte affichées dans la liste Clés de carte sans correspondance ne sont pas rendues dans la visualisation.

Comparer les valeurs des cartes et des données. La clé de carte et la clé de données sont reliées l'une à l'autre pour créer la visualisation de carte. Les valeurs de ces deux clés doivent correspondre ou vous ne pourrez pas créer de visualisation de carte. Cliquez sur Comparer pour savoir si les valeurs de la clé de données et de la clé de carte correspondent. L'icône qui apparaît vous informe de l'état de la comparaison. Ces icônes sont décrites ci-après. Si une comparaison a été effectuée et qu'il existe des valeurs de clé de données sans valeurs de clé de carte correspondantes, les valeurs de clé de données apparaissent dans la liste Clés de données sans correspondance. Dans la liste Clés de carte sans correspondance, vous pouvez également voir quelles valeurs de clés de carte n'ont pas de valeurs de clés de données correspondantes. Si Afficher toutes les fonctions sur des cartes dans la visualisation n'est pas coché, les fonctions identifiées par ces valeurs de clés de carte ne seront pas rendues.

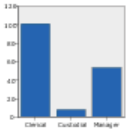
Table 5-1
Icônes de comparaison

Icône	Description
	Aucune comparaison n'a été effectuée. Il s'agit de l'état par défaut avant de cliquer sur Comparer. Vous devez procéder avec prudence car vous ne savez pas si les valeurs de clé de données et de clé de carte correspondent.
	Une comparaison a été effectuée et les valeurs de clé de données et de clé de carte correspondent entièrement. Pour chaque valeur de la clé de données, il existe une fonction correspondante identifiée par la clé de carte.
	Une comparaison a été effectuée et certaines valeurs de clé de données et de clé de carte ne correspondent pas. Pour certaines des valeurs de clé de données, il n'existe pas de fonction correspondante identifiée par la clé de carte. Vous devez procéder avec prudence. Si vous continuez, la visualisation de carte ne contiendra pas toutes les valeurs de données.
	Une comparaison a été effectuée et les valeurs de clé de données et de clé de carte ne correspondent pas. Choisissez une autre clé de données ou clé de carte car aucune carte ne sera proposée si vous continuez.

Types de visualisation des Représentations graphiques intégrées disponibles

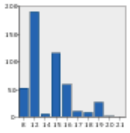
Vous pouvez créer plusieurs types différents de visualisations. Tous les types intégrés suivants sont disponibles dans les onglets de base et détaillé. Certaines des descriptions de ces modèles (particulièrement les modèles de cartes) identifient les champs (variables) spécifiés sur l'onglet Détaillé à l'aide de texte spécial.

Table 5-2
Types de diagrammes disponibles



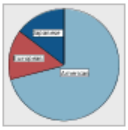
Bâtons
Calcule une statistique récapitulative d'un champ numérique continu et affiche les résultats de chaque modalité d'un champ qualitatif sous la forme de bâtons.

Requiert : Un champ qualitatif et un champ continu.



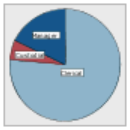
Effectifs en bâtons
Affiche la proportion de lignes/d'observations dans chaque modalité d'un champ qualitatif sous la forme de bâtons. Vous pouvez aussi utiliser le noeud de diagramme de distribution pour générer ce diagramme. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, reportez-vous à la section Noeud Proportion sur p. 311.

Requiert : Un champ qualitatif unique.



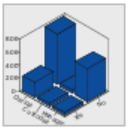
Secteur
Calcule la somme d'un champ numérique continu et affiche la proportion de cette somme distribuée dans chaque modalité d'un champ qualitatif sous la forme de secteurs.

Requiert : Un champ qualitatif et un champ continu.



Secteurs d'effectifs
Affiche la proportion de lignes/d'observations dans chaque catégorie d'un champ qualitatif sous la forme de secteurs.

Requiert : Un champ qualitatif unique.

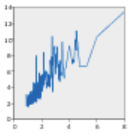


Bâtons 3D
Calcule une statistique récapitulative d'un champ numérique continu et affiche les résultats de l'intersection de modalités de deux champs qualitatifs.

Requiert : Une paire de champs qualitatifs et un champ continu.

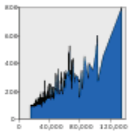
Secteurs 3-D
Il est identique au diagramme en secteur à l'exception de l'effet 3-D supplémentaire.

Requiert : Un champ qualitatif et un champ continu.



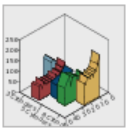
Courbes
Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace une ligne reliant les valeurs. Vous pouvez aussi utiliser le noeud de diagramme de traçage pour générer un diagramme en courbe. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, reportez-vous à la section Noeud Nuage sur p. 303.

Requiert : Une paire de champs de n'importe quel type.



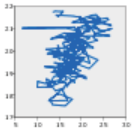
Aires
Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace une aire reliant les valeurs. La différence entre une courbe et une aire est infime en ce sens que l'aire ressemble à une courbe dont l'espace situé en dessous est colorié. Cependant, si vous utilisez un type esthétique en couleur, le résultat est un découpage de la courbe et un empilement de l'aire.

Requiert : Une paire de champs de n'importe quel type.



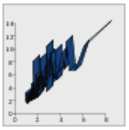
Aire 3-D
Affiche les valeurs d'un champ représenté par rapport aux valeurs d'un autre, et le découpe par champ qualitatif. Un élément d'aire est tracé pour chaque modalité.

Requiert : Un champ qualitatif et une paire de champs de n'importe quel type.



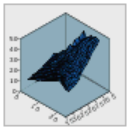
Chemin
Affiche les valeurs d'un champ représentées par rapport aux valeurs d'un autre champ, sous la forme d'une courbe reliant les valeurs dans l'ordre où elles apparaissent dans l'ensemble de données d'origine. L'ordre est la principale différence entre un chemin et une courbe.

Requiert : Une paire de champs de n'importe quel type.



Ruban
Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace un ruban reliant les valeurs. Un ruban est essentiellement une courbe dotée d'un effet 3-D. Il ne s'agit pas d'un véritable diagramme 3-D.

Requiert : Une paire de champs de n'importe quel type.



Surface
Affiche les valeurs de trois champs représentés par rapport à leurs valeurs réciproques, sous la forme d'une surface reliant les valeurs.

Requiert : Trois champs de n'importe quel type.

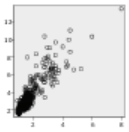


Diagramme de dispersion
Affiche les valeurs d'un champ représentées par rapport aux valeurs d'un autre champ. Ce diagramme peut mettre en évidence la relation entre les champs (s'il en existe). Vous pouvez aussi utiliser le noeud de diagramme de traçage pour générer une dispersion. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, reportez-vous à la section Noeud Nuage sur p. 303.

Requiert : Une paire de champs de n'importe quel type.

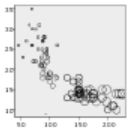
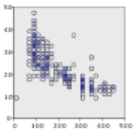


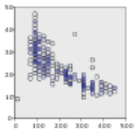
Diagramme à bulles
Tout comme une dispersion de base, il affiche les valeurs d'un champ représentées par rapport aux valeurs d'un autre champ. La différence tient en ce que les valeurs d'un troisième champ sont utilisées pour faire varier la taille de chaque point.

Requiert : Trois champs de n'importe quel type.



Dispersion groupée
Tout comme une dispersion de base, il affiche les valeurs d'un champ représentées par rapport aux valeurs d'un autre champ. La différence tient en ce que des valeurs similaires sont regroupées et qu'un type esthétique de couleur ou de taille est utilisé pour indiquer le nombre d'observations de chaque groupe.

Requiert : Une paire de champs continus.



Dispersion en groupes hexagonaux
Voir la description des dispersion groupées. La différence réside dans la forme des groupes sous-jacents qui ont la forme d'hexagones plutôt que de cercles. La dispersion en groupes hexagonaux résultant ressemble à la dispersion groupée. Cependant, le nombre de valeurs de chaque groupe est différent pour chaque diagramme à cause de la forme des groupes sous-jacents.

Requiert : Une paire de champs continus.

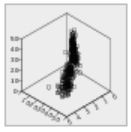
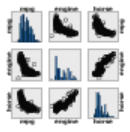


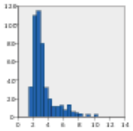
Diagramme de dispersion 3D
Affiche les valeurs de trois champs représentés les uns par rapport aux autres. Ce diagramme peut mettre en évidence la relation entre les champs (s'il en existe). Vous pouvez aussi utiliser le noeud de diagramme de traçage pour générer une dispersion en 3-D. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, reportez-vous à la section Noeud Nuage sur p. 303.



Matrice de dispersion (SPLOM)
Affiche les valeurs d'un champ représentées par rapport aux valeurs d'un autre champ pour chaque champ. Un SPLOM ressemble à un tableau de dispersions. Le SPLOM comprend aussi un histogramme de chaque champ.

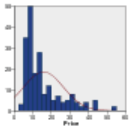
Requiert : Deux champs continus ou plus.

Requiert : Trois champs de n'importe quel type.



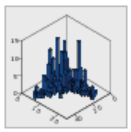
Histogramme
Affiche la distribution d'effectifs d'un champ. Un histogramme peut vous aider à déterminer le type de distribution et à voir si la distribution est asymétrique. Vous pouvez aussi utiliser le noeud de diagramme d'histogramme pour générer ce diagramme. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, reportez-vous à la section Onglet Nuage d'histogramme sur p. 317.

Requiert : Un seul champ de n'importe quel type.



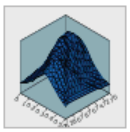
Histogramme avec distribution normale
Affiche la distribution d'effectifs d'un champ continu avec une courbe surimposée de la distribution normale.

Requiert : Un champ continu unique.



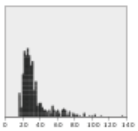
Histogramme 3-D
Affiche la distribution d'effectifs d'une paire de champs continus.

Requiert : Une paire de champs continus.



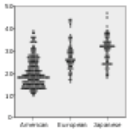
Densité 3-D
Affiche la distribution d'effectifs d'une paire de champs continus. Il est similaire à un histogramme 3-D, l'unique différence résidant dans le fait qu'une surface est utilisée à la place des bâtons pour afficher la distribution.

Requiert : Une paire de champs continus.



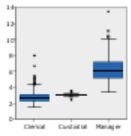
Graphique en points
Affiche les observations/lignes individuelles et les empile aux points de données distincts de l'axe *x*. Ce diagramme est similaire à un histogramme en ce fait qu'il affiche la distribution des données, mais il affiche chaque observation/ligne plutôt qu'un effectif agrégé d'un groupe spécifique (intervalle de valeurs).

Requiert : Un seul champ de n'importe quel type.



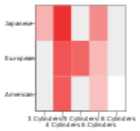
Tracé de points 2-D
Affiche les observations/lignes individuelles et les empile aux points de données distincts de l'axe *y* pour chaque modalité d'un champ qualitatif.

Requiert : Un champ qualitatif et un champ continu.



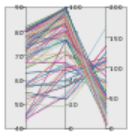
Boîtes à moustaches
Calcule les cinq statistiques (minimum, premier quartile, médiane, troisième quartile et maximum) d'un champ continu pour chaque modalité d'un champ qualitatif. Les résultats sont affichés sous la forme d'éléments de boîte à moustache/schéma. Les boîtes à moustaches peuvent vous aider à voir comment la distribution de données continues varie au sein des modalités.

Requiert : Un champ qualitatif et un champ continu.



Carte thermique
Calcule la moyenne d'un champ continu pour l'intersection de modalités de deux champs qualitatifs.

Requiert : Une paire de champs qualitatifs et un champ continu.



Parallèle
Crée des axes parallèles pour chaque champ et trace une ligne à travers la valeur de champ pour chaque ligne/observation des données.

Requiert : Deux champs continus ou plus.

Choroplèthe des totaux
Calcule l'effectif de chaque modalité d'un champ qualitatif (Clé de données) et dessine une carte qui utilise la saturation de couleur pour réprésenter les effectifs dans les fonctionnalités des cartes qui correspondent aux modalités.

Requiert : Un champ qualitatif. Un fichier carte dont la clé correspond aux modalités Clé de données.

Carte choroplèthe des moyennes/médianes/sommes
Calcule la moyenne, la médiane ou la somme d'un champ continu (Couleur) pour chaque modalité d'un champ qualitatif (Clé de données) et dessine une carte qui utilise la saturation de couleur pour représenter les statistiques calculées dans les fonctionnalités de la carte qui correspondent à ces modalités.

Requiert : Un champ qualitatif et un champ continu. Un fichier carte dont la clé correspond aux modalités Clé de données.

Choroplèthe de valeurs
Dessine une carte qui utilise la couleur pour représenter les valeurs d'un champ qualitatif (Couleur) pour les fonctionnalités de la carte qui correspondent aux valeurs définies par un autre champ qualitatif (Clé de données). S'il existe plusieurs valeurs qualitatives du champ Couleur pour chaque fonctionnalité, la valeur modale est utilisée.

Requiert : Une paire de champs qualitatifs. Un fichier carte dont la clé correspond aux modalités Clé de données.

Coordonnées sur une choroplèthe d'effectifs
Cette fonction est semblable à la carte chorophlète d'effectifs sauf qu'elle contient deux champs continus supplémentaires (Longitude et Latitude) qui identifient les coordonnées pour dessiner les points sur la carte choroplèthe.

Requiert : Un champ qualitatif et une paire de champs continus. Un fichier carte dont la clé correspond aux modalités Clé de données.

Coordonnées sur une carte choroplèthe des moyennes/médianes/sommes
Cette fonction est semblable à la carte chorophlète des moyennes/médianes/sommes sauf qu'elle contient deux champs continus supplémentaires (Longitude et Latitude) qui identifient les coordonnées pour dessiner les points sur la carte choroplèthe.

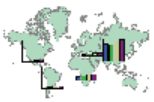
Requiert : Un champ qualitatif et trois champs continus. Un fichier carte dont la clé correspond aux modalités Clé de données.

Coordonnées sur une choroplèthe de valeurs
Cette fonction est semblable à la carte chorophlète de valeurs sauf qu'elle contient deux champs continus supplémentaires (Longitude et Latitude) qui identifient les coordonnées pour dessiner les points sur la carte choroplèthe.

Requiert : Une paire de champs qualitatifs et une paire de champs continus. Un fichier carte dont la clé correspond aux modalités Clé de données.

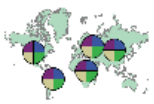
Barres d'effectifs sur une carte
Calcule la proportion de lignes/observations dans chaque modalité d'un champ qualitatif (Modalités) pour chaque fonctionnalité de carte (Clé de données) et dessine une carte et les diagrammes en bâtons au centre de chaque fonctionnalité de carte.

Requiert : Une paire de champs qualitatifs. Un fichier carte dont la clé correspond aux modalités Clé de données.



Barres sur une carte
Calcule une statistique récapitulative pour un champ continu (Valeurs) et affiche les résultats de chaque modalité d'un champ qualitatif (Modalités) pour chaque fonctionnalité de carte (Clé de données) sous la forme de diagrammes en bâtons placés au centre de chaque fonctionnalité de carte.

Requiert : Une paire de champs qualitatifs et un champ continu. Un fichier carte dont la clé correspond aux modalités Clé de données.



Diagrammes en secteurs des effectifs d'une carte
Affiche la proportion de lignes/observations dans chaque modalité d'un champ qualitatif (Modalités) pour chaque fonctionnalité de carte (Clé de données) et dessine une carte et les proportions sous la forme d'un diagramme en secteurs au centre de chaque fonctionnalité de carte.

Requiert : Une paire de champs qualitatifs. Un fichier carte dont la clé correspond aux modalités Clé de données.

Diagramme en secteurs sur une carte
Calcule la somme d'un champ continu (Valeurs) dans chaque modalité d'un champ qualitatif (Modalités) pour chaque fonctionnalité de carte (Clé de données) et dessine une carte et les sommes sous la forme d'un diagramme en secteurs au centre de chaque fonctionnalité de carte.

Requiert : Une paire de champs qualitatifs et un champ continu. Un fichier carte dont la clé correspond aux modalités Clé de données.

Diagramme curviligne sur une carte
Calcule une statistique récapitulative pour un champ continu (Y) pour chaque valeur d'un autre champ (X) pour chaque fonctionnalité de carte (Clé de données) et dessine une carte et des diagrammes curvilignes qui relient les valeurs au centre de chaque fonctionnalité de carte.

Requiert : Un champ qualitatif et une paire de champs de n'importe quel type. Un fichier carte dont la clé correspond aux modalités Clé de données.

Coordonnées sur une carte de référence
Dessine une carte et des points à l'aide des champs continus (Longitude et Latitude) qui identifient les coordonnées pour ces points.

Requiert : Une paire de champs d'intervalle. Un fichier carte.

Flèches sur une carte de référence
Dessine une carte et des flèches à l'aide de champs continus qui identifient les points de départ (Long de départ et Lat de départ) les points d'arrivée (Long de fin et Lat de fin) pour chaque flèche. Chaque enregistrement/observation dans les résultats de données dans une flèche de la carte.

Requiert : Quatre champs continus. Un fichier carte.

Carte de superposition de points
Dessine une carte de référence et la superpose à une autre carte à points avec les points en couleur en fonction du champ qualitatif (Couleur).

Requiert : Une paire de champs qualitatifs. Un fichier carte à points dont la clé correspond aux modalités Clé de données. Un fichier carte de référence.

Carte superposée polygone
Dessine une carte de référence et la superpose à une autre carte polygone avec les polygones en couleur en fonction du champ qualitatif (Couleur).

Requiert : Une paire de champs qualitatifs. Un fichier carte polygone dont la clé correspond aux modalités Clé de données. Un fichier carte de référence.

Carte de superposition de ligne
Dessine une carte de référence et la superpose à une autre carte curviligne avec les lignes en couleur en fonction du champ qualitatif (Couleur).

Requiert : Une paire de champs qualitatifs. Un fichier carte curviligne dont la clé correspond aux modalités Clé de données. Un fichier carte de référence.

Création de visualisations de carte

Pour de nombreuses visualisations, vous n'avez que deux choix à faire : les champs (variables) souhaités et un modèle pour visualiser ces champs. Aucun choix ou action supplémentaire n'est nécessaire. Les visualisations de carte nécessitent au moins une étape supplémentaire : sélectionnez un fichier carte qui définit les informations géographiques pour la visualisation de carte.

Les étapes de base pour créer une carte simple sont les suivantes :

- Sélectionnez les champs souhaités dans l'onglet de base. Pour des informations sur le type et le nombre de champs nécessaires aux différentes visualisations de carte, consultez Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.
- Sélectionnez un modèle de carte.
- Cliquez sur l'onglet Détaillé.
- Vérifiez que la Clé de données et les autres listes déroulantes nécessaires sont définies sur les champs appropriés.
- Dans le groupe Fichiers cartes, cliquez sur Sélectionner un fichier carte.
- Utilisez la boîte de dialogue Sélectionner les cartes pour choisir le fichier carte et la clé de carte. Les valeurs de la clé de carte doivent correspondre aux valeurs du champ spécifié par la clé de données. Vous pouvez utiliser le bouton Comparer pour comparer ces valeurs. Si vous sélectionnez le modèle de carte superposée, vous aurez également besoin de choisir une carte de référence. La carte de référence ne se trouve pas dans les données. Elle sert d'arrière-plan pour la carte principale. Pour des informations supplémentaires sur la boîte de dialogue Sélectionner les cartes, consultez Sélection des fichiers cartes pour les visualisations de carte sur p. 261.
- Cliquez sur OK pour fermer la boîte de dialogue Sélectionner les cartes.
- Dans le Sélecteur de modèles de représentations graphiques, cliquez sur Exécuter pour créer la visualisation de carte.

Représentation graphique - Exemples

Cette section comporte plusieurs exemples différents pour faire une démonstration des options disponibles. Les exemples présentent également des informations relatives à l'interprétation des visualisations finales.

Ces exemples utilisent le flux intitulé *graphboard.str* qui fait référence aux fichiers de données *employee_data.sav*, *customer_subset.sav* et *worldsales.sav*. Ces fichiers sont disponibles à partir du dossier *Demos* de n'importe quelle installation du client IBM® SPSS® Modeler. Ils sont accessibles depuis le groupe de programmes SPSS Modeler dans le menu Démarrer de Windows. Le fichier *graphboard.str* se trouve dans le dossier *streams*.

Nous vous conseillons de lire les exemples dans leur ordre de présentation. Les exemples postérieurs s'appuient sur les précédents.

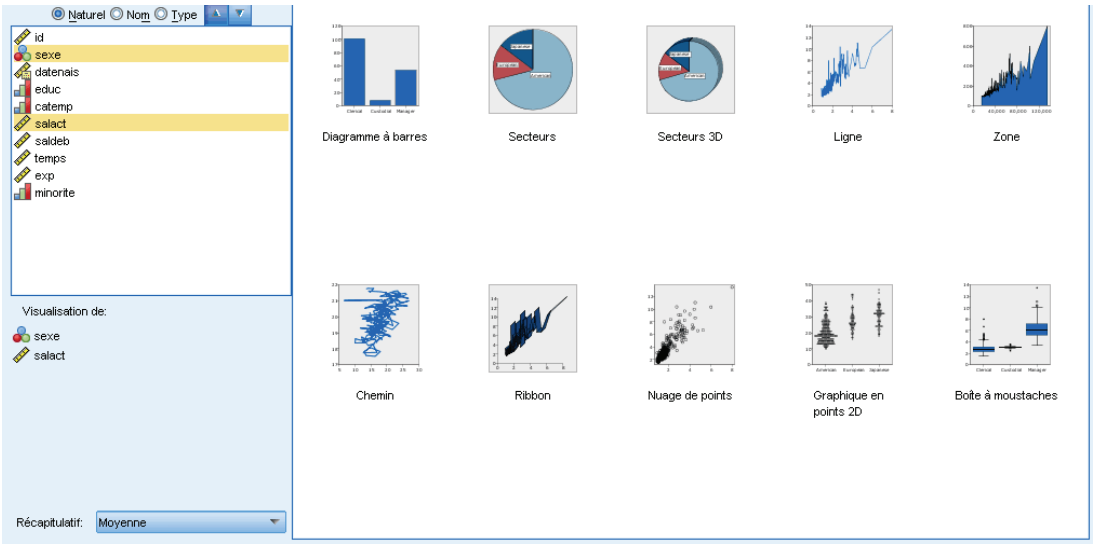
Exemple : Diagramme en bâtons avec statistique récapitulative

Nous allons créer un diagramme en bâtons qui résume un champ ou une variable numérique continue pour chaque catégorie d'une variable d'ensemble/catégorielle. En particulier, nous allons créer un diagramme en bâtons qui indique le salaire moyen des hommes et des femmes.

Cet exemple ainsi que plusieurs des exemples suivants utilisent *Employee data*, un ensemble de données fictif contenant des informations sur les employés d'une société.

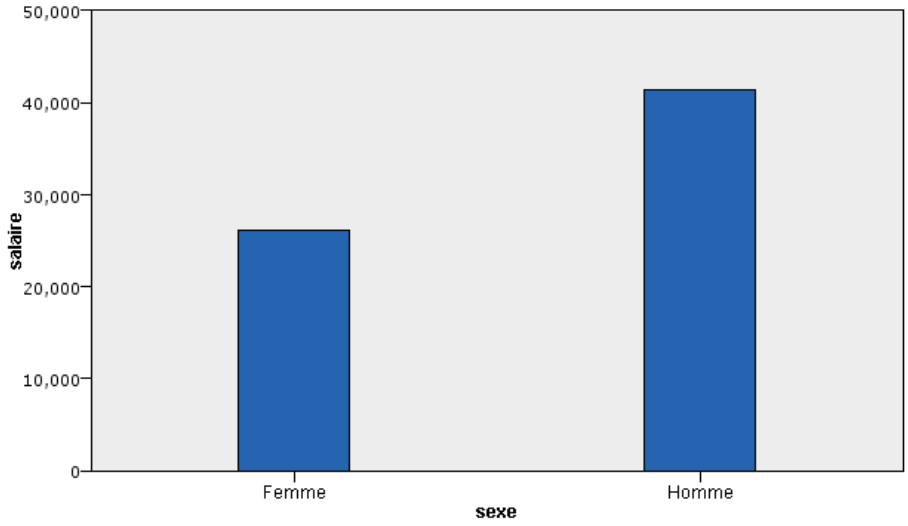
- Ajoutez un noeud source Statistics qui pointe vers *employee_data.sav*.
- Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- Dans l'onglet Base, sélectionnez *Sexe* et *Salaire actuel*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
- Sélectionnez Barres.
- Dans la liste déroulante Récapitulatif, sélectionnez Moyenne.

Figure 5-11
Sélections de l'onglet Base, diagramme en bâtons avec statistique récapitulative



- Cliquez sur Exécuter.
- Sur l'écran qui s'affiche, cliquez sur le bouton « Afficher les étiquettes de champ et de valeur » de la barre d'outils (le second bouton du groupe de deux situé au centre de la barre d'outils).

Figure 5-12
Diagramme en bâtons avec statistique récapitulative



Nous pouvons remarquer que :

- En fonction de la hauteur des barres, il est clair que le salaire moyen des hommes et supérieur au salaire moyen des femmes.

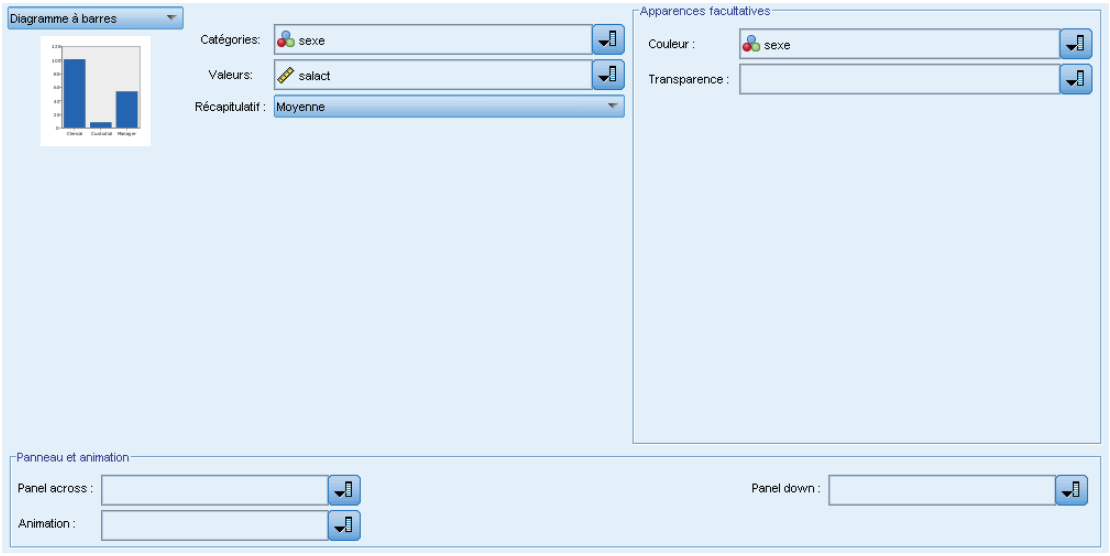
Exemple : Diagramme à barres groupé en classes avec statistique récapitulative

Nous allons maintenant créer un graphique en bâtons groupé en classes (bâtons juxtaposés) pour voir si la différence de salaire moyen entre les hommes et les femmes dépend du type d'emploi. Les femmes gagnent peut-être plus que les hommes, en moyenne, pour certains types d'emploi.

Remarque : Cet exemple utilise *Employee data*.

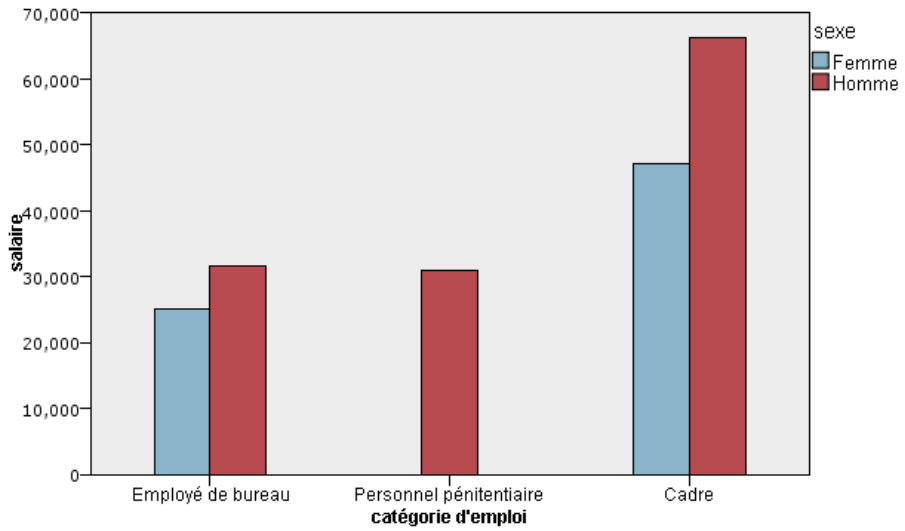
- Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- Dans l'onglet Base, sélectionnez *Catégorie d'emploi* et *Salaire actuel*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
- Sélectionnez Barres.
- Dans la liste Récapitulatif, sélectionnez Moyenne.
- Cliquez sur l'onglet Détaillé. Remarque : vos sélections sur l'onglet précédent se reflètent ici.
- Dans le groupe Apparences en option, choisissez *sexe* dans la liste déroulante Couleur.

Figure 5-13
Sélections de l'onglet Détaillé, diagramme en bâtons groupé en classes



▶ Cliquez sur Exécuter.

Figure 5-14
Diagramme en bâtons juxtaposés



Nous pouvons remarquer que :

- La différence entre les salaires moyens pour chaque type d'emploi ne semble pas aussi importante qu'elle l'était dans le diagramme en bâtons qui comparait les salaires moyens de tous les hommes et femmes. Peut-être y a-t-il un nombre variable d'hommes et de femmes dans chaque groupe. Vous pourriez le vérifier en créant un diagramme en bâtons de nombres.
- Quelque soit le type d'emploi, le salaire moyen des hommes est toujours supérieur au salaire moyen des femmes.

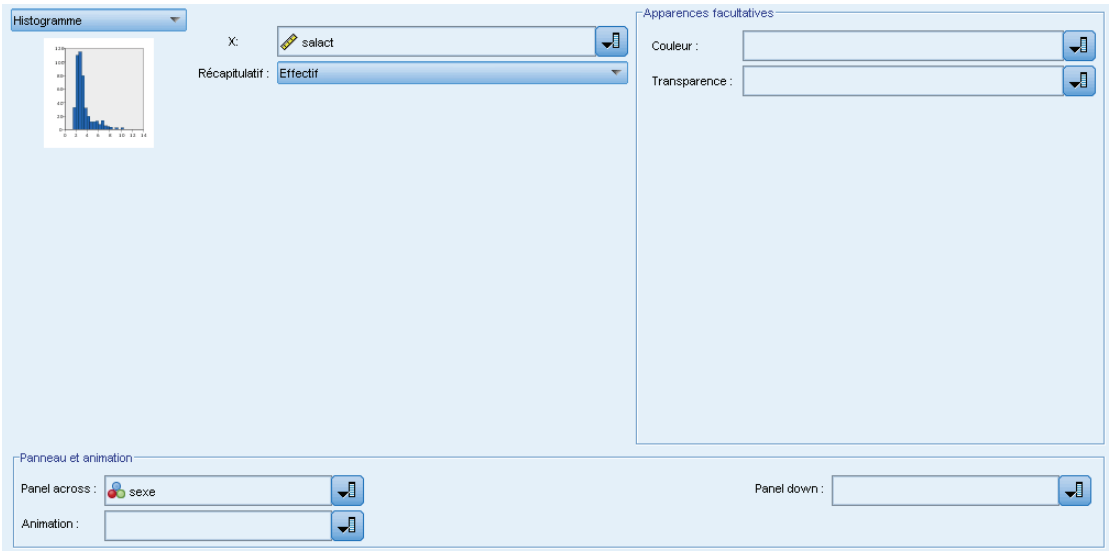
Exemple : Histogramme panélisé

Nous allons créer un histogramme panélisé par sexe afin de pouvoir comparer les distributions des fréquences de salaire pour les hommes et les femmes. La distribution des fréquences indique combien d'observations/lignes se trouvent à l'intérieur d'intervalles de salaire spécifiques. L'histogramme panélisé peut nous aider à analyser plus en détails la différence de salaire entre les sexes.

Remarque : Cet exemple utilise *Employee data*.

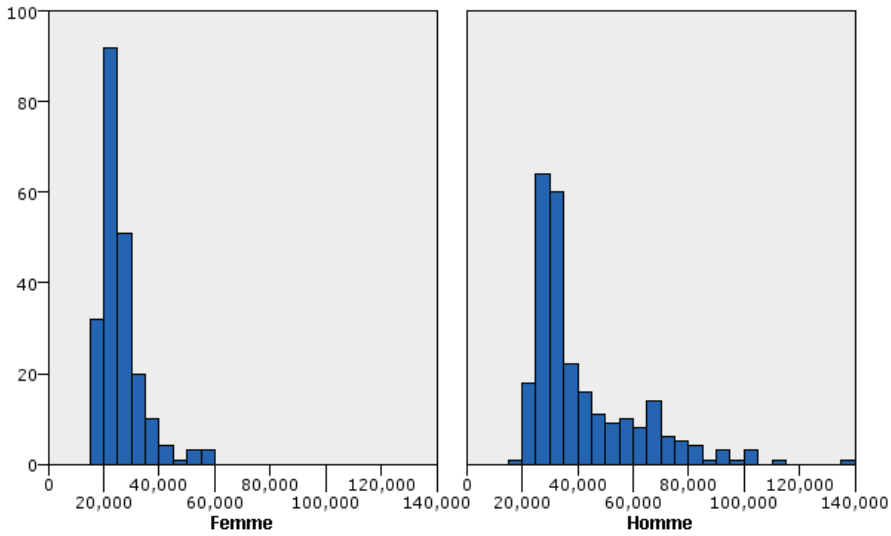
- Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- Dans l'onglet Base, sélectionnez *Salaire actuel*.
- Sélectionnez Histogramme.
- Cliquez sur l'onglet Détaillé.
- Dans le groupe Panneaux et Animation, choisissez *sexe* dans la liste déroulante Panel Across.

Figure 5-15
Sélections de l'onglet Détaillé, histogramme panélisé



- Cliquez sur Exécuter.

Figure 5-16
Histogramme panélisé



Nous pouvons remarquer que :

- Aucune distribution de fréquences n'est une distribution normale. Autrement dit, les histogrammes ne ressemblent pas à des courbes en cloche, comme ce serait le cas si les données étaient distribuées normalement.
- Les barres les plus hautes sont situées sur le côté gauche de chaque graphique. Par conséquent, les hommes comme les femmes sont plus nombreux à avoir des salaires plus bas que des salaires plus élevés.
- Les distributions des fréquences de salaire parmi les hommes et les femmes ne sont pas égales. Observez la forme des histogrammes. Il y a plus d'hommes à avoir des salaires plus élevés que de femmes à avoir des salaires plus élevés.

Exemple : Graphique en points panélisé

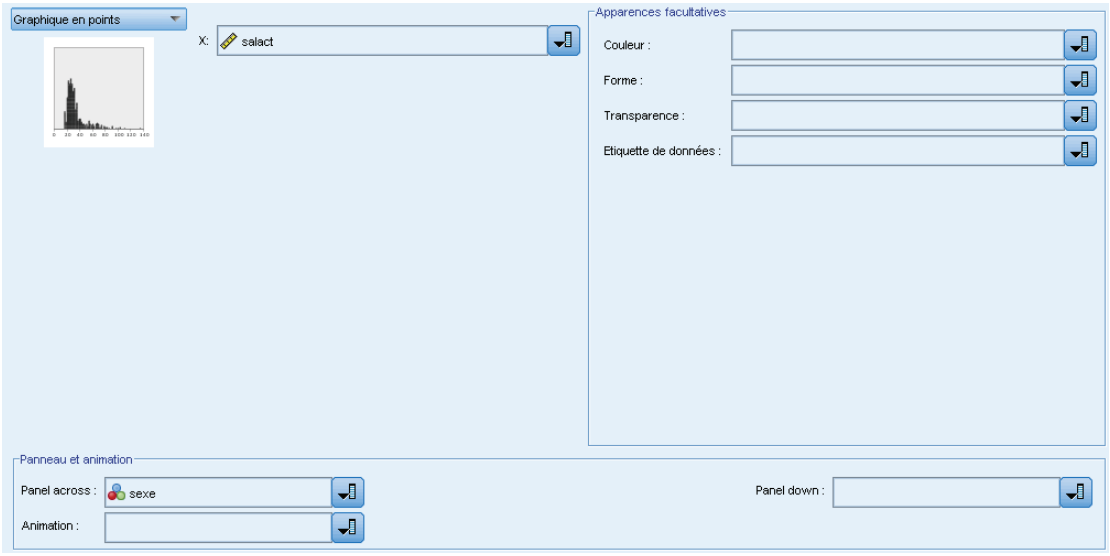
De même qu'un histogramme, un graphique en points indique la distribution d'un intervalle numérique continu. A la différence d'un histogramme, qui indique les nombres d'intervalles de données mis en intervalles, un graphique en points montre toutes les lignes/observations contenues dans les données. Par conséquent, un graphique en points offre une granularité supplémentaire comparé à l'histogramme. En fait, l'utilisation d'un graphique en points peut constituer le point de départ privilégié lors de l'analyse des distributions des fréquences.

Remarque : Cet exemple utilise *Employee data*.

- ▶ Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- ▶ Dans l'onglet Base, sélectionnez *Salaire actuel*.
- ▶ Sélectionnez Graphique en points.

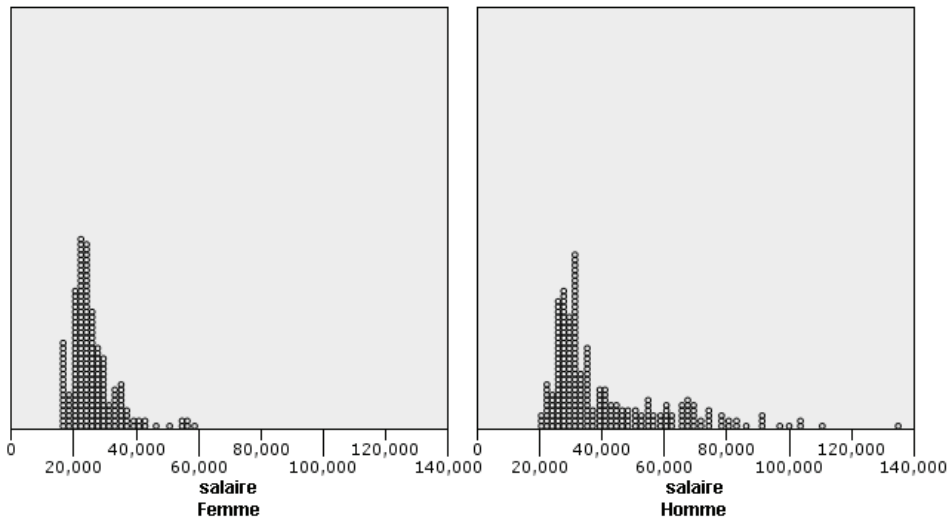
- Cliquez sur l'onglet Détaillé.
- Dans le groupe Panneaux et Animation, choisissez *sexe* dans la liste déroulante Panel Across.

Figure 5-17
Sélections de l'onglet Détaillé, graphique en points panélisé



- Cliquez sur Exécuter.
- Agrandissez la fenêtre de résultats afin de voir le diagramme plus distinctement.

Figure 5-18
Graphique en points panélisé



Comparé à l'histogramme (voirExemple : Histogramme panélisé sur p. 274), nous pouvons observer les éléments suivants :

- Le plus haut niveau à 20 000 qui apparaissait dans l'histogramme pour les femmes est moins sensible dans le graphique à points. Il existe plusieurs observations/lignes concentrées autour de cette valeur, mais la plupart de ces valeurs sont plus proches de 25 000. Ce niveau de granularité n'est pas apparent dans l'histogramme.
- Bien que l'histogramme des hommes suggère que le salaire moyen des hommes diminue progressivement après 40 000, le graphique à points montre que la distribution est plutôt uniforme après cette valeur, jusqu'à 80 000. A chaque valeur de salaire dans cet intervalle, il existe trois hommes ou plus qui gagnent ce salaire en particulier.

Exemple : Boîtes à moustaches

Une boîte à moustaches est un autre graphique utile pour la visualisation de la distribution des données. Une boîte à moustaches contient plusieurs mesures statistiques que nous allons explorer après avoir créé la visualisation.

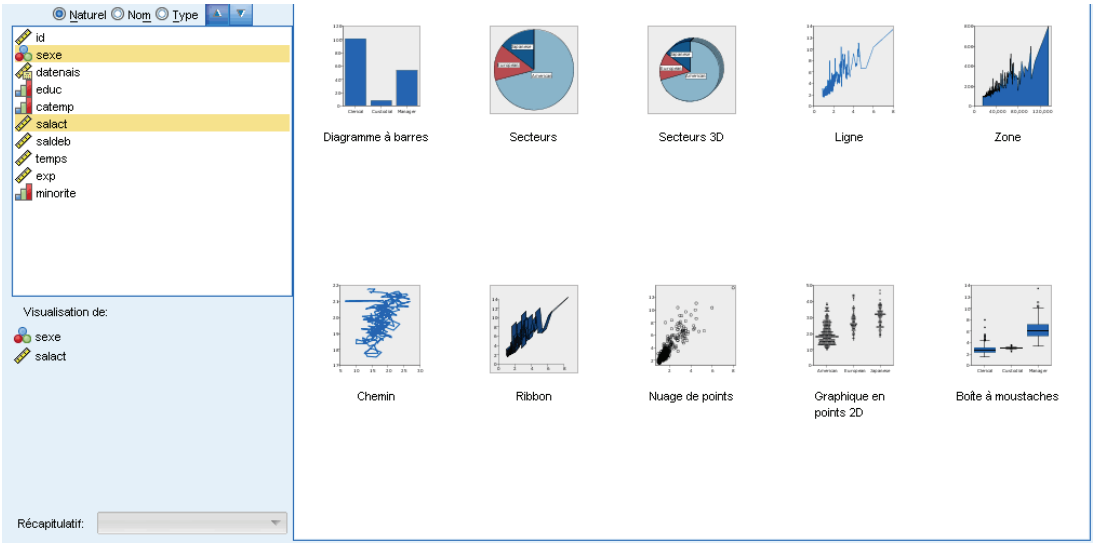
Remarque : Cet exemple utilise *Employee data*.

▶ Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.

▶ Dans l'onglet Base, sélectionnez *Sexe* et *Salaire actuel*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)

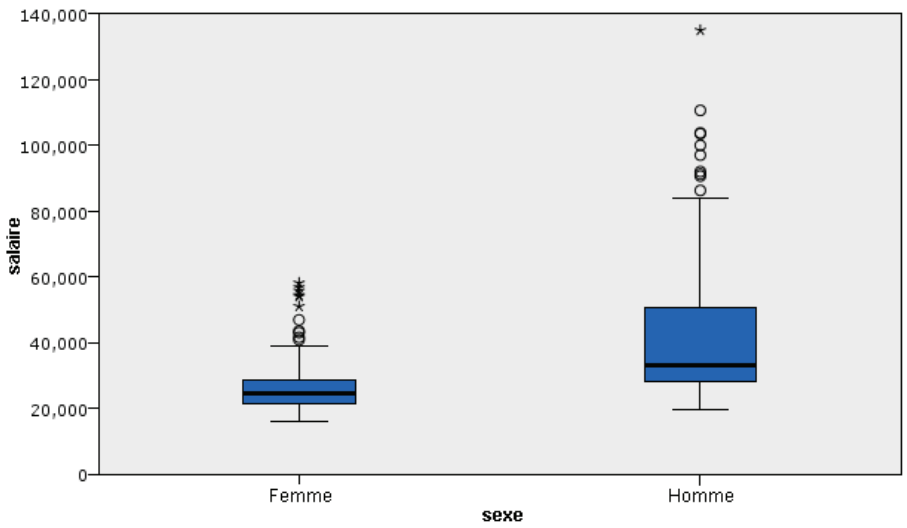
▶ Sélectionnez Boîte à moustaches.

Figure 5-19
Sélections de l'onglet Base, diagramme à surfaces



▶ Cliquez sur Exécuter.

Figure 5-20
Boîtes à moustaches



Etudions les différentes parties de la boîte à moustaches :

- La ligne sombre au milieu des boîtes est la médiane du *salaire*. La moitié des observations /lignes a une valeur supérieure à la médiane, et la moitié a une valeur inférieure. Comme la moyenne, la médiane est une mesure de la tendance centrale. Contrairement à la moyenne, elle est moins influencée par les observations/lignes avec des valeurs extrêmes. Dans cet exemple, la médiane est inférieure à la moyenne (comparez avec Exemple : Diagramme en bâtons avec statistique récapitulative sur p. 271). Le fait que la moyenne et la médiane soient différentes indique qu'il existe quelques observations/lignes avec des valeurs extrêmes qui élèvent la moyenne. Autrement dit, il existe quelques employés qui gagnent des salaires élevés.
- Le bas de la boîte indique le 25ème centile. Vingt-cinq pour cent des observations/lignes ont des valeurs au-dessous du 25ème centile. Le haut de la boîte indique le 75ème centile. Vingt-cinq pour cent des observations/lignes ont des valeurs au-dessus du 75ème centile. Cela signifie que 50 % des observations/lignes sont situées dans la boîte. La boîte est beaucoup plus petite pour les femmes que pour les hommes. Ceci indique que le *salaire* varie moins pour les femmes que pour les hommes. Le haut et le bas de la boîte sont souvent appelés **charnières**.
- Les barres en T qui partent des boîtes sont appelées **limites internes** ou **moustaches**. Elles s'étendent jusqu'à 1,5 fois la hauteur de la zone ou, si aucune observation/ligne n'a une valeur comprise dans cet intervalle, jusqu'aux valeurs minimum ou maximum. Si les données sont distribuées normalement, environ 95 % des données doivent être situées entre les limites internes. Dans cet exemple, les limites internes sont moins étendues pour les femmes que pour les hommes, ce qui constitue une autre indication que le *salaire* varie moins pour les femmes que pour les hommes.
- Les points sont des **valeurs éloignées**. Il s'agit de valeurs qui n'entrent pas dans les limites internes. Les valeurs éloignées sont des valeurs extrêmes. Les astérisques ou les étoiles sont des **valeurs éloignées extrêmes**. Elles représentent des observations/lignes qui ont des valeurs égales à plus de trois fois la hauteur des boîtes. Il existe plusieurs valeurs éloignées

pour les femmes et les hommes. Souvenez-vous que la moyenne est supérieure à la médiane. La moyenne est plus élevée du fait de ces valeurs éloignées.

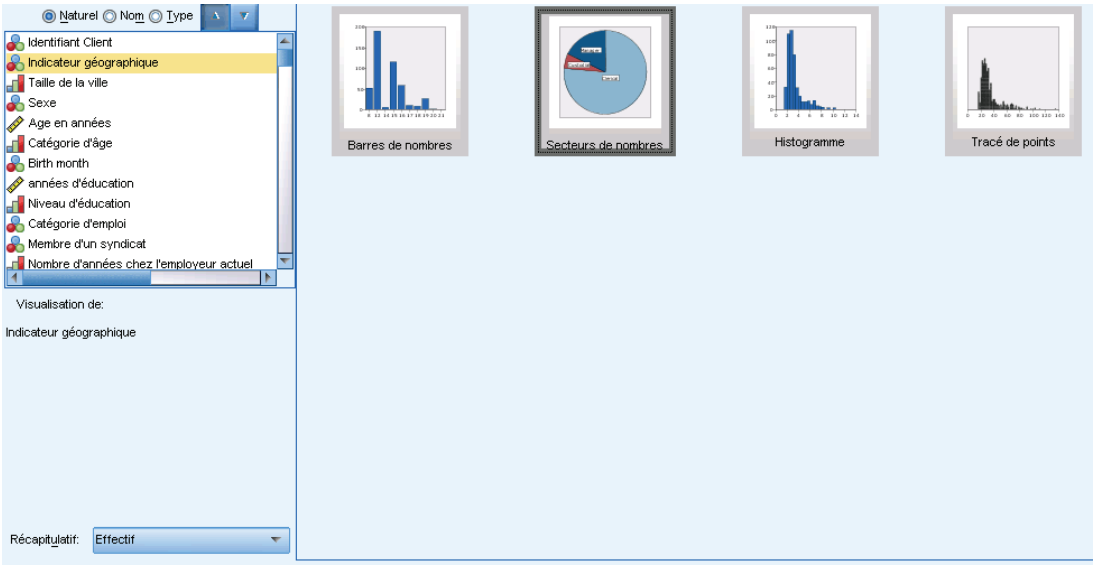
Exemple : Diagramme en secteurs

Nous allons désormais utiliser un ensemble de données différent pour explorer d'autres types de visualisation. L'ensemble de données est *customer_subset*, un fichier de données fictif contenant des informations sur des clients.

Nous allons d'abord créer un diagramme en secteurs pour vérifier la proportion de clients résidant dans différentes régions géographiques.

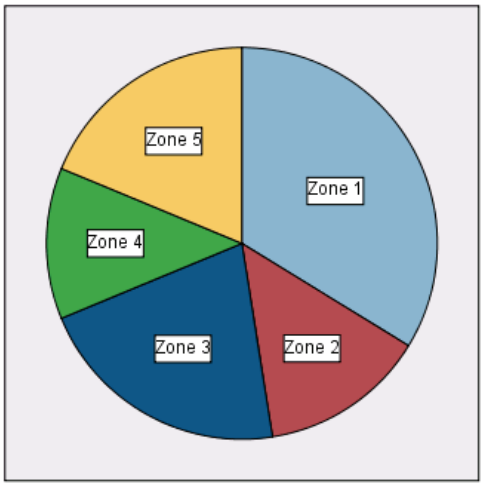
- Ajoutez un noeud source Statistics qui pointe vers *customer_subset.sav*.
- Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- Dans l'onglet Base, sélectionnez *Indicateur géographique*.
- Sélectionnez Secteurs de nombres.

Figure 5-21
Sélections de l'onglet Base, graphique sectoriel



- Cliquez sur Exécuter.

Figure 5-22
Diagramme en secteurs



Nous pouvons remarquer que :

- Zone 1 a davantage de clients que chacune des autres zones.
- Les clients sont équitablement répartis entre les autres zones.

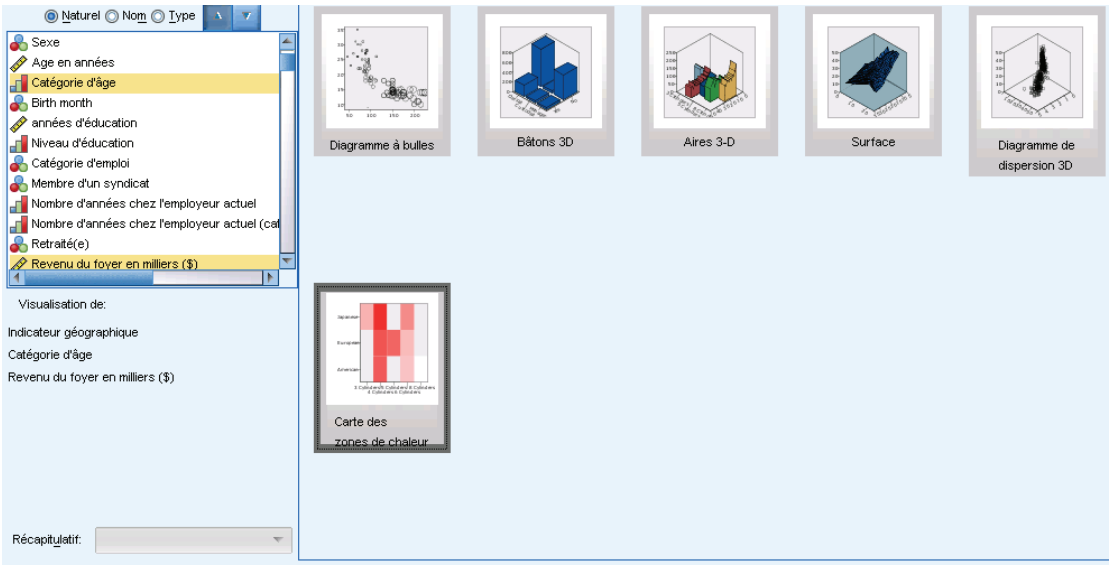
Exemple : Carte thermique

Nous allons maintenant créer une carte de zones de chaleur catégorielle pour vérifier le revenu moyen des clients de différentes régions géographiques et de différentes tranches d'âge.

Remarque : Cet exemple utilise *customer_subset*.

- ▶ Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- ▶ Dans l'onglet Base, sélectionnez *Indicateur géographique*, *Catégorie d'âge* et *Revenu du ménage en milliers*, dans cet ordre. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
- ▶ Sélectionnez Zones de chaleur.

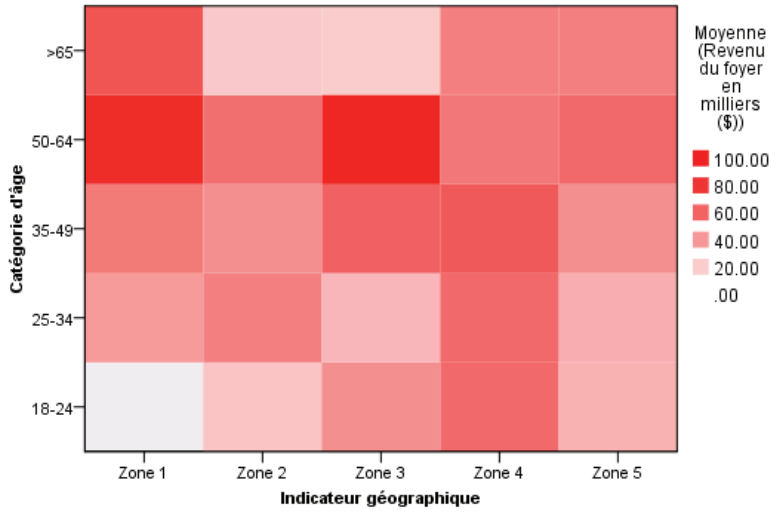
Figure 5-23
Sélections de l'onglet Base, zones de chaleur



▶ Cliquez sur Exécuter.

▶ Sur la fenêtre de résultats, cliquez sur le bouton « Afficher les étiquettes de champ et de valeur » de la barre d'outils (le bouton à droite parmi les deux au centre de la barre d'outils).

Figure 5-24
Carte de zones de chaleur catégorielle



Nous pouvons remarquer que :

- Une carte des zones de chaleur est semblable à une table utilisant des couleurs plutôt que des numéros pour représenter les valeurs des cellules. Le rouge brillant et profond indique la valeur la plus élevée, tandis que le gris indique une valeur basse. La valeur de chaque cellule est la moyenne du champ ou de la variable continue pour chaque paire de catégories.

- Excepté dans les Zones 2 et 5, le groupe de clients situés dans la tranche d'âge 50-64 ans a un revenu moyen du ménage supérieur à ceux des autres groupes.
- Il n'y a pas de clients situés de la tranche d'âge 25-34 ans dans la Zone 4.

Exemple : Matrice de dispersion (SPLOM)

Nous allons créer une matrice de diagramme de dispersion de plusieurs variables différentes afin de déterminer s'il existe des relations entre les variables de l'ensemble de données.

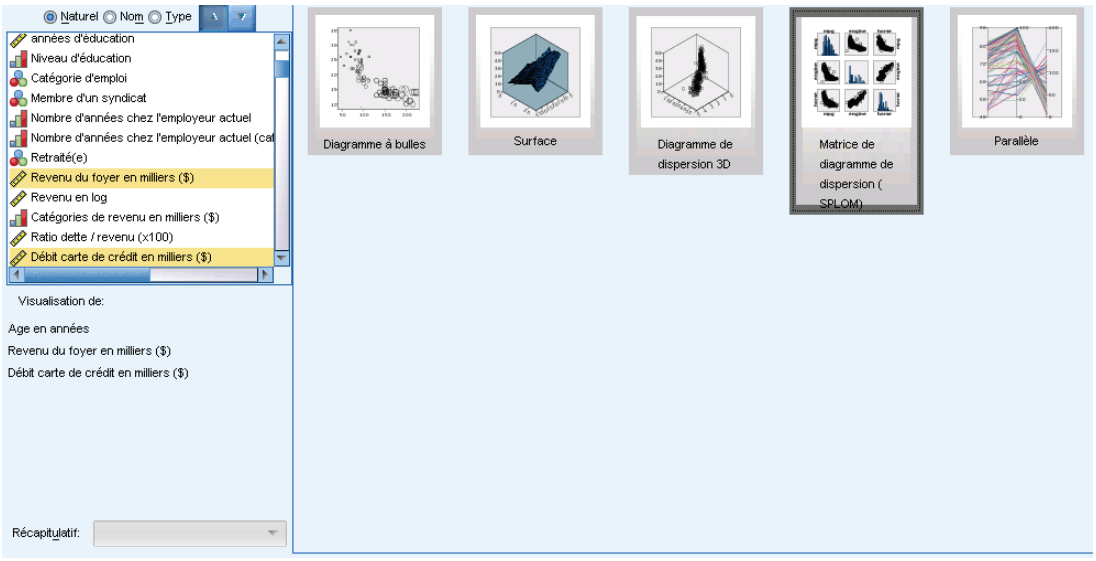
Remarque : Cet exemple utilise *customer_subset*.

▶ Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.

▶ Dans l'onglet Base, sélectionnez *Âge en années*, *Revenu du ménage en milliers* et *Dette de la carte de crédit en milliers*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)

▶ Sélectionnez SPLOM.

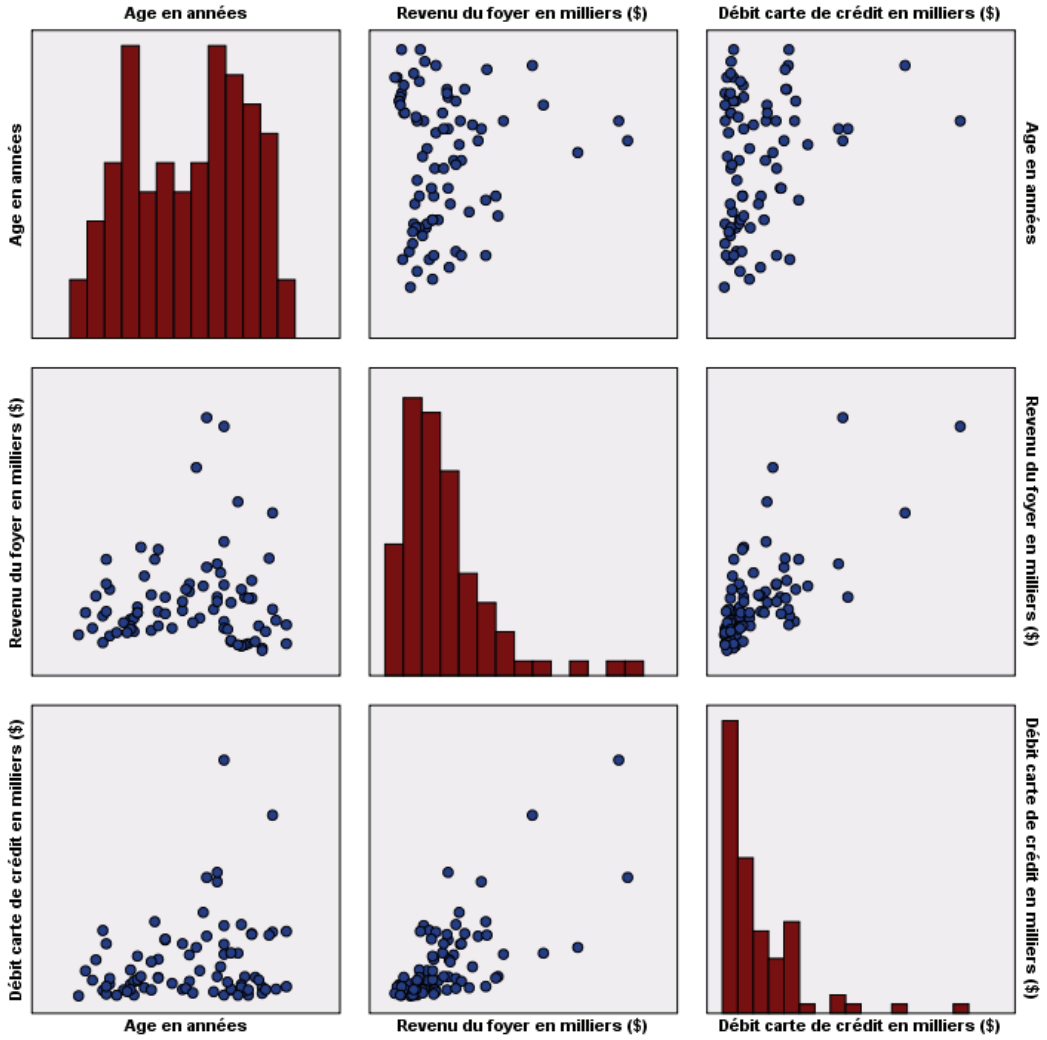
Figure 5-25
Sélections de l'onglet Base, SPLOM



▶ Cliquez sur Exécuter.

▶ Agrandissez la fenêtre de résultats afin de voir la matrice plus distinctement.

Figure 5-26
Matrice de dispersion (SPLOM)



Nous pouvons remarquer que :

- Les histogrammes affichés sur la diagonale montrent la répartition de chaque variable dans la SPLOM. L'histogramme de l'*âge* apparaît dans la cellule en haut à gauche, celui du *revenu* dans la cellule du centre, et celui de la *dettcred* dans la cellule en bas à droite. Aucune des variables ne semble être distribuée normalement. Autrement dit, aucun histogramme ne ressemble à une courbe en cloche. Notez en outre que les histogrammes du *revenu* et de la *dettcred* sont asymétriques.
- Il ne semble pas y avoir de relation entre l'*âge* et les autres variables.
- il existe une relation linaire entre le *revenu* et la *dettcred*. En effet, la *dettcred* augmente à mesure que le *revenu* augmente. Vous pouvez créer des diagrammes de dispersion individuels de ces variables et des autres variables liées pour étudier les relations plus en détails.

Exemple : Choroplèthe (carte de couleur) des sommes

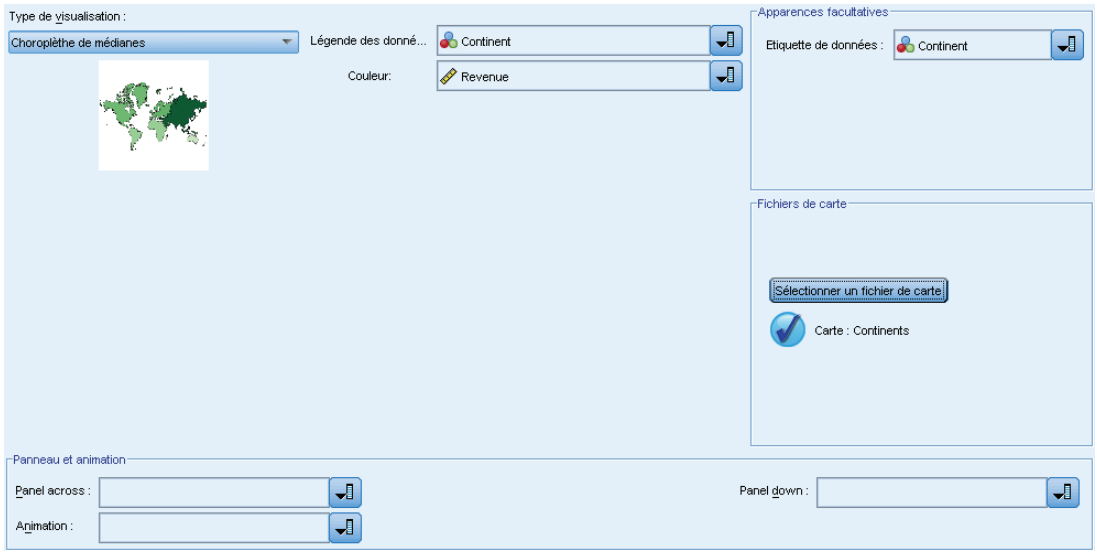
Nous ne créerons pas de visualisation de carte. Par conséquent, dans l'exemple suivant, nous créerons une variation de cette visualisation. L'ensemble de données est *ventes mondiales* qui est un fichier de données hypothétiques qui contient les revenus des ventes par continent et par produit.

- Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
- Dans l'onglet Base, sélectionnez *Continent* et *Revenus*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
- Sélectionnez Choroplèthe des sommes.
- Cliquez sur l'onglet Détaillé.
- Dans le groupe Apparences en option, choisissez *Continent* dans la liste déroulante Etiquetage de données.
- Dans le groupe Fichiers cartes, cliquez sur Sélectionner un fichier carte.
- Dans la boîte de dialogue Sélectionner les cartes, vérifiez que Carte est défini sur *Continents* et Clé de carte est défini sur *CONTINENT*.
- Dans les groupes Comparer la carte et Valeurs de données, cliquez sur Comparer pour vérifier que les clés de carte correspondent aux clés de données. Dans cet exemple, toutes les valeurs de clés de données ont des clés et des fonctionnalités de carte correspondantes. Nous pouvons également voir qu'il n'y a pas de données pour l'Océanie.

Figure 5-27
Boîte de dialogue Sélectionner des cartes

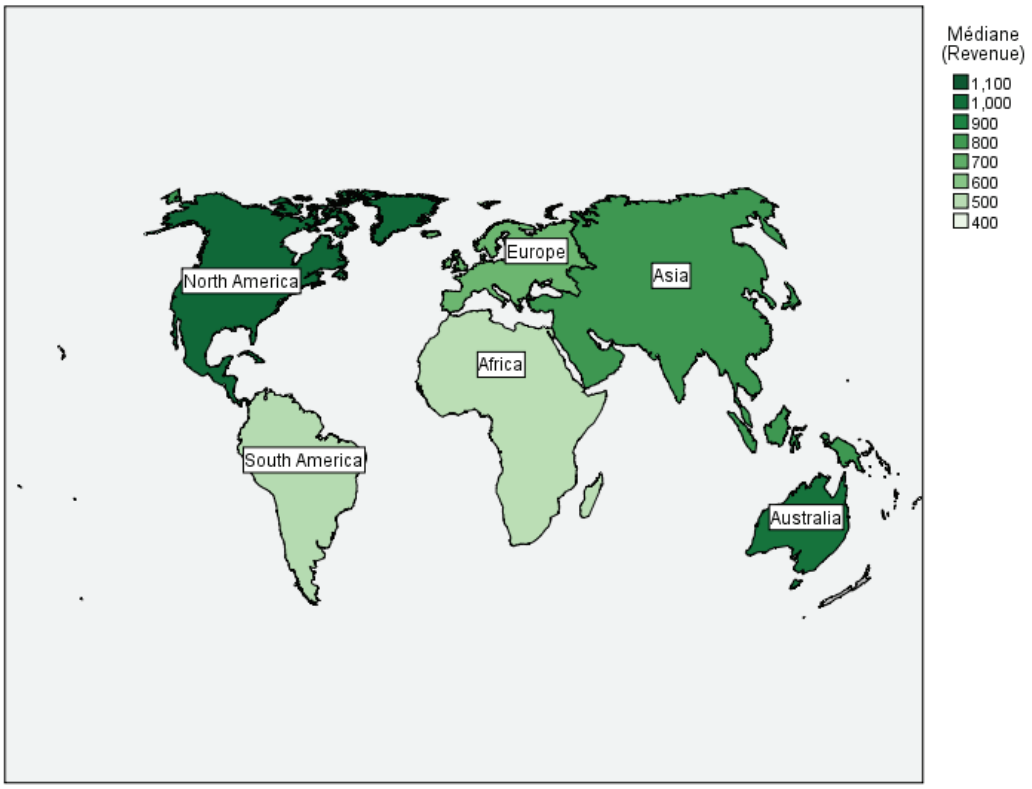
▶ Dans la boîte de dialogue Sélectionner les cartes, cliquez sur OK.

Figure 5-28
Sélections de l'onglet Base, choroplèthe des sommes



▶ Cliquez sur Exécuter.

Figure 5-29
Choroplèthe des sommes



Dans cette visualisation de carte, nous pouvons facilement voir que le revenu est plus élevé en Amérique du Nord qu'en Amérique du Sud et en Afrique. Chaque continent est étiqueté parce que nous avons utilisé *Continent* comme apparence d'étiquette de données.

Exemple : Diagrammes en bâtons sur une carte

Cet exemple montre la façon dont les revenus sont divisés en produit dans chaque continent.

Remarque : Cet exemple utilise *ventes mondiales*.

▶ Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.

▶ Dans l'onglet Base, sélectionnez *Continent*, *Produit* et *Revenus*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)

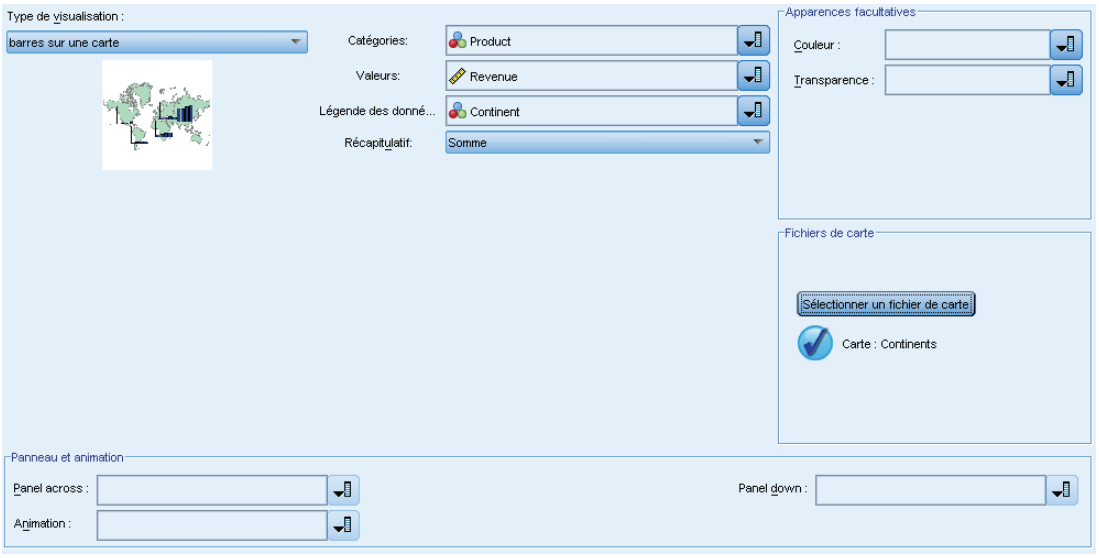
▶ Sélectionnez Bâtons sur une carte.

▶ Cliquez sur l'onglet Détaillé.

Lorsque vous utilisez plusieurs champs de type spécifique, il est important de vérifier que chaque champ est affecté à l'emplacement correspondant.

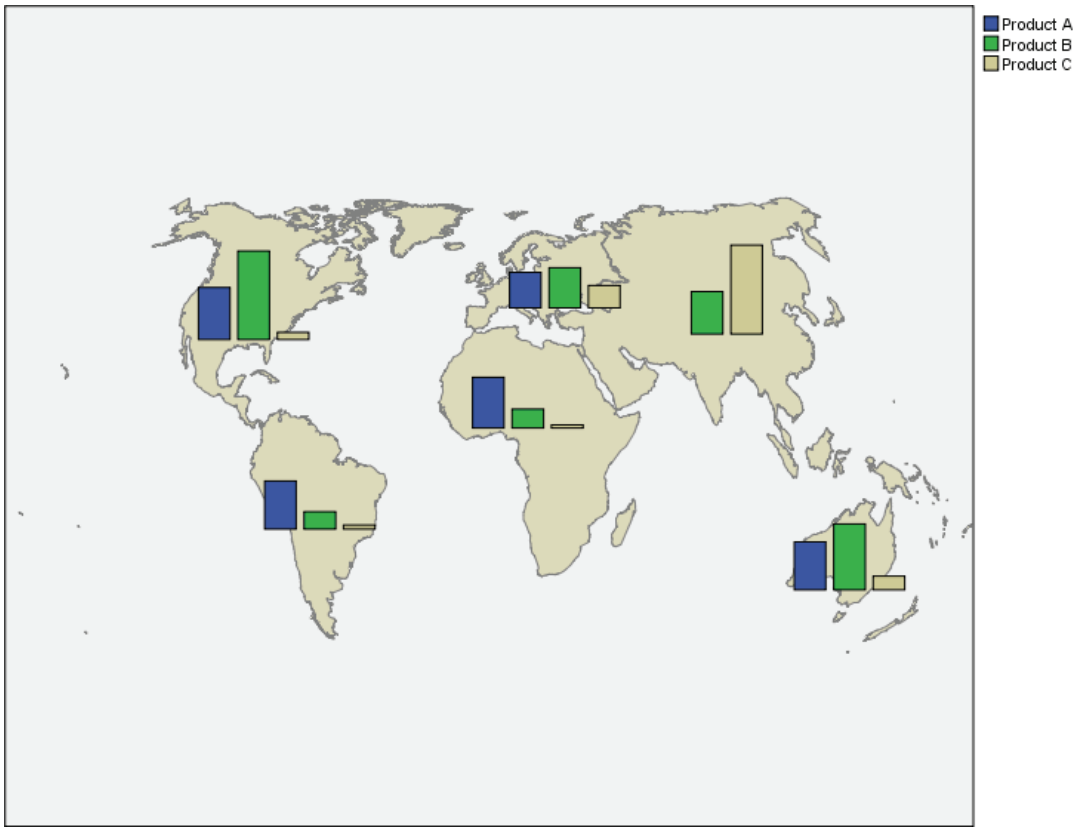
- Sélectionnez *Produit* dans la liste déroulante Catégories.
- Sélectionnez *Revenus* dans la liste déroulante Valeurs.
- Sélectionnez *Continent* dans la liste déroulante Clé de données.
- Dans la liste déroulante Récapitulatif, sélectionnez *Somme*.
- Dans le groupe Fichiers cartes, cliquez sur Sélectionner un fichier carte.
- Dans la boîte de dialogue Sélectionner les cartes, vérifiez que Carte est défini sur *Continents* et Clé de carte est défini sur *CONTINENT*.
- Dans les groupes Comparer la carte et Valeurs de données, cliquez sur Comparer pour vérifier que les clés de carte correspondent aux clés de données. Dans cet exemple, toutes les valeurs de clés de données ont des clés et des fonctionnalités de carte correspondantes. Nous pouvons également voir qu'il n'y a pas de données pour l'Océanie.
- Dans la boîte de dialogue Sélectionner les cartes, cliquez sur OK.

Figure 5-30
Sélections de l'onglet Base, bâtons sur une carte



- Cliquez sur Exécuter.
- Agrandissez la fenêtre de résultats afin de voir l'affichage plus distinctement.

Figure 5-31
Diagrammes en bâtons sur une carte



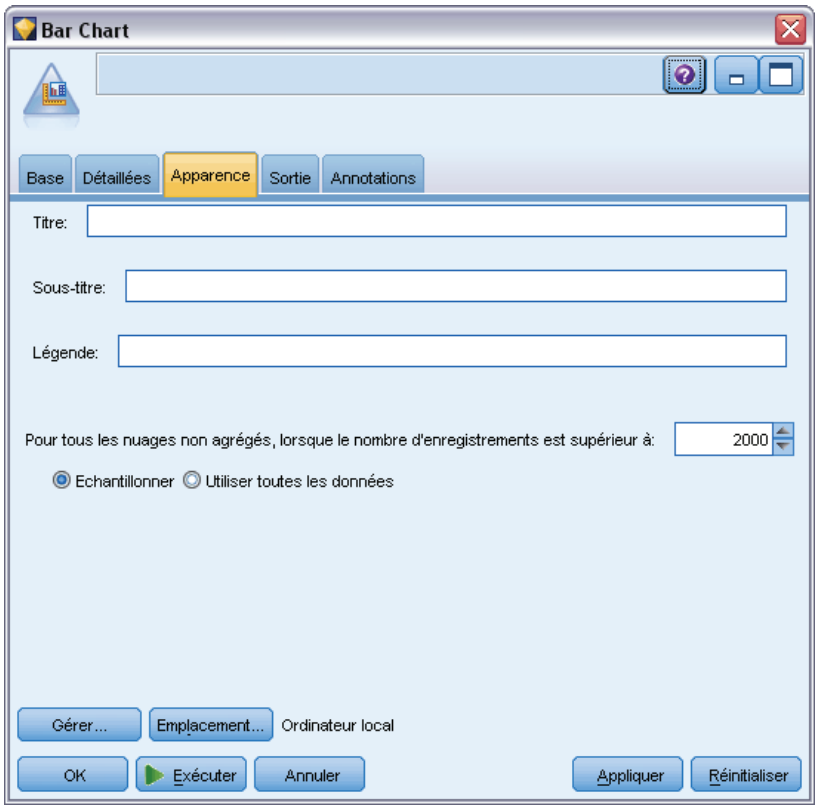
Nous pouvons remarquer que :

- La répartition du total des revenus par produit est à peu près la même en Amérique du Sud et en Afrique.
- *Produit C* génère le revenu le moins élevé partout sauf en Asie.
- Il n'y a pas ou quasi pas de revenu du *Produit A* en Asie.

Onglet Apparence du panneau des représentations graphiques

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-32
Paramètres de l'onglet Apparence pour un noeud Représentation Graphique



Options générales d'apparence

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Echantillonnage. Spécifiez la méthode pour les ensembles de données volumineux. Vous pouvez spécifier le nombre de modalités maximales des ensembles de données ou utiliser le nombre d'enregistrements par défaut. Lorsque vous sélectionnez l'option Echantillon, les performances des ensembles de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant Utiliser toutes les données, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

Options d'apparence des feuilles de style

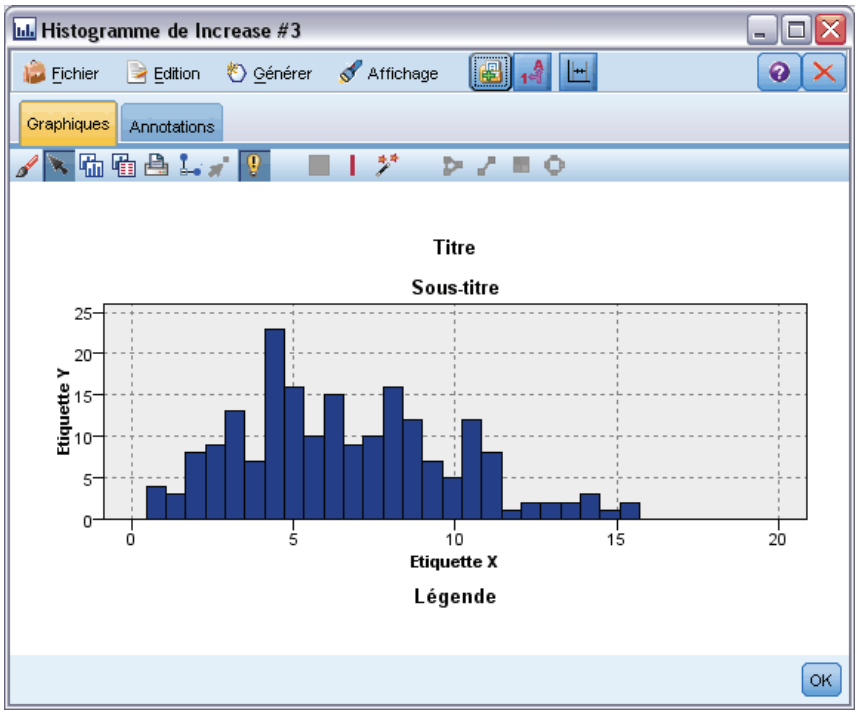
Deux boutons vous permettent de contrôler les modèles de visualisation (et les feuilles de style et les cartes) disponibles :

Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, reportez-vous à la section Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est noté à droite du bouton. Pour plus d'informations, reportez-vous à la section Définition de l'emplacement des modèles, des feuilles de style et des cartes sur p. 290.

L'exemple suivant montre l'emplacement des options d'apparence sur le graphique. (*Remarque* : Tous les graphiques n'utilisent pas toutes ces options.)

Figure 5-33
Position des différentes options d'apparence du graphique



Définition de l'emplacement des modèles, des feuilles de style et des cartes

Les modèles et feuilles de style de visualisation et les fichiers de cartes sont stockés dans un dossier local spécifique ou dans le IBM® SPSS® Collaboration and Deployment Services Repository. Lors de la sélection des modèles, des feuilles de styles et des cartes, seuls les éléments intégrés de cet emplacement sont affichés. En conservant tous les modèles, feuilles de style et cartes à un seul endroit, les applications IBM SPSS peuvent facilement y accéder. Pour plus d'informations sur

l'ajout de modèles, de feuilles de style et de fichiers cartes supplémentaires à cet emplacement, reportez-vous à Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Définir l'emplacement des modèles, des feuilles de style et des fichiers de carte

- Dans une boîte de dialogue de modèle ou de feuille de style, cliquez sur Emplacement... pour afficher la boîte de dialogue Modèles, feuilles de style et cartes.
- Sélectionnez une option pour l'emplacement par défaut des fichiers de modèles, des feuilles de style et des cartes :

  Ordinateur local. Les modèles, les feuilles de style et les fichiers de carte sont situés dans un dossier spécifique sur votre ordinateur local. Sous Windows XP, ce dossier est *C:\Documents and Settings\<utilisateur>\Application Data\SPSSInc\Graphboard*. Le dossier ne peut pas être changé.

  IBM® SPSS® Collaboration and Deployment Services Repository. Les modèles, les feuilles de style et les fichiers de carte sont situés dans un dossier spécifié par l'utilisateur dans le IBM SPSS Collaboration and Deployment Services Repository. Pour identifier le dossier spécifique, cliquez sur Dossier. Pour plus d'informations, consultez Utilisation du IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers de carte. sur p. 291.

- Cliquez sur OK.

Utilisation du IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers de carte.

Les modèles de visualisation et les feuilles de style peuvent être stockés dans le IBM® SPSS® Collaboration and Deployment Services Repository. Cet emplacement est un dossier spécifique dans le IBM SPSS Collaboration and Deployment Services Repository. S'il est défini comme l'emplacement par défaut, tous les modèles, feuilles de style et fichiers de carte situés à cet emplacement sont disponibles pour la sélection.

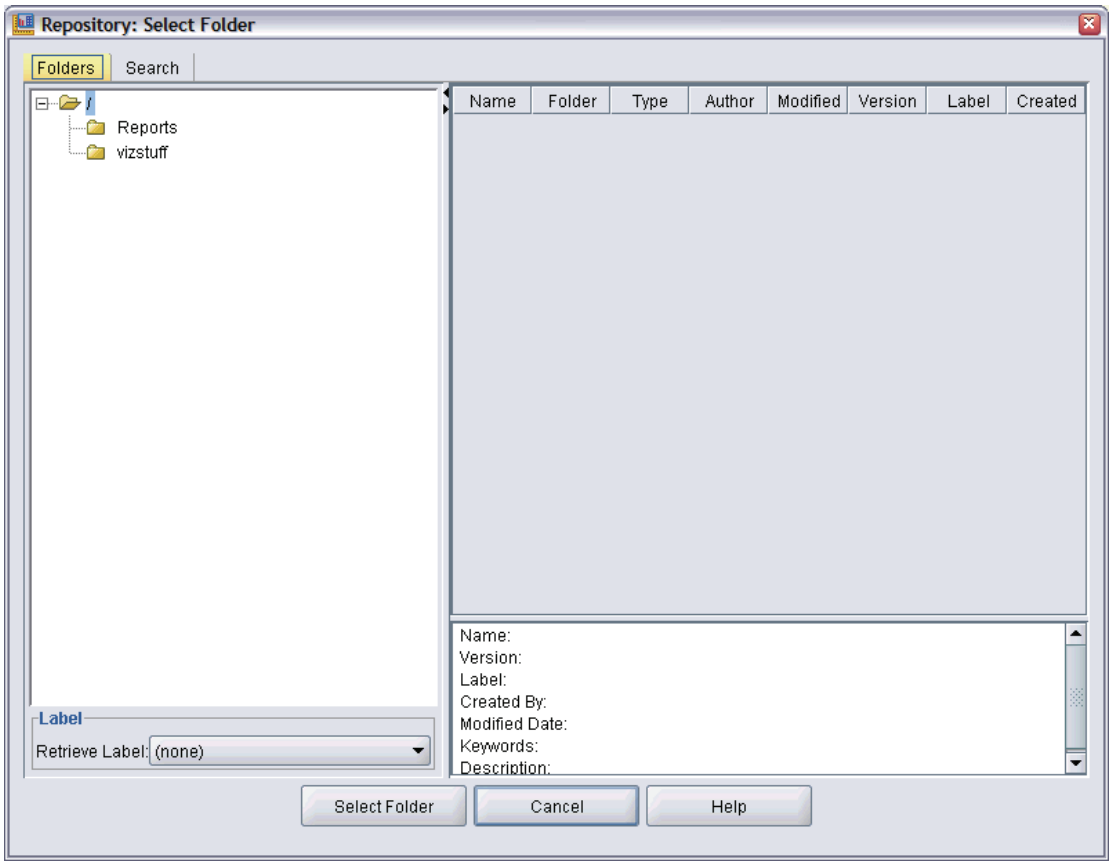
Définir un dossier dans IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers de carte

- Dans une boîte de dialogue comprenant le bouton Emplacement, cliquez sur Emplacement....
- Sélectionnez IBM® SPSS® Collaboration and Deployment Services Repository.
- Cliquez sur Dossier.

  Remarque : Si vous n'êtes pas encore connecté au IBM SPSS Collaboration and Deployment Services Repository, le système vous invite à entrer vos informations de connexion.

- Dans la boîte de dialogue Sélectionner un dossier, sélectionnez le dossier dans lequel les modèles, les feuilles de style et les fichiers de carte sont stockés.

Figure 5-34
Boîte de dialogue Sélectionner un dossier



- Si vous le souhaitez, sélectionnez une étiquette depuis l'option Récupérer l'étiquette. Seuls les modèles, les feuilles de style et les fichiers de carte avec cette étiquette seront affichés.

- Si vous cherchez un dossier qui contient un modèle, une feuille de style ou une carte en particulier, utilisez l'onglet Recherche. La boîte de dialogue Sélectionner un dossier sélectionne automatiquement le dossier dans lequel le modèle, la feuille de style ou le fichier de carte est situé.

- Cliquez sur Sélectionner un dossier.

Gestion des modèles, des feuilles de style et des fichiers cartes

Vous pouvez gérer les modèles, les feuilles de style et les fichiers cartes en local sur votre ordinateur à l'aide de la boîte de dialogue Gérer les modèles, les feuilles de style et les cartes. Cette boîte de dialogue vous permet d'importer, d'exporter, de renommer et de supprimer les modèles de visualisation, les feuilles de style et les fichiers cartes depuis votre ordinateur en local.

- Cliquez sur Gérer... dans l'une des boîtes de dialogue dans laquelle vous sélectionnez les modèles, les feuilles de style ou les cartes.

Gérer la boîte de dialogue des modèles, des feuilles de style et des cartes

L'onglet Modèle répertorie tous les modèles locaux. L'onglet Feuille de style répertorie toutes les feuilles de style locales et affiche des visualisations à partir d'exemples de données. Vous pouvez choisir une des feuilles de style pour l'appliquer aux exemples de visualisation. Pour plus d'informations, reportez-vous à la section Application des feuilles de style sur p. 393. La carte répertorie tous les fichiers cartes locaux. Cet onglet affiche également les clés de carte qui contiennent des exemples de valeur, un commentaire si l'un d'eux a été fourni pendant la création de la carte et un aperçu de la carte.

Les boutons suivants sont disponibles sur l'onglet actif.

Importer. Importer un modèle de visualisation, une feuille de style ou un fichier carte à partir du système de fichiers. L'importation d'un modèle, d'une feuille de style ou d'un fichier carte rend ceux-ci disponibles à l'application IBM SPSS. Si un autre utilisateur vous a envoyé un modèle, une feuille de style ou un fichier carte, vous devez l'importer avant de pouvoir l'utiliser avec votre application.

Exporter. Exporter un modèle de visualisation, une feuille de style ou un fichier carte dans le système de fichiers. Exportez un modèle, une feuille de style ou un fichier carte lorsque vous souhaitez l'envoyer à un autre utilisateur.

Renommer. Renommer le modèle de visualisation, la feuille de style ou le fichier carte sélectionné. Vous ne pouvez pas donner un nom à un modèle déjà utilisé.

Exporter la clé de carte. Exportez les clés de cartes comme un fichier (CSV) de valeurs séparées par des virgules. Ce bouton est uniquement activé dans l'onglet Carte.

Supprimer. Supprimer le(s) modèle(s) de visualisation, la(les) feuille(s) de style ou le(s) fichier(s) carte(s) sélectionné(s). Vous pouvez sélectionner plusieurs modèles, feuilles de style ou fichiers cartes en cliquant sur Ctrl. L'action de suppression ne peut pas être annulée, utilisez-la avec précaution.

Conversion et distribution des fichiers de formes Carte

Le sélectionneur de modèles de représentations graphiques vous permet de créer des visualisations de carte en combinant un modèle de visualisation et un fichier SMZ. Les fichiers SMZ sont semblables aux fichiers de formes ESRI (format de fichier SHP). En effet, ils contiennent des informations géographiques permettant de dessiner une carte (par exemple, les frontières d'un pays) mais il sont optimisés pour les visualisations de carte. Le sélectionneur de modèles de représentations graphiques est préinstallé avec un nombre précis de fichiers SMZ. Si vous possédez un fichier de formes ESRI existant que vous souhaitez utiliser pour les visualisations de cartes, vous devez d'abord convertir le fichier de formes en fichier SMZ à l'aide de l'utilitaire de conversion des cartes. L'utilitaire de conversion des cartes prend en charge les fichiers de formes ESRI (types de formes 1, 3 et 5) avec points, polylignes er polygones contenant une couche unique.

En plus de convertir les fichiers de formes ESRI, l'utilitaire de conversion des cartes vous permet de modifier le niveau de détails des cartes, de modifier les étiquettes des fonctionnalités, de fusionner et de déplacer les fonctionnalités, en autres options. Vous pouvez également utiliser

l'utilitaire de conversion des cartes pour modifier un fichier SMZ existant (y compris ceux déjà installés).

Modification des fichiers SMZ pré-installés

- Exportez le fichier SMZ depuis le système de gestion. Pour plus d'informations, reportez-vous à la section Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.
- Utilisez l'utilitaire de conversion des cartes pour ouvrir et modifier le fichier SMZ exporté. Nous vous recommandons d'enregistrer le fichier sous un nom différent. Pour plus d'informations, reportez-vous à la section Utilisation de l'utilitaire de conversion des cartes sur p. 295.
- Importez le fichier SMZ modifié dans le système de gestion. Pour plus d'informations, reportez-vous à la section Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Ressources supplémentaires pour les fichiers cartes

Les données géospatiales au format de fichier SHP, qui pourraient être utilisées pour prendre en charge vos besoins de mappage, sont disponibles à partir de nombreuses ressources privées et publiques. Vérifiez les sites Web gouvernementaux locaux pour trouver des données gratuites. De nombreux modèles dans ce produit sont basés sur les données publiques obtenues sur GeoCommons (*http://www.geocommons.com*) et auprès du U.S. Census Bureau (*http://www.census.gov*). Une autre source pour les données géospatiales fédérales, régionales et locales américaines est le U.S. Geological Survey (*http://www.geodata.gov*).

REMARQUE IMPORTANTE : les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits. Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel de ces sites Web ne fait pas partie du matériel de ce programme IBM, sauf mention contraire dans un fichier Remarques qui accompagne ce programme IBM et l'utilisation du matériel de ces sites se fait à vos propres risques.

Concepts principaux des cartes

Comprendre certains concepts principaux associés aux fichiers de formes vous aidera à utiliser efficacement l'utilitaire de conversion des cartes.

Un **fichier de formes** propose les informations géographiques permettant de dessiner une carte. Il existe trois types de fichiers de formes que l'utilitaire de conversion des cartes prend en charge :

- **Point.** Le fichier de formes identifie les emplacements des points, tels que les villes.

- **Polyligne.** Le fichier de formes identifie les chemins d'accès et leurs emplacements des points, tels que les rivières.
- **Polygone.** Le fichier de formes identifie les contours des régions et leurs emplacements, tels que les pays.

La plupart du temps, vous utiliserez un fichier de formes polygone. Les cartes choroplèthes sont créées à partir des fichiers de formes polygones. Les cartes choroplèthes utilisent la couleur pour représenter une valeur dans des polygones individuels (régions). Les fichiers de formes avec points et polygones sont généralement superposés sur un fichier de formes polygone. Un exemple en est un fichier de formes avec points des villes américaines superposé sur un fichier de formes polygones des États américains.

Un fichier de formes est composé de **fonctionnalités**. Les fonctionnalités sont des entités géographiques individuelles. Par exemple, les fonctionnalités peuvent être des pays, des États, des villes, etc. Le fichier des formes contient également des données sur les fonctionnalités. Ces données sont stockées dans les **attributs**. Les attributs sont semblables aux champs ou aux variables dans un fichier de données. Il existe toujours au moins un attribut qui est la **clé de carte** de cette fonctionnalité. La clé de carte peut être une étiquette, comme un pays ou le nom d'une région. La clé de carte est ce que vous relierez à une variable/champ dans un fichier de données pour créer une visualisation de carte.

Veuillez noter que vous pourrez uniquement conserver le ou les attributs clés dans le fichier SMZ. L'utilitaire de conversion des cartes ne prend pas en charge l'enregistrement des attributs supplémentaires. Cela signifie que vous devrez créer plusieurs fichiers SMZ si vous souhaitez effectuer une agrégation à différents niveaux. Par exemple, si vous souhaitez agréger les États et les régions américaines, vous aurez besoin de fichiers SMZ distincts. un fichier contenant une clé qui identifie les États et un fichier contenant une clé qui identifie les régions.

Utilisation de l'utilitaire de conversion des cartes

Démarrage de l'utilitaire de conversion des cartes

▶ A partir des menus, sélectionnez :
Outils > Utilitaire de conversion des cartes

Il existe quatre écrans principaux (étapes) dans l'utilitaire de conversion des cartes. Une des étapes inclut également des sous-étapes pour un contrôle plus complet de l'édition des fichiers cartes.

Etape 1 : choisir les fichiers de destination et source

Vous devez d'abord sélectionner un fichier carte source et un fichier de destination pour le fichier carte converti. Vous aurez besoin à la fois du fichier *.shp* et du fichier *.dbf* pour le fichier de formes.

Sélectionnez un fichier .shp (ESRI) ou .smz pour la conversion. Recherchez un fichier carte existant sur votre ordinateur. Il s'agit du fichier que vous convertirez et que vous enregistrerez comme fichier SMZ. Le fichier *.dbf* pour le fichier de formes *doit* être stocké au même emplacement et

avoir un nom de fichier de base qui correspond au fichier *.shp*. Le fichier *.dbf* est nécessaire parce qu'il contient les informations sur les attributs pour le fichier *.shp*.

Définissez une destination et un nom de fichier pour le fichier carte converti. Saisissez un chemin d'accès et un nom de fichier pour le fichier SMZ qui sera créé à partir du fichier source carte d'origine.

Etape 2 : choisir une clé de carte

Vous devez maintenant choisir les clés de carte à inclure au fichier SMZ. Vous pourrez ensuite modifier certaines options qui auront un effet sur la représentation de la carte. Les étapes suivantes de l'utilitaire de conversion des cartes contiennent un aperçu de la carte. Les options de réprésentation que vous choisissez seront utilisées pour générer l'aperçu de la carte.

Choisissez la clé de carte principale. Sélectionnez l'attribut qui sera la clé principale pour identifier et étiqueter les fonctionnalités de la carte. Par exemple, la clé principale d'une carte du monde pourrait être l'attribut identifiant les noms des pays. La clé principale reliera également vos données aux fonctionnalités de la carte. Par conséquent, vérifiez que les valeurs (étiquettes) de l'attribut que vous avez choisies correspondent aux valeurs de vos données. Des exemples d'étiquettes sont affichés lorsque vous choisissez un attribut. Si vous avez besoin de modifier ces étiquettes, vous pourrez le faire dans une autre étape.

Choisir des clés supplémentaires à ajouter. En plus de la clé de carte principale, vérifiez tous les autres attributs clés que vous souhaitez inclure dans le fichier SMZ Généré. Par exemple, certains attributs peuvent contenir des étiquettes traduites. Si vous savez que vos données sont codées dans d'autres langues, il peut être nécessaire de conserver ces attributs. Veuillez noter que vous pouvez uniquement choisir des clés supplémentaires qui représentent les mêmes fonctionnalités que la clé principale. Par exemple, si la clé principale contient les noms complets des États américains, vous pouvez uniquement sélectionner les autres clés qui représentent les États américains, comme les abréviations des États.

Lisser automatiquement la carte. Les fichiers de formes avec des polygones contiennent généralement trop de points de données et trop de détails pour les visualisations de cartes statistiques. Trop de détails peut avoir un effet distrayant et un impact négatif sur les performances. Vous pouvez réduire le niveau des détails et généraliser la carte grâce au lissage. La carte sera plus claire et sera plus rapidement représentée. Lorsque la carte est automatiquement lissée, l'angle maximum est de 15 degrés et le pourcentage à conserver est de 99. Pour des informations sur ces paramètres, consultez Lisser la carte sur p. 297. Veuillez noter que vous avez la possibilité d'appliquer un lissage supplémentaire ultérieurement dans une autre étape.

Supprimer les bordures entre des polygones attenants dans la même fonctionnalité. Certaines fonctionnalités peuvent contenir des sous-fonctionnalités qui contiennent des bordures internes dans les fonctionnalités principales. Par exemple, une carte du monde avec les continents peut contenir des bordures internes des pays qui composent chaque continent. Si vous choisissez cette option, les bordures internes n'apparaîtront pas sur le plan. Dans le cas de la carte du monde avec les continents, choisir cette option supprime les frontières des pays tout en conservant les frontières des continents.

Etape 3 : modifier la carte

Maintenant que vous avez spécifié les options de base de la carte, vous pouvez en ajouter d'autres et les affiner. Ces modifications sont facultatives. Cette étape de l'utilitaire de conversion des cartes vous guide dans les tâches associées et affiche un aperçu de la carte pour que vous puissiez vérifier vos modifications. Certaines tâches peuvent ne pas être disponibles en fonction du type du fichier de formes (avec points, polylignes ou polygones) et du système de coordonnées.

Chaque tâche possède les commandes communes suivantes à gauche de l'utilitaire de conversion des cartes.

Afficher les étiquettes sur la carte. Par défaut, les étiquettes des fonctionnalités n'apparaissent pas dans l'aperçu. Vous pouvez choisir d'afficher les étiquettes. Bien que les étiquettes puissent aider à identifier les fonctionnalités, elles peuvent interférer avec la sélection directe sur l'aperçu de la carte. Activez cette option lorsque vous en avez besoin, par exemple lorsque vous modifiez les étiquettes des fonctionnalités.

Colorer l'aperçu de la carte. Par défaut, l'aperçu de la carte affiche des zones avec une couleur solide. Toutes les fonctionnalités sont de la même couleur. Vous pouvez choisir d'avoir un assortiment de couleurs attribuées aux fonctionnalités individuelles de la carte. Cette option peut aider à distinguer différentes fonctionnalités de la carte. Cela est particulièrement utile lorsque vous fusionnez des fonctionnalités et que vous souhaitez voir ces nouvelles fonctionnalités représentées dans l'aperçu.

Chaque tâche possède également la commande commune suivante à droite de l'utilitaire de conversion des cartes.

Annuler. Si vous effectuez une modification non désirée, cliquez sur Annuler pour revenir à l'état précédent. Vous pouvez annuler un maximum de 100 modifications.

Lisser la carte

Les fichiers de formes avec des polygones contiennent généralement trop de points de données et trop de détails pour les visualisations de cartes statistiques. Trop de détails peut avoir un effet distrayant et un impact négatif sur les performances. Vous pouvez réduire le niveau des détails et généraliser la carte grâce au lissage. La carte sera plus claire et sera plus rapidement représentée. Cette option n'est pas disponible pour les cartes avec points et polygones.

Angle max. L'angle maximum, qui doit se situer entre une valeur de 1 et 20, spécifie la tolérance pour lisser les ensembles de points qui sont presque linéaires. Une valeur plus importante permet une tolérance plus élevée pour le lissage linéaire et donnera plus de points et par conséquent, une carte plus généralisée. Pour appliquer un lissage linéaire, l'utilitaire de conversion des cartes vérifie l'angle interne formé par chaque ensemble de trois points sur la carte. Si 180 moins l'angle est inférieur à la valeur spécifiée, l'utilitaire de conversion des cartes ignore le point central. Par exemple, l'utilitaire de conversion des cartes vérifie si la ligne formée par les trois points est presque droite. Si c'est le cas, l'utilitaire de conversion des cartes considère la ligne comme une ligne droite entre les points terminaux et ignore le point central.

Pourcentage à conserver. Le pourcentage à conserver, qui doit être une valeur comprise entre 90 et 100, détermine la quantité de zone terrestre à conserver lorsque la carte est lissée. Cette option a uniquement un effet sur les fonctionnalités qui contiennent plusieurs polygones comme dans le cas d'une fonctionnalité incluant plusieurs îles. Si la zone totale de la fonctionnalité moins un polygone est supérieure au pourcentage spécifié de la zone d'origine, l'utilitaire de conversion des cartes ignore le polygone de la carte. L'utilitaire de conversion des cartes ne supprimera jamais tous les polygones de la fonctionnalité. C'est-à-dire qu'il y aura au moins un polygone pour la fonctionnalité, quelle que soit la quantité de lissage appliqué.

Après avoir choisi un angle et un pourcentage maximum à conserver, cliquez sur Appliquer. L'aperçu met à jour les modifications de lissage. Si vous avez besoin de lisser de nouveau la carte, répétez cette action jusqu'au niveau de lissage désiré. Veuillez noter qu'il existe une limite au lissage. Si vous effectuez plusieurs lissages, arrivera un moment où vous ne pourrez plus lisser la carte.

Modifier les étiquettes des fonctionnalités

Vous pouvez modifier les étiquettes des fonctionnalités (peut-être pour qu'elles correspondent aux données attendues) et également repositionner les étiquettes sur la carte. Même si vous ne pensez pas avoir besoin de modifier les étiquettes, vérifiez-les avant de créer les visualisations à partir de la carte. Parce que les étiquettes n'apparaissent pas par défaut dans l'aperçu, vous pouvez également sélectionner l'option Afficher les étiquettes sur la carte pour les afficher.

Clés. Sélectionnez la clé contenant les étiquettes des fonctionnalités à consulter et/ou modifier.

Caractéristiques. Cette liste affiche les étiquettes des fonctionnalités contenues dans la clé sélectionnée. Pour modifier l'étiquette, faites un double clic sur cette liste. Si les étiquettes apparaissent sur la carte, vous pouvez également faire un double clic sur les étiquettes des fonctionnalités directement dans l'aperçu de la carte. Si vous souhaitez comparer les étiquettes à un fichier de données réel, cliquez sur Comparer.

X/Y. Ces zones de texte répertorient le point central actuel de l'étiquette de la fonctionnalité sélectionnée sur la carte. Les unités apparaissent dans les coordonnées de la carte. Il peut s'agir de coordonnées locales cartésiennes (par exemple, le SPCS - State Plane Coordinate System-) (où X est la longitude et Y la latitude). Saisissez les coordonnées pour la nouvelle position de l'étiquette. Si les étiquettes apparaissent, vous pouvez également cliquer et faire glisser une étiquette sur la carte pour la déplacer. Les zones de texte seront mises à jour avec la nouvelle position.

Comparer. Si vous avez un fichier de données qui contient des valeurs de données censées correspondre aux étiquettes des fonctionnalités pour une clé particulière, cliquez sur Comparer pour afficher la boîte de dialogue Comparer à une source de données externe. Dans cette boîte de dialogue, vous pourrez ouvrir le fichier de données et comparer ses valeurs directement avec celles qui se trouvent dans les étiquettes de fonctionnalités de la clé de carte.

Boîte de dialogue Comparer à une source de données externe

La boîte de dialogue Comparer à une source de données externe vous permet d'ouvrir un fichier de valeurs au format tabulé (avec une extension *.txt*) ou un fichier de valeurs séparées par des virgules (avec une extension *.csv*). Lorsque le fichier est ouvert, vous pouvez sélectionner un

champ dans le fichier de données pour comparer les étiquettes des fonctionnalités dans une clé de carte spécifique. Vous pouvez ensuite corriger les incohérences du fichier carte.

Champs dans le fichier de données. Choisissez le champ dont vous souhaitez comparer les valeurs aux étiquettes de fonctionnalités. Si la première ligne du fichier *.txt* ou *.csv* contient des étiquettes de description pour chaque champ, cochez Utiliser la première ligne comme étiquette de colonne. Sinon, chaque champ sera identifié par sa position dans le fichier de données (par exemple, “Colonne 1”, “Colonne 2”, etc.).

Clé à comparer. Choisissez la clé de carte dont vous souhaitez comparer les étiquettes de fonctionnalités aux valeurs des champs des fichier de données.

Comparer. Cliquez lorsque vous êtes prêt à comparer les valeurs.

Résultats des comparaisons. Par défaut, le tableau Résultats des comparaisons ne répertorie que les valeurs de champ sans correspondance dans le fichier de données. L’application essaie de trouver une étiquette de fonctionnalité associée, en vérifiant généralement les espaces ajoutés ou manquants. Cliquez sur la liste déroulante dans la colonne *Etiquette de carte* pour trouver l’étiquette de fonctionnalité dans le fichier carte correspondant à la valeur de champ affichée. Si’il n’existe aucune étiquette de fonctionnalité correspondante dans votre fichier carte, choisissez l’option *Laisser les étiquettes sans correspondance*. Si vous souhaitez afficher toutes les valeurs de champ, même celles qui correspondent déjà à une étiquette de fonctionnalité, décochez l’option Afficher uniquement les observations sans correspondance. Cette possibilité peut vous être utile pour remplacer une ou plusieurs correspondances.

Chaque fonctionnalité ne peut être utilisée qu’une seule fois pour correspondre à une valeur de champ. Si vous souhaitez faire correspondre plusieurs fonctionnalités à une seule valeur de champ, il est possible de fusionner les fonctionnalités puis de faire correspondre la nouvelle fonctionnalité fusionnée à la valeur de champ. Pour plus d’informations sur les fonctionnalités de fusion, reportez-vous à Fusionner les fonctionnalités sur p. 299.

Fusionner les fonctionnalités

La fusion des fonctionnalités permet de créer de grandes régions sur une carte. Par exemple, si vous convertissez une carte des États, vous pouvez fusionner les États (les fonctionnalités dans cet exemple) en régions Nord, Sud, Ouest et Est de plus grande taille.

Clés. Sélectionnez la clé de carte contenant les étiquettes de fonctionnalités qui vous aident à identifier les fonctionnalités à fusionner.

Caractéristiques. Cliquez sur la première fonctionnalité à fusionner. Cliquez sur Ctrl pour sélectionner les autres fonctionnalités à fusionner. Veuillez noter que les fonctionnalités seront également sélectionnées dans l’aperçu de la carte. Vous pouvez cliquer directement sur les fonctionnalités puis sur Ctrl dans l’aperçu de la carte en plus de les sélectionner dans la liste.

Après avoir sélectionné les fonctionnalités à fusionner, cliquez sur Fusionner pour afficher la boîte de dialogue Nommer la fonctionnalité fusionné dans laquelle vous pourrez appliquer une étiquette à la nouvelle fonctionnalité. Vous pouvez cocher Colorer l’aperçu de la carte après avoir fusionné les fonctionnalités pour vous assurer que les résultats sont corrects.

Après avoir fusionné les fonctionnalités, vous pouvez également déplacer l'étiquette de la nouvelle fonctionnalité. Pour ce faire, utilisez la tâche *Modifier les étiquettes des fonctionnalités*. Pour plus d'informations, reportez-vous à la section Modifier les étiquettes des fonctionnalités sur p. 298.

Boîte de dialogue Nommer la fonctionnalité fusionné

La boîte de dialogue Nommer la fonctionnalité fusionnée vous permet d'attribuer des étiquettes à la nouvelle fonctionnalité fusionnée.

Le tableau Etiquettes affiche les informations pour chaque clé dans le fichier carte et vous permet d'attribuer une étiquette à chaque clé.

Nouvelle étiquette. Saisissez une nouvelle étiquette pour la fonctionnalité fusionnée à attribuer à la clé de carte spécifique.

Clé. La clé de carte à laquelle vous attribuez la nouvelle étiquette.

Anciennes étiquettes. Les étiquettes des fonctionnalités qui seront fusionnées dans la nouvelle fonctionnalité.

Supprimer les bordures entre les polygones attenants. Cocher cette option pour supprimer les bordures des fonctionnalités ayant été fusionnées. Par exemple, si vous fusionnez des États dans des zones géographiques, cette option supprime les bordures autour des États individuels.

Déplacer les fonctionnalités

Vous pouvez déplacer les fonctionnalités dans la carte. Cette option peut être utile lorsque vous souhaitez rassembler des fonctionnalités, comme le continent et les îles environnantes.

Clés. Sélectionnez la clé de carte contenant les étiquettes de fonctionnalités qui vous aident à identifier les fonctionnalités à déplacer.

Caractéristiques. Cliquez sur la première fonctionnalité à déplacer. Veuillez noter que la fonctionnalité sera sélectionnée dans l'aperçu de la carte. Vous pouvez également cliquer directement sur la fonctionnalité dans l'aperçu de la carte.

X/Y. Ces zones de texte répertorient le point central actuel de la fonctionnalité sur la carte. Les unités apparaissent dans les coordonnées de la carte. Il peut s'agir de coordonnées locales cartésiennes (par exemple, le SPCS - State Plane Coordinate System-) (où X est la longitude et Y la latitude). Saisissez les coordonnées pour la nouvelle position de la fonctionnalité. Vous pouvez également cliquer sur une fonctionnalité et la déplacer sur la carte. Les zones de texte seront mises à jour avec la nouvelle position.

Supprimer les fonctionnalités

Vous pouvez supprimer les fonctionnalités indésirables de la carte. Cela peut être utile lorsque vous souhaitez supprimer certaines répétitions en supprimant des fonctionnalités qui ne vous intéressent pas dans la visualisation de carte.

Clés. Sélectionnez la clé de carte contenant les étiquettes de fonctionnalités qui vous aident à identifier les fonctionnalités à supprimer.

Caractéristiques. Cliquez sur la fonctionnalité à supprimer. Si vous souhaitez supprimer plusieurs fonctionnalités en même temps, cliquez sur les fonctionnalités supplémentaires en appuyant sur Ctrl. Veuillez noter que les fonctionnalités seront également sélectionnées dans l'aperçu de la carte. Vous pouvez cliquer directement sur les fonctionnalités puis sur Ctrl dans l'aperçu de la carte en plus de les sélectionner dans la liste.

Supprimer les éléments individuels

En plus de supprimer des fonctionnalités entières, vous pouvez supprimer certains éléments individuels qui composent ces fonctionnalités, comme des lacs et de petites îles. Cette option n'est pas disponible pour les cartes avec points.

Eléments. Cliquez sur les éléments à supprimer. Si vous souhaitez supprimer plusieurs éléments en même temps, cliquez sur les éléments supplémentaires en appuyant sur Ctrl. Veuillez noter que les éléments seront également sélectionnés dans l'aperçu de la carte. Vous pouvez cliquer directement sur les éléments puis sur Ctrl dans l'aperçu de la carte en plus de les sélectionner dans la liste. Parce que la liste des noms d'éléments n'est pas descriptive (chaque élément est doté d'un chiffre dans la fonctionnalité), vérifiez la sélection dans l'aperçu de la carte pour vous assurer que vous avez bien sélectionné les éléments désirés.

Définir la projection

La projection de carte spécifie la façon dont la terre en trois dimensions est représentée en deux dimensions. Toutes les projections provoquent des distorsions. Cependant, certaines projections sont plus adaptées en fonction du type de carte utilisé : mondiale ou locale. De plus, certaines projections préservent la forme des fonctionnalités d'origine. Les projections qui préservent la forme sont des projections conformes. Cette option est disponible uniquement pour les cartes avec des coordonnées géographiques (longitude et latitude).

Contrairement aux autres options de l'utilitaire de conversion des cartes, la projection peut être modifiée après la création d'une visualisation de carte.

Projection. Sélectionnez une projection de carte. Si vous créez une carte mondiale ou des hémisphères, utilisez les projections *Locale*, de *Mercator* ou *Winkel Tripel*. Pour les zones plus petites, utilisez les projections *Locale*, *conique conforme de Lambert* ou de *Mercator transverse*. Toutes les projections utilisent l'ellipsoïde WGS83 pour les données.

- La projection **Locale** est toujours utilisée lorsque la carte est créée avec un système de coordonnées local, tel que le SPCS (State Plane Coordinate System). Ces systèmes de coordonnées sont définis par des coordonnées cartésiennes plutôt que par des coordonnées géographiques (longitude et latitude). Dans la projection locale, les lignes horizontales et verticales sont espacées de la même manière dans un système de coordonnées cartésiennes. La projection locale n'est pas conforme.
- La projection de **Mercator** est une projection conforme pour les cartes mondiales. Les lignes horizontales et verticales sont droites et toujours perpendiculaires. Veuillez noter que la projection de Mercator s'étend à l'infini en se rapprochant des pôles Nord et Sud et ne peut donc pas être utilisée si votre carte contient le pôle Nord ou le pôle Sud. La distorsion est la plus importante lorsque la carte se rapproche de ces limites.

- La projection de **Winkel Tripel** est une projection non conforme pour les cartes mondiales. Bien qu'elle ne soit pas conforme, elle offre un bon compromis entre la forme et la taille. A l'exception du méridien de l'équateur et du méridien origine, toutes les lignes sont courbes. Si votre carte mondiale contient le pôle Nord ou Sud, cette projection est un choix adapté.
- Comme son nom l'indique, la projection **conique conforme de Lambert** est une projection conforme utilisée pour les cartes des continents ou de zones terrestres plus petites qui sont plus longues à l'Est et à l'Ouest qu'au Nord et au Sud.
- La projection de **Mercator transverse** est une autre projection conforme pour les cartes continentales ou les zones terrestres plus petites. Utilisez cette projection pour les zones terrestres qui sont plus longues au Nord et au Sud qu'à l'Est et à l'Ouest.

Etape 4 : fin

A ce moment, vous pouvez ajouter un commentaire pour décrire le fichier carte et également créer un fichier de données d'échantillon à partir des clés de carte.

Clés de carte. S'il existe plusieurs clés dans le fichier carte, sélectionnez une clé de carte dont vous souhaitez afficher les étiquettes de fonctionnalités dans l'aperçu. Si vous créez un fichier de données à partir de la carte, ces étiquettes seront utilisées pour les valeurs des données.

Commentaire. Entrez un commentaire qui décrit la carte ou offre des informations supplémentaires pouvant être utiles à vos utilisateurs, comme les sources des fichiers de formes d'origine. Ce commentaire apparaît dans le système de gestion du sélectionneur de modèles de représentations graphiques.

Créer un ensemble de données à partir des étiquettes des fonctionnalités. Cochez cette option si vous souhaitez créer un fichier de données texte à partir des étiquettes de fonctionnalités affichées. Lorsque vous cliquez sur Parcourir…, vous pouvez spécifier un emplacement et un nom de fichier. Si vous ajoutez une extension *.txt*, le fichier sera enregistré sous la forme d'un fichier avec valeurs séparées par des tabulations. Si vous ajoutez une extension *.csv*, le fichier sera enregistré sous la forme d'un fichier avec valeurs séparées par des virgules. CSV est le format par défaut lorsqu'aucune extension n'est spécifiée.

Distribution des fichiers cartes

Lors de la première étape de l'utilitaire de conversion des cartes, vous avez choisi un emplacement où enregistrer le fichier SMZ converti. Vous pouvez également choisir d'ajouter la carte au système de gestion pour le sélectionneur de modèles de représentations graphiques. Si vous choisissez d'enregistrer dans le système de gestion, la carte sera disponible dans n'importe quel produit IBM SPSS que vous exécutez sur le même ordinateur.

Pour distribuer la carte à d'autres utilisateurs, vous devrez leur envoyer le fichier SMZ. Ces utilisateurs pourront ensuite utiliser le système de gestion pour importer la carte. Vous pouvez simplement envoyer le fichier dont vous avez spécifié l'emplacement à l'étape 1. Si vous souhaitez envoyer un fichier qui se trouve dans le système de gestion, vous devez d'abord l'exporter :

- ▶ Dans le sélectionneur de modèles, cliquez sur Gérer...
- ▶ Cliquez sur l'onglet Carte.

- Sélectionnez la carte à distribuer.
- Cliquez sur Exporter... et choisissez un emplacement où enregistrer le fichier.

Vous pouvez envoyer le fichier carte physique à d'autres utilisateurs. Les utilisateurs devront refaire le processus à l'envers et importer la carte dans le système de gestion.

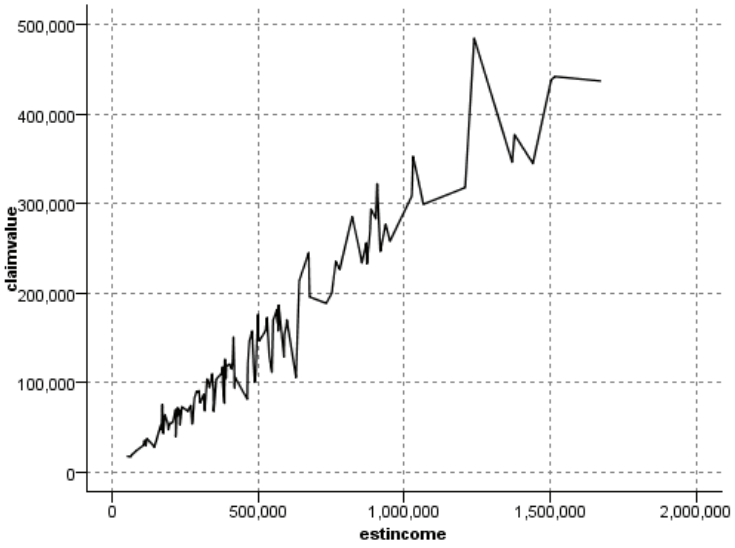
Noeud Nuage

Les noeuds Nuage montrent les relations existant entre les champs numériques. Vous pouvez créer un graphique Nuage à l'aide de points (on parle alors également de diagramme de dispersion) ou à l'aide de lignes. Vous pouvez créer trois types de nuages de lignes en définissant le mode X dans la boîte de dialogue.

Mode X = Trier

Paramétrez le mode X sur Trier pour trier par valeur les données du champ représenté sur l'axe *x*. Une ligne unique allant de gauche à droite apparaît sur le graphique. Si vous utilisez un champ nominal en tant que superposition, vous obtenez plusieurs lignes de différentes nuances, allant de gauche à droite sur le graphique.

Figure 5-35
Nuage de lignes avec le mode X paramétré sur Trier

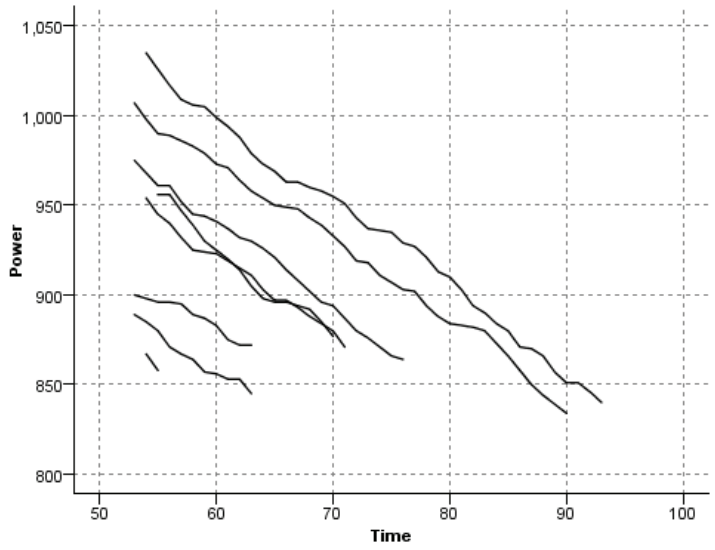


Mode X = Superposer

Paramétrez le mode X sur Superposer pour créer plusieurs nuages de lignes sur le même graphique. Dans un nuage de superposition, les données ne sont pas triées. Tant que les valeurs de l'axe *x* augmentent, les données sont représentées sur une seule ligne. Si les valeurs diminuent,

une nouvelle ligne apparaît. Par exemple, si la valeur de *x* s'accroît de 0 à 100, les valeurs *y* sont représentées par une ligne unique. Si la valeur de *x* passe en dessous de 100, une nouvelle ligne est tracée. Le nuage terminé peut contenir plusieurs nuages, ce qui est pratique pour comparer plusieurs séries de valeurs *y*. Ce type de nuage est utile pour les données intégrant une composante temporelle périodique, telle que la consommation électrique sur des périodes successives de vingt-quatre heures.

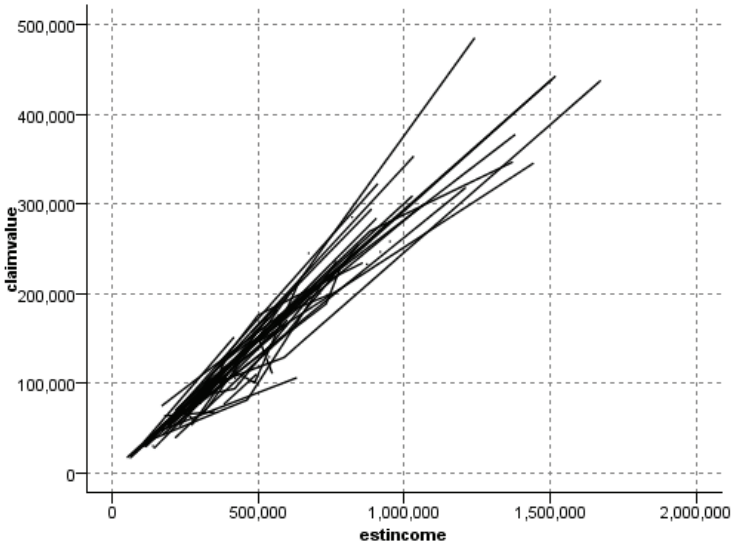
Figure 5-36
Nuage de lignes avec le mode X paramétré sur Superposer



Mode X = Selon lecture

Paramétrez le mode X sur Selon lecture pour représenter les valeurs *x* et *y* telles qu'elles sont lues dans la source de données. Cette option est utile pour les données intégrant une série temporelle et pour lesquelles vous vous intéressez aux tendances ou aux motifs dépendant de l'ordre des données. Il faut parfois trier les données avant de créer ce type de nuage. Il peut également être intéressant de comparer deux nuages similaires dont le mode X est respectivement paramétré sur Trier et sur Selon lecture afin de déterminer dans quelle mesure le tri influence le motif.

Figure 5-37
Nuage de lignes affiché précédemment avec le paramétrage Trier, exécuté à nouveau avec le mode X paramétré sur Selon lecture

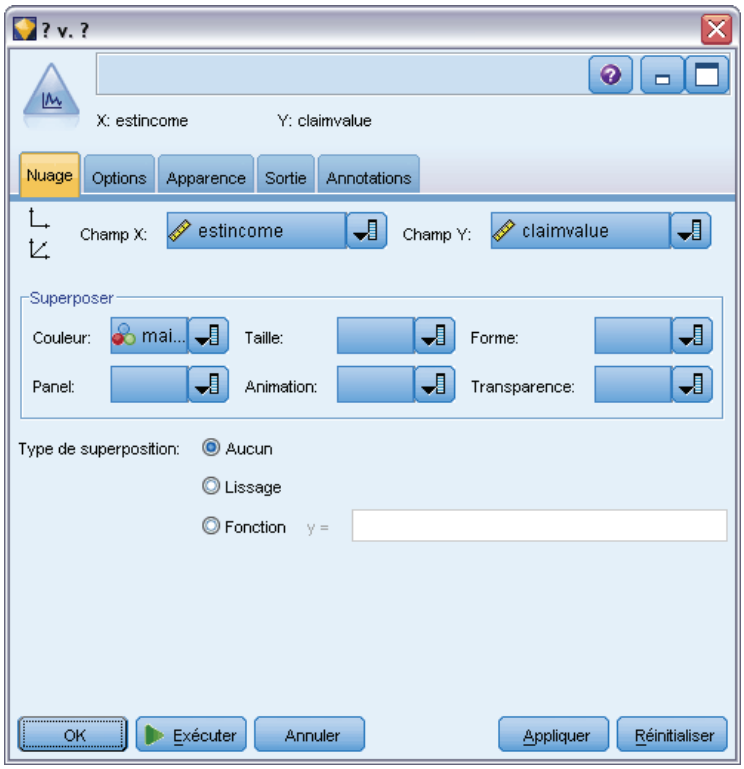


Vous pouvez aussi utiliser le noeud Représentation graphique pour produire des diagrammes de dispersion et des nuages de lignes. Néanmoins, vous pouvez choisir parmi davantage d'options dans ce noeud. Pour plus d'informations, reportez-vous à la section Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.

Onglet Noeud nuage

Les graphiques Nuage comparent les valeurs d'un champ *Y* à celles d'un champ *X*. En général, ces champs correspondent respectivement à une variable dépendante et à une variable indépendante.

Figure 5-38
Paramètres de l'onglet Nuage pour un noeud Nuage



Champ X. Dans la liste, sélectionnez le champ à afficher sur l'axe *x* horizontal.

Champ Y. Dans la liste, sélectionnez le champ à afficher sur l'axe *y* vertical.

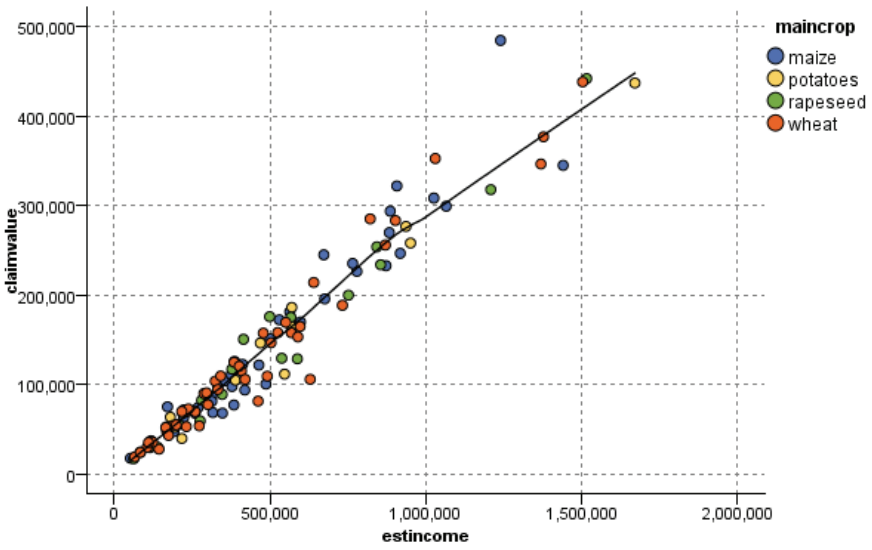
Champ Z. Lorsque vous cliquez sur le bouton 3D du diagramme, vous pouvez ensuite sélectionner un champ dans la liste à afficher sur l'axe *z*.

Superposer. Il existe plusieurs méthodes pour mettre en évidence les catégories des valeurs de données. Par exemple, vous pouvez utiliser *récolteprincipale* en tant que superposition de couleurs afin d'indiquer les valeurs *revenuest* et *valeurréclamation* de la récolte principale cultivée par les demandeurs. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Type de superposition. Indique si une fonction de superposition ou un lissage apparaît. Les fonctions de lissage et de superposition sont toujours calculées comme fonctions de y.

- **Aucune.** Aucune superposition n'est affichée.
- **Lissage.** Affiche une ligne lissée, calculée à l'aide d'une régression des moindres carrés itérative et robuste pondérée localement (LOESS). Cette méthode permet de calculer efficacement une série de régressions, chacune étant axée sur une petite zone du nuage. Une série de droites de régression « locale » est alors obtenue ; ces droites sont ensuite reliées pour créer une courbe lissée.

Figure 5-39
Nuage avec superposition par lissage LOESS



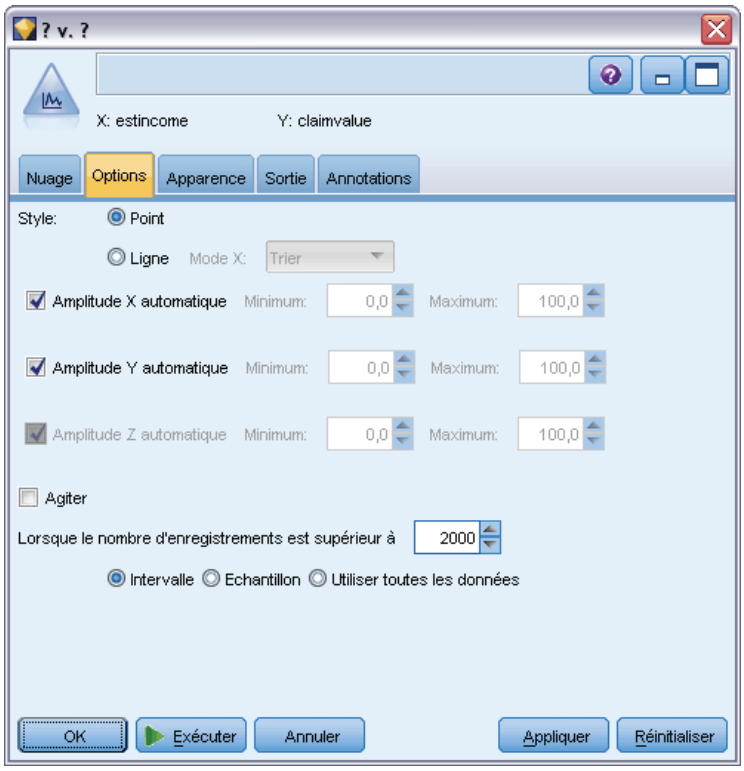
- **Fonction.** Sélectionnez cette option pour indiquer la fonction connue à comparer aux valeurs réelles. Par exemple, pour comparer des valeurs réelles à des valeurs prédites, vous pouvez représenter la fonction y = x sous la forme d'une superposition. Indiquez une fonction y = dans la zone de texte. La fonction par défaut est y = x, mais vous pouvez définir toutes sortes de fonctions, telles qu'une fonction quadratique ou une expression arbitraire en termes de x.

  Remarque : les fonctions de superposition ne sont pas disponibles pour les panneaux ou les graphiques animés.

Une fois les options du nuage définies, vous pouvez exécuter le nuage directement à partir de la boîte de dialogue en cliquant sur Exécuter. Si vous le souhaitez, vous pouvez cependant utiliser l'onglet Options pour ajouter des spécifications, telles que la création d'intervalles, le mode X et le style.

Onglet Options nuage

Figure 5-40
Paramètres de l'onglet Options du noeud Nuage



Style. Sélectionnez le style de nuage Point ou Ligne. Si vous sélectionnez Ligne, l'option Mode X est activée. La sélection de Point utilise un symbole plus (+) comme forme de point par défaut. Une fois le graphique créé, vous pouvez modifier la forme des points et leur taille.

Mode X. Pour les nuages de lignes, vous devez choisir un mode X pour définir le style du nuage. Sélectionnez Trier, Superposer ou Selon lecture. Pour les options Superposer ou Selon lecture, vous devez spécifier le nombre de modalités maximales des ensembles de données à utiliser pour échantillonner les *n* premiers enregistrements. Sinon, les 2,000 enregistrements par défaut sont utilisés.

Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs Minimum et Maximum spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Amplitude Y automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs Minimum et Maximum spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Amplitude Z automatique. Seulement quand un graphique en 3D est spécifié sur l'onglet Nuage. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs Minimum et Maximum spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Effet d'agitation.L'**agitation** est utile pour les nuages de points représentant un ensemble de données dans lequel plusieurs valeurs sont récurrentes. Pour obtenir une proportion plus claire des valeurs, vous pouvez utiliser un effet d'agitation afin de distribuer les points de façon aléatoire autour de la valeur réelle.

Remarque à l'attention des utilisateurs des anciennes versions de SPSS Modeler : la valeur de l'effet d'agitation appliquée dans un nuage utilise une mesure différente dans cette version de IBM® SPSS® Modeler. Dans les versions précédentes, la valeur était un nombre réel. Il s'agit désormais d'une proportion de la taille du cadre. Autrement dit, les valeurs d'agitation des anciens flux risquent d'être trop élevées. Dans cette version, toute valeur d'agitation autre que zéro prendra la valeur 0,2.

Nombre maximal d'enregistrements à représenter graphiquement. Spécifiez la méthode de représentation des ensembles de données volumineux. Vous pouvez spécifier le nombre de modalités maximales des ensembles de données ou utiliser les 2 000 enregistrements par défaut. Lorsque vous sélectionnez les options Intervalle ou Echantillon, les performances des ensembles de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant Utiliser toutes les données, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

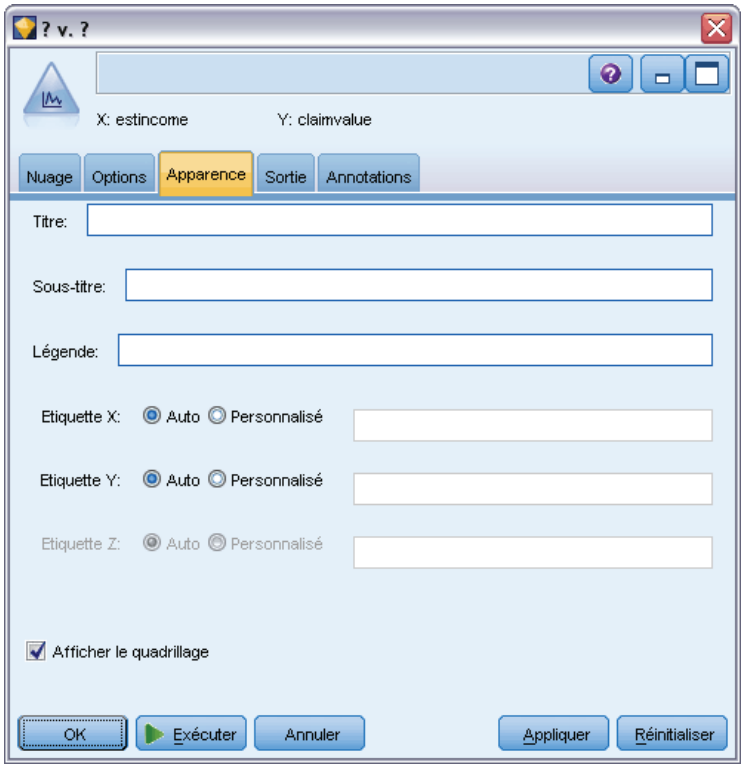
Remarque : lorsque le mode X est paramétré sur Superposer ou sur Selon lecture, ces options sont désactivées et seuls les *n* premiers enregistrements sont utilisés.

- **Intervalle.** Sélectionnez cette option pour permettre la création d'intervalles lorsque l'ensemble de données contient plus d'enregistrements que le nombre spécifié. La création d'intervalles applique une fine grille au graphique avant que soit effectué le traçage réel et compte le nombre de points apparaissant dans chacune des cellules de la grille. Dans le graphique final, un point est représenté dans chaque cellule, au niveau du centroïde de l'intervalle (moyenne de tous les emplacements de point de l'intervalle). La taille des symboles représentés indique le nombre de points dans cette zone (sauf si vous avez utilisé la taille en tant que superposition). L'utilisation du centroïde et de la taille pour représenter le nombre de points fait du nuage mis en intervalles un excellent moyen de représenter les ensembles de données volumineux. En effet, elle permet d'éviter la superposition des tracés dans les zones denses (masses de couleur impossibles à différencier) et de réduire le nombre d'artefacts de symbole (motifs de densité artificiels). Les artefacts de symbole se produisent lorsque certains symboles (notamment le symbole [+]) entrent en conflit, créant ainsi des zones denses qui n'existaient pas dans les données brutes.
- **Exemple :** Sélectionnez cette option pour échantillonner de façon aléatoire les données dans le nombre d'enregistrements saisi dans le champ de texte. La valeur par défaut est 2 000.

Onglet Apparence nuage

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-41
Paramètres de l'onglet Apparence du noeud Nuage



Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

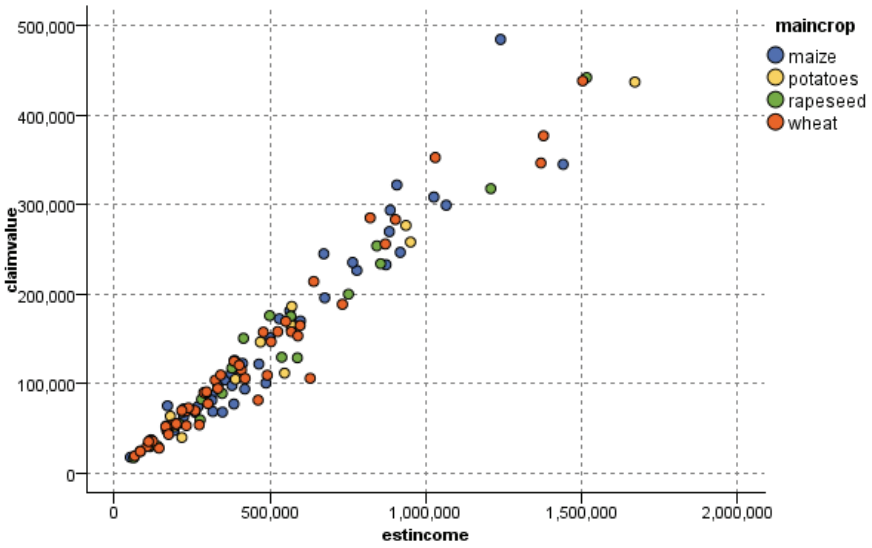
Etiquette Z. Disponible uniquement pour les graphiques en 3D. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *z*, soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un graphique Nuage

Les graphiques Nuage et Courbes sont principalement basés sur la comparaison des valeurs *X* par rapport aux valeurs *Y*. Par exemple, si vous étudiez une fraude potentielle en matière de demande de subventions agricoles (illustrée dans le fichier *fraud.str* du dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler), vous pouvez comparer, par le biais d'un réseau de neurones, le revenu réclamé dans la demande à son estimation. L'utilisation d'une superposition, telle que le type de culture, permettra de démontrer s'il existe un lien entre la demande (valeur ou nombre) et le type de culture.

Figure 5-42
Nuage représentant la relation entre le revenu estimé et la valeur de la demande, le type de culture principale servant de superposition



Les graphiques Nuage, Courbes et Evaluation étant des illustrations en deux dimensions de la comparaison entre *Y* et *X*, il est facile d'interagir avec ceux-ci en définissant des zones, en marquant un élément, ou en traçant des bandes. Vous pouvez également générer des noeuds pour les données représentées par ces zones, bandes, ou éléments. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360.

Noeud Proportion

Les graphiques ou les tableaux Proportion montrent l'occurrence, dans un ensemble de données, de valeurs symboliques (non numériques) comme un type de prêt hypothécaire ou le sexe d'un individu. Les noeuds Proportion servent souvent à montrer les déséquilibres des données, déséquilibres pouvant être rectifiés grâce à l'utilisation d'un noeud Equilibrer avant la création d'un modèle. Vous pouvez générer automatiquement un noeud Equilibrer à l'aide du menu Générer de la fenêtre d'un graphique ou d'un tableau Proportion.

Vous pouvez aussi utiliser le noeud Représentation graphique pour produire des graphiques à barres. Néanmoins, vous pouvez choisir parmi davantage d'options dans ce noeud. Pour plus d'informations, reportez-vous à la section Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.

Remarque : pour montrer l'occurrence de valeurs numériques, utilisez de préférence un noeud Histogramme.

Onglet Nuage de proportion

Figure 5-43
Paramètres de l'onglet Nuage pour un noeud Proportion

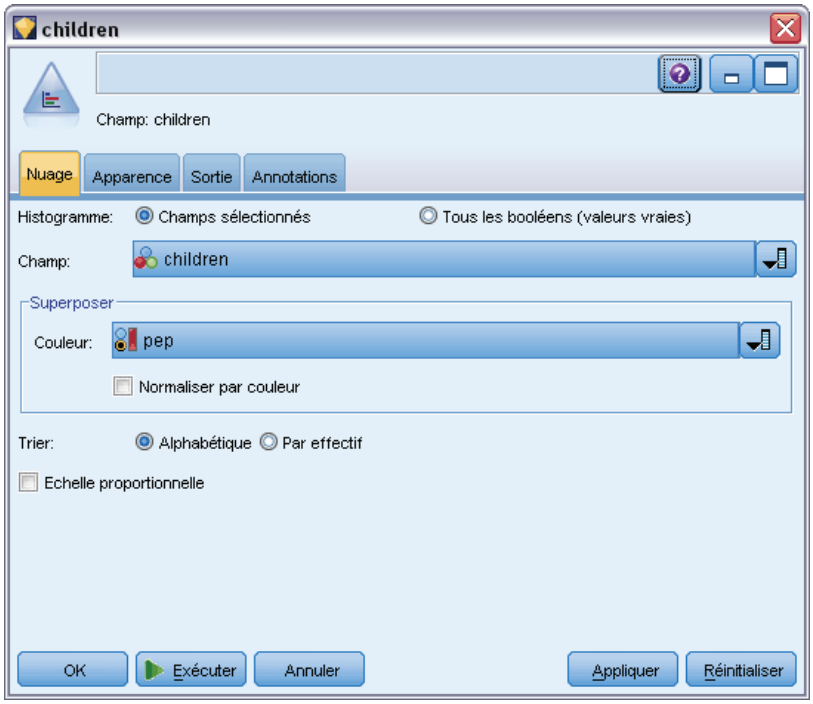


Diagramme : Sélectionnez le type de proportion. Sélectionnez Champs sélectionnés pour afficher la proportion du champ sélectionné. Sélectionnez Tous les booléens (valeurs vraies) pour afficher la proportion des valeurs true (vrai) des champs booléens de l'ensemble de données.

Champ. Sélectionnez le champ nominal ou le champ booléen dont vous souhaitez montrer la proportion des valeurs. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés dans la liste.

Superposer. Sélectionnez le champ nominal ou le champ booléen à utiliser en tant que superposition de couleurs, illustrant la proportion de ses valeurs au sein de chaque valeur du champ spécifié. Par exemple, vous pouvez utiliser la réponse à une campagne de marketing (*pep*) comme superposition au nombre d'enfants (*enfant*) afin d'illustrer la réactivité en fonction de la taille de la famille. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Normaliser par couleur. Sélectionnez cette option pour mettre les barres à l'échelle de sorte que toutes les barres occupent la totalité du graphique. Les valeurs de superposition correspondent à une proportion de chaque barre, ce qui facilite les comparaisons entre les catégories.

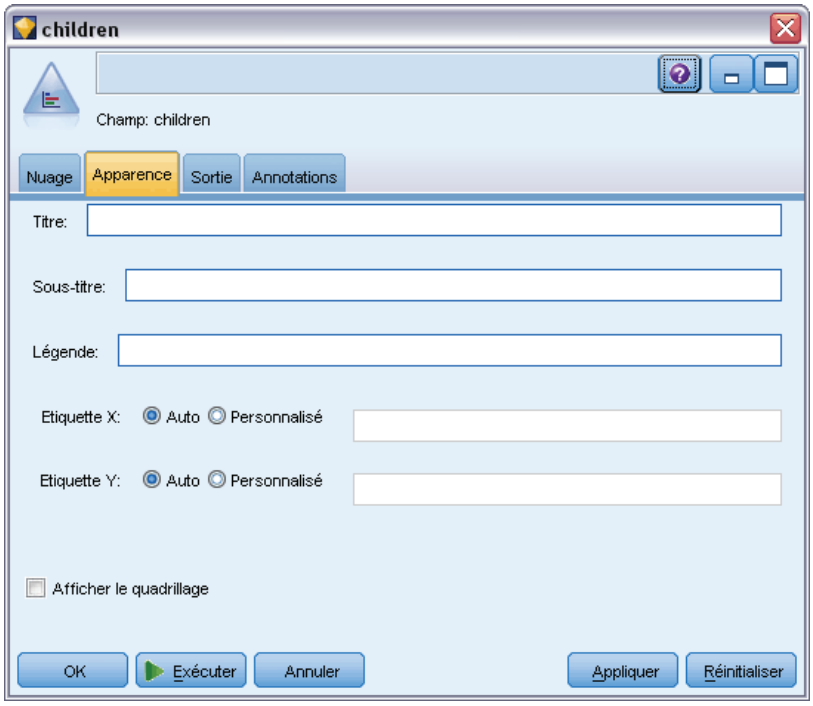
Trier. Sélectionne la méthode utilisée pour afficher les valeurs sur le graphique Proportion. Sélectionnez Alphabétique pour utiliser le classement par ordre alphabétique ou Par effectif pour répertorier les valeurs par ordre décroissant d'occurrence.

Echelle proportionnelle. Sélectionnez cette option pour mettre la proportion des valeurs à l'échelle de sorte que la valeur la plus représentée occupe la totalité du nuage. Toutes les autres barres sont mises à l'échelle en fonction de cette valeur. Désélectionnez cette option pour mettre les barres à l'échelle en fonction du total de chaque valeur.

Onglet Apparence de proportion

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-44
Paramètres de l'onglet apparence



Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un noeud Proportion

Les noeuds Proportion sont utilisés pour montrer la proportion des valeurs symboliques dans un ensemble de données. Ils sont fréquemment utilisés avant les noeuds de manipulation pour explorer les données et corriger les déséquilibres. Par exemple, si les instances des personnes sans enfant interrogées sont beaucoup plus nombreuses que celles des autres types de personne interrogée, vous souhaiterez peut-être réduire le nombre de ces instances afin de pouvoir générer une règle plus utile pour vos opérations de Data mining ultérieures. Les noeuds Proportion vous aident à examiner ces déséquilibres et à prendre des décisions les concernant.

Le noeud Proportion est particulier car il produit à la fois un graphique et un tableau pour analyser vos données.

Figure 5-45
Graphique Proportion affichant le nombre de personnes avec ou sans enfants qui ont répondu à une campagne de marketing

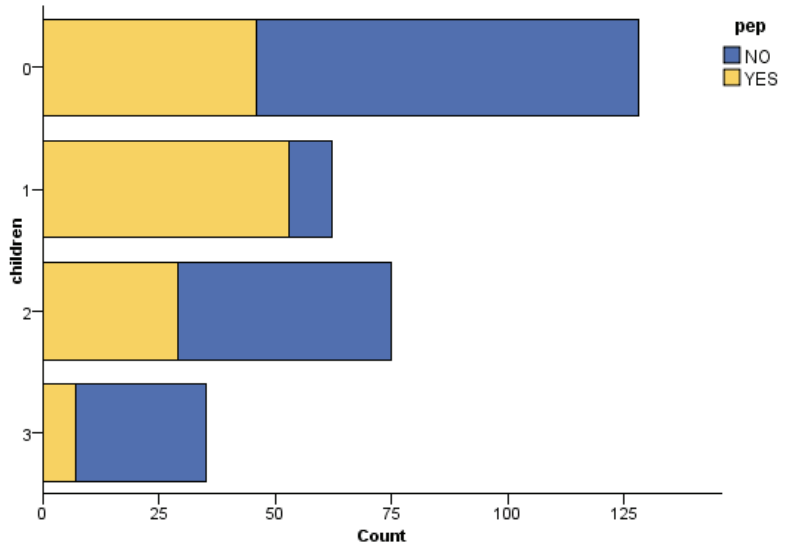
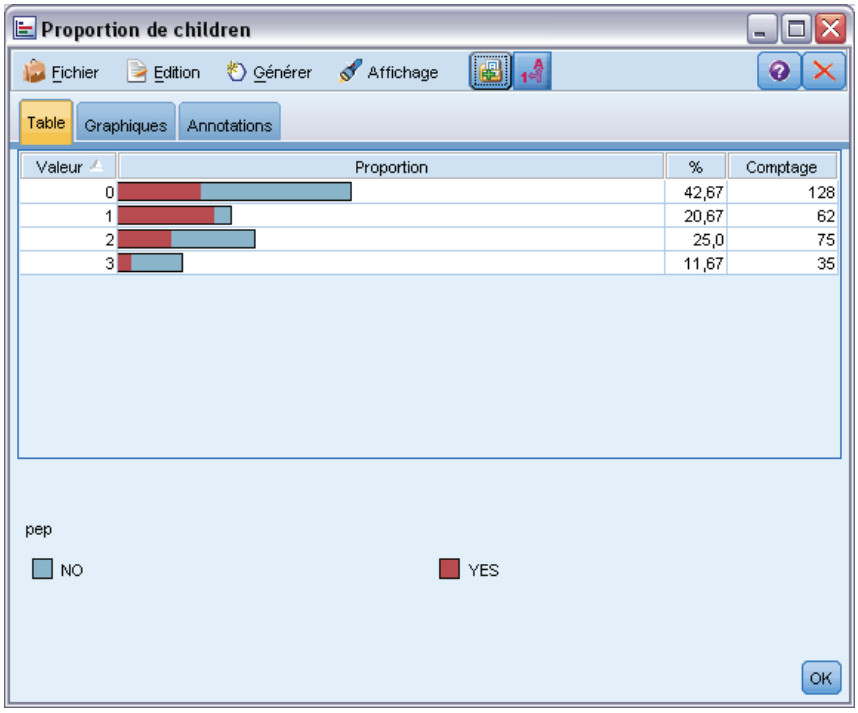


Figure 5-46
Tableau Proportion affichant la proportion de personnes avec ou sans enfants qui ont répondu à une campagne de marketing



Une fois que vous avez créé un graphique et un tableau Proportion et examiné les résultats, vous pouvez utiliser les options des menus pour regrouper et copier les valeurs, ainsi que pour générer un certain nombre de noeuds pour la préparation des données. En outre, vous pouvez copier ou exporter les informations du graphique et du tableau pour les utiliser dans d'autres applications, par exemple MS Word ou MS PowerPoint. Pour plus d'informations, reportez-vous à la section Impression, enregistrement, copie et exportation de graphiques sur p. 395.

Pour sélectionner et copier des valeurs à partir d'un tableau Proportion

- Cliquez sur les lignes tout en maintenant le bouton de la souris enfoncé afin de sélectionner un ensemble de valeurs. Vous pouvez aussi utiliser l'option Sélectionner tout du menu Edition pour sélectionner toutes les valeurs.
- Dans le menu Edition, sélectionnez Copier la table ou Copier la table (avec les noms de champ).
- Collez les valeurs dans le Presse-papiers ou dans l'application souhaitée.

  Remarque : Les barres ne sont pas copiées directement. A la place, ce sont les valeurs du tableau qui sont copiées. Autrement dit, les valeurs superposées ne figurent pas dans le tableau copié.

Pour regrouper des valeurs à partir d'un tableau Proportion

- Sélectionnez les valeurs à regrouper à l'aide de la méthode Ctrl+clic.
- Dans le menu Edition, sélectionnez Associer.

Remarque : Lorsque vous associez et dissociez des valeurs, le graphique de l'onglet Graphique est automatiquement redessiné pour refléter les modifications.

Vous pouvez également :

- Dissocier les valeurs d'un groupe en sélectionnant le nom de ce groupe dans la liste des proportions et en choisissant Dissocier dans le menu Edition.
- Editer un groupe en sélectionnant son nom dans la liste des proportions et en choisissant Editer le groupe dans le menu Edition. Dans la boîte de dialogue qui apparaît, vous pouvez ajouter des valeurs au groupe ou les en supprimer.

Figure 5-47
Boîte de dialogue Editer le groupe

Options du menu Générer

Vous pouvez utiliser les options du menu Générer pour sélectionner un sous-ensemble de données, calculer un champ booléen, regrouper des valeurs, reclassifier des valeurs ou équilibrer les données d'un graphique ou d'un tableau. Ces opérations génèrent un noeud Préparation des données et le placent dans l'espace de travail de flux. Pour utiliser le noeud généré, connectez-le à un flux existant. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

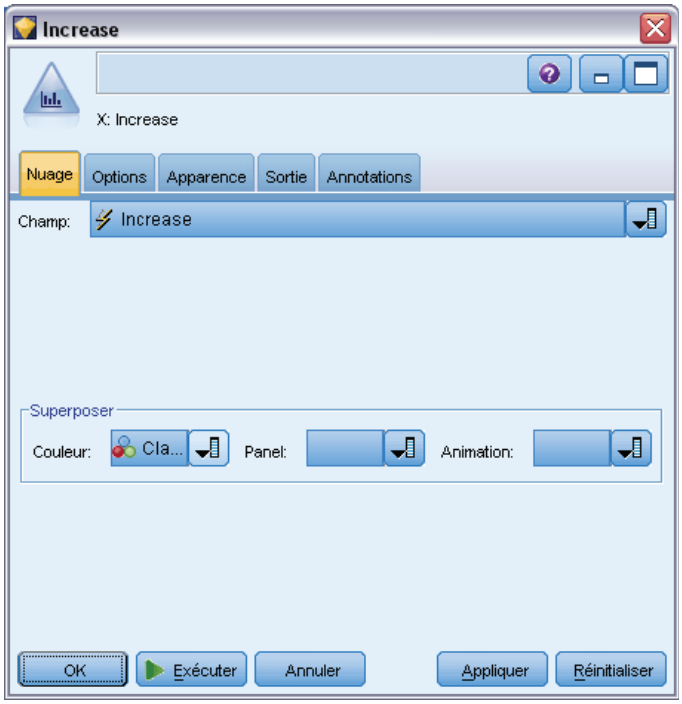
Noeud Histogramme

Les noeuds Histogramme montrent l'occurrence des valeurs des champs numériques. Ils sont souvent utilisés pour explorer les données avant toute création de modèles ou manipulation. Semblables au noeud Proportion, les noeuds Histogramme servent souvent à montrer les déséquilibres des données. Vous pouvez utiliser le noeud Représentation graphique pour produire un histogramme, mais vous pouvez également sélectionner davantage d'options dans ce noeud. Pour plus d'informations, reportez-vous à la section Types de visualisation des Représentations graphiques intégrées disponibles sur p. 263.

Remarque : Pour montrer l'occurrence des valeurs des champs symboliques, utilisez un noeud Proportion.

Onglet Nuage d'histogramme

Figure 5-48
Paramètres de l'onglet Nuage du noeud Histogramme

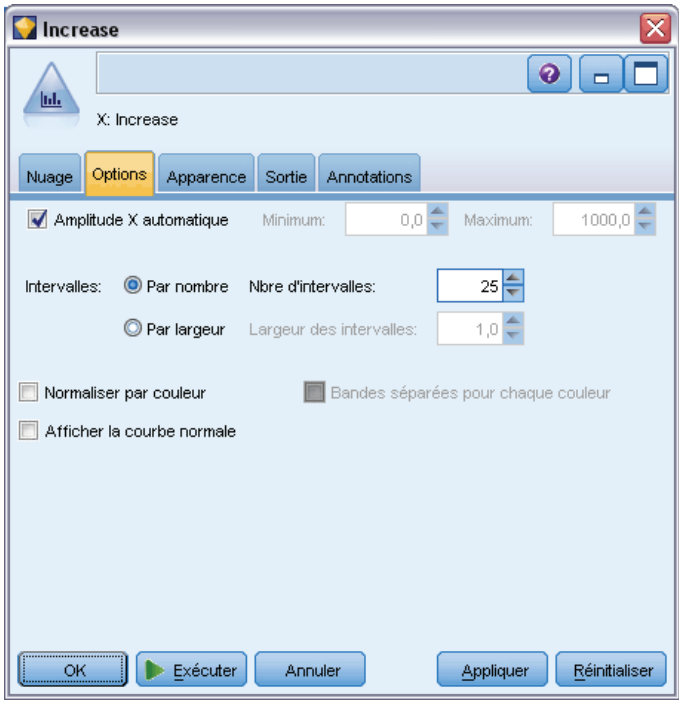


Champ. Sélectionnez le champ numérique dont vous souhaitez montrer la proportion des valeurs. Seuls les champs n'ayant pas été explicitement définis comme symboliques (catégoriels) sont répertoriés.

Superposer. Sélectionnez un champ symbolique afin de montrer les catégories de valeurs du champ spécifié. La sélection d'un champ de superposition transforme le graphique Histogramme en un graphique aux valeurs superposées. Les couleurs servent à représenter les différentes catégories du champ de superposition. Le noeud Histogramme met à disposition trois types de superpositions : couleur, panneau et animation. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Onglet Options d'histogramme

Figure 5-49
Paramètres de l'onglet Options du noeud Histogramme



Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs Minimum et Maximum spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Intervalles. Sélectionnez soit Par nombre soit Par largeur.

- Sélectionnez Par nombre pour afficher un nombre fixe de barres dont la largeur dépend de l'amplitude et du nombre d'intervalles spécifiés. Indiquez le nombre d'intervalles à utiliser dans le graphique dans l'option Nbre d'intervalles. Utilisez les flèches pour rectifier le nombre.
- Sélectionnez Par largeur pour créer un graphique formé de barres de largeur fixe. Le nombre d'intervalles dépend de la largeur indiquée et de l'amplitude de valeurs. Indiquez la largeur des barres dans l'option Largeur d'intervalle.

Normaliser par couleur. Sélectionnez cette option pour attribuer la même hauteur à toutes les barres, les valeurs superposées étant affichées sous la forme d'un pourcentage de la totalité des observations dans chaque barre.

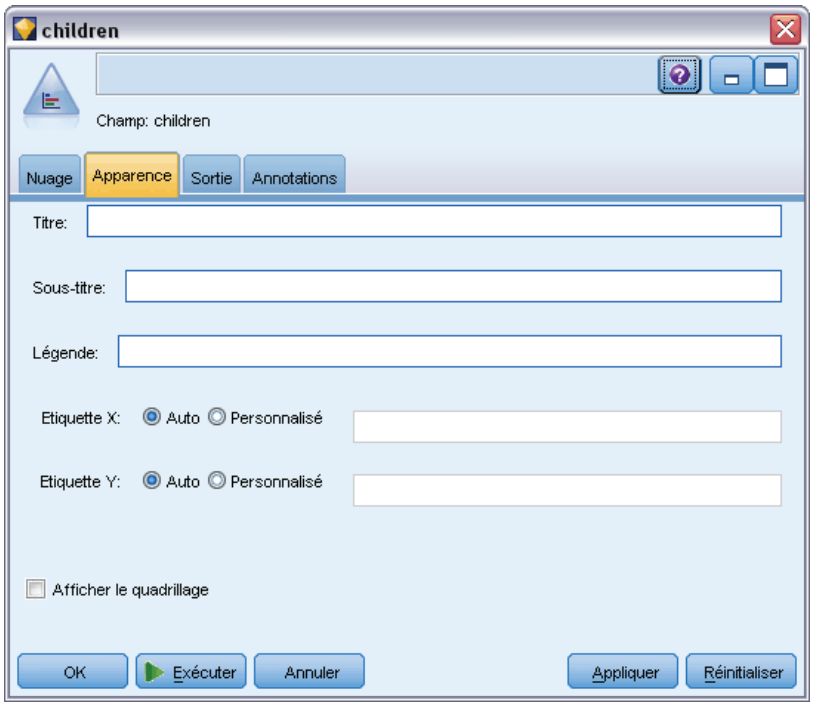
Afficher la courbe normale. Sélectionnez cette option pour ajouter une courbe normale au graphique affichant la moyenne et la variance des données.

Bandes séparées pour chaque couleur. Sélectionnez cette option pour afficher chaque valeur superposée sous la forme d'une bande distincte sur le graphique.

Onglet Apparence d'histogramme

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-50
Paramètres de l'onglet Apparence pour la plupart des noeuds Graphiques



Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

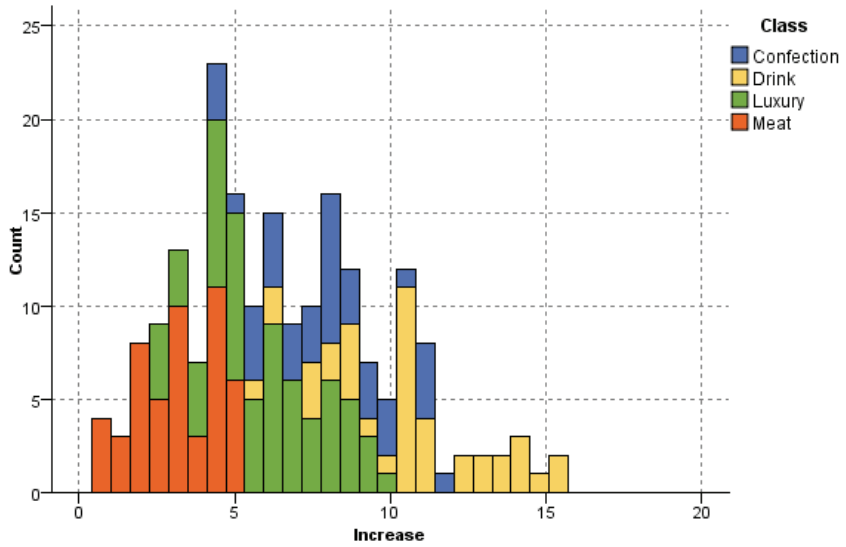
Utilisation des histogrammes

Les graphiques Histogramme montrent la proportion des valeurs d'un champ numérique dont les valeurs sont situées sur l'axe *x*. Les histogrammes fonctionnent de la même manière que les graphiques Résumés. Les graphiques Résumé montrent la proportion des valeurs d'un champ numérique *par rapport à celles d'un autre*, plutôt que l'occurrence de valeurs d'un champ unique.

Lorsque vous avez créé un graphique, vous pouvez observer les résultats et déterminer des bandes pour diviser les valeurs le long de l'axe *x* ou définir des régions. Vous pouvez également marquer des éléments dans le graphique. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360.

Vous pouvez utiliser les options du menu Générer pour créer des noeuds Equilibrer, Sélectionner ou Calculer à l'aide des données du graphique ou plus spécifiquement dans les bandes, régions ou les éléments marqués. Ce type de graphique est souvent utilisé avant les noeuds de manipulation pour explorer les données et corriger les éventuels déséquilibres en générant un noeud Equilibrer à partir du graphique à utiliser dans le flux. Vous pouvez également générer un noeud booléen Calculer pour ajouter un champ montrant à quelle bande appartient chaque enregistrement ou un noeud Sélectionner pour sélectionner tous les enregistrements appartenant à un ensemble ou à une amplitude spécifique de valeurs. Ces opérations vous aident à vous concentrer sur un sous-ensemble particulier de données pour procéder à une exploration plus approfondie. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Figure 5-51
Graphique Histogramme montrant la proportion par catégorie de la hausse d'achat due à une promotion

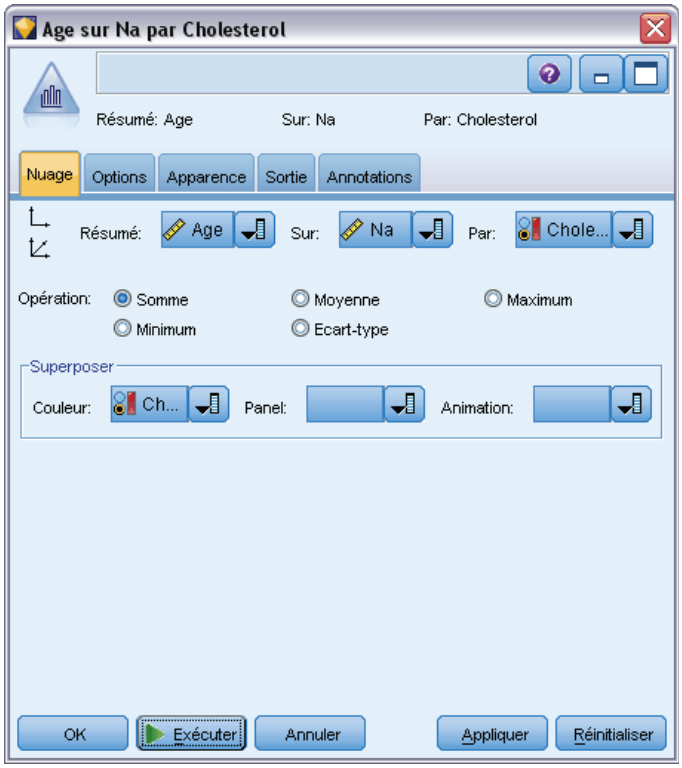


Noeud Résumé

Les graphiques Résumé sont semblables aux graphiques Histogramme, excepté que les graphiques Résumé montrent la proportion des valeurs d'un champ numérique par rapport à celles d'un autre, plutôt que l'occurrence de valeurs d'un champ unique. Les graphiques Résumé sont utiles pour illustrer une variable ou un champ dont les valeurs changent avec le temps. Grâce à la représentation graphique en 3D, vous pouvez en outre inclure un axe symbolique affichant les proportions par catégorie. Des Résumés bidimensionnels sont affichés sous forme de diagrammes en bâtons empilés, avec des superpositions le cas échéant. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Onglet nuage de Résumé

Figure 5-52
Paramètres de l'onglet Nuage du noeud Résumé



Résumé. Sélectionnez le champ dont les valeurs doivent être résumées et affichées en fonction de l'amplitude de valeurs du champ défini dans Sur. Seuls les champs n'ayant pas été indiqués comme symboliques sont répertoriés.

Sur. Sélectionnez le champ dont les valeurs doivent être utilisées pour afficher le champ défini dans Résumé.

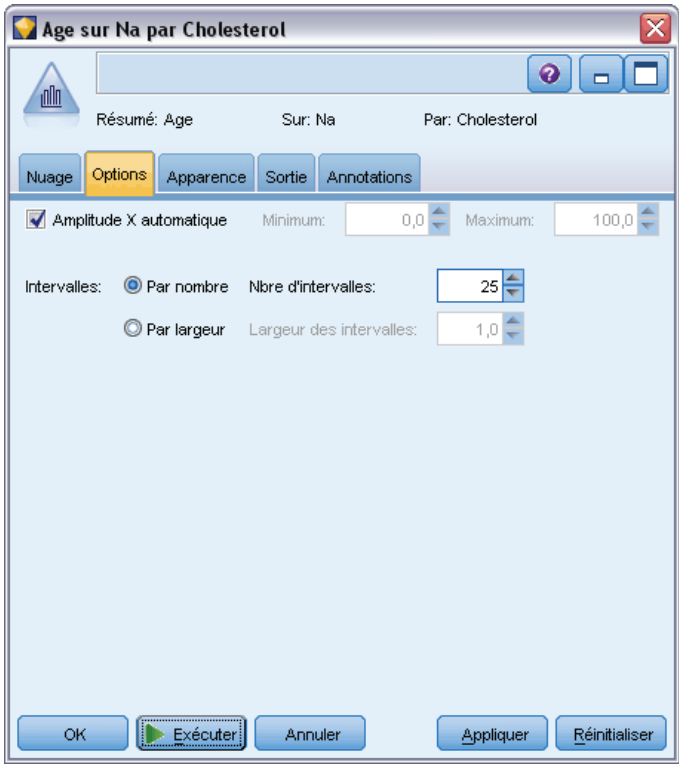
Par. Activée lors de la création de graphiques en 3D, cette option vous permet de sélectionner le champ nominal ou le champ booléen utilisé pour l'affichage du champ de résumé par catégorie.

Opération. Permet de sélectionner ce que représente chaque barre du graphique Résumé. Les options disponibles sont Somme, Moyenne, Maximum, Minimum et Ecart- type.

Superposer. Sélectionnez un champ symbolique afin de montrer les catégories de valeurs du champ sélectionné. La sélection d'un champ de superposition transforme le graphique Résumé et crée plusieurs barres de différentes couleurs pour chaque catégorie. Ce noeud dispose de trois types de superposition : couleur, panneau et animation. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Onglet Options de résumé

Figure 5-53
Paramètres de l'onglet Options du noeud Résumé



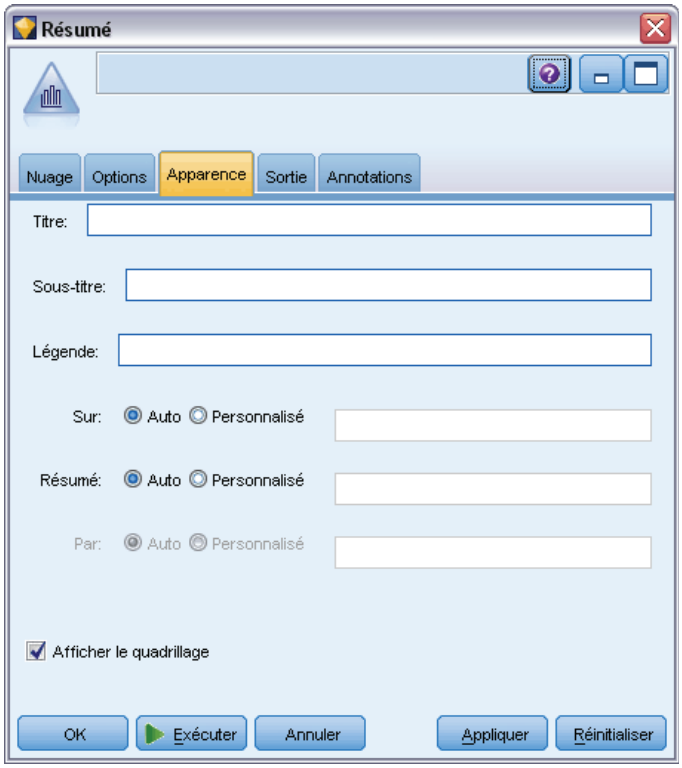
Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs Minimum et Maximum spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Intervalles. Sélectionnez soit Par nombre soit Par largeur.

- Sélectionnez Par nombre pour afficher un nombre fixe de barres dont la largeur dépend de l'amplitude et du nombre d'intervalles spécifiés. Indiquez le nombre d'intervalles à utiliser dans le graphique dans l'option Nbre d'intervalles. Utilisez les flèches pour rectifier le nombre.
- Sélectionnez Par largeur pour créer un graphique formé de barres de largeur fixe. Le nombre d'intervalles dépend de la largeur indiquée et de l'amplitude de valeurs. Indiquez la largeur des barres dans l'option Largeur d'intervalle.

Onglet Apparence de résumé

Figure 5-54
Paramètres de l'onglet Apparence pour un noeud Résumé



Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Sur l'étiquette Vous pouvez soit accepter l'étiquette générée automatiquement, soit sélectionner Personnalisé pour indiquer une étiquette.

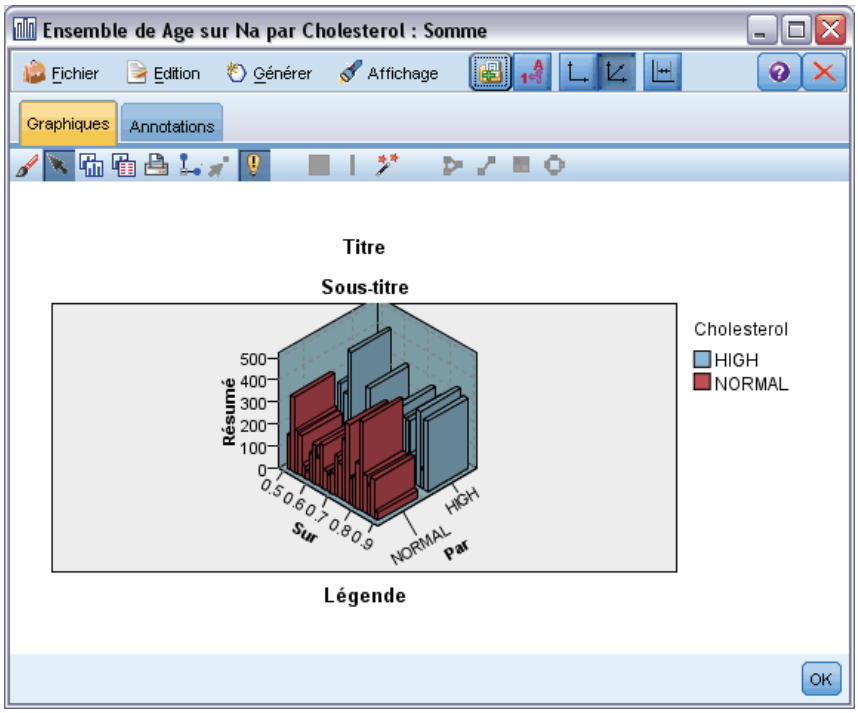
Collecte des étiquettes. Vous pouvez soit accepter l'étiquette générée automatiquement, soit sélectionner Personnalisé pour indiquer une étiquette.

Par étiquette. Vous pouvez soit accepter l'étiquette générée automatiquement, soit sélectionner Personnalisé pour indiquer une étiquette.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Les exemples suivants montrent où sont placées les options d'apparence dans une version 3D du graphique.

Figure 5-55
Position des options d'apparence du graphique sur un graphique résumé en 3D



Utilisation d'un graphique Résumé

Les graphiques Résumé montrent la proportion des valeurs d'un champ numérique *par rapport à celles d'un autre*, plutôt que l'occurrence de valeurs d'un champ unique. Les histogrammes fonctionnent de la même manière que les graphiques Résumés. Les graphiques Histogramme montrent la proportion des valeurs d'un champ numérique dont les valeurs sont situées sur l'axe *x*.

Lorsque vous avez créé un graphique, vous pouvez observer les résultats et déterminer des bandes pour diviser les valeurs le long de l'axe *x* ou définir des régions. Vous pouvez également marquer des éléments dans le graphique. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360.

Vous pouvez utiliser les options du menu Générer pour créer des noeuds Equilibrer, Sélectionner ou Calculer à l'aide des données du graphique ou plus spécifiquement dans les bandes, régions ou les éléments marqués. Ce type de graphique est souvent utilisé avant les noeuds de manipulation pour explorer les données et corriger les éventuels déséquilibres en générant un noeud Equilibrer à partir du graphique à utiliser dans le flux. Vous pouvez également générer un noeud booléen Calculer pour ajouter un champ montrant à quelle bande appartient chaque enregistrement ou un noeud Sélectionner pour sélectionner tous les enregistrements appartenant à un ensemble ou à une amplitude spécifique de valeurs. Ces opérations vous aident à vous concentrer sur un sous-ensemble particulier de données pour procéder à une exploration plus approfondie. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Figure 5-56
Graphique Résumé en 3D montrant la somme Na_sur_K par rapport à l'âge pour les niveaux de cholestérol élevés et normaux

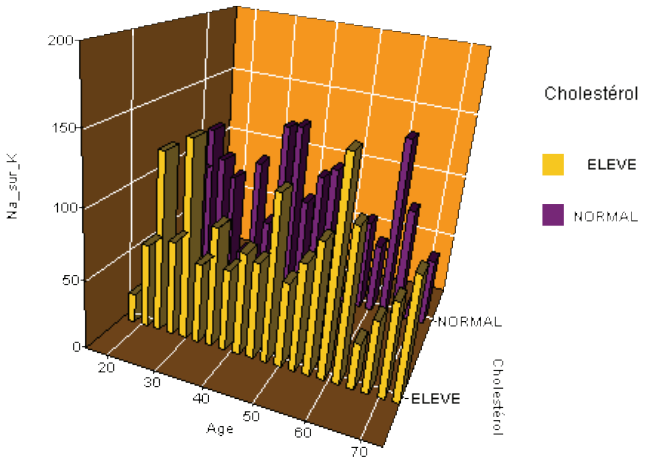
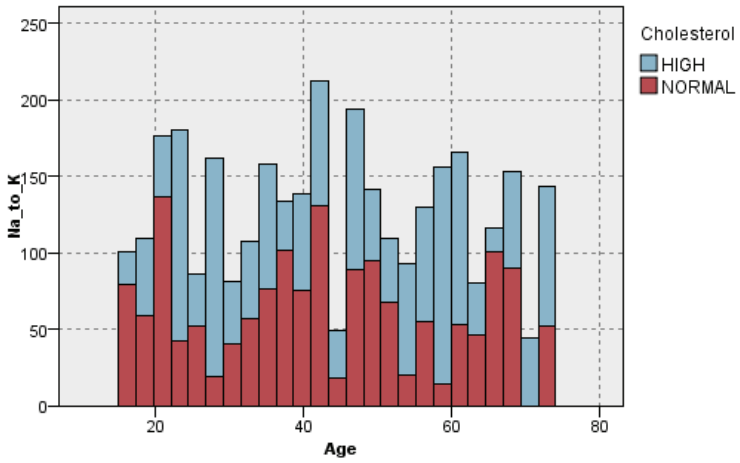


Figure 5-57
Graphique Résumé sans axe z, mais avec le taux de cholestérol affiché sous la forme d'une superposition de couleurs

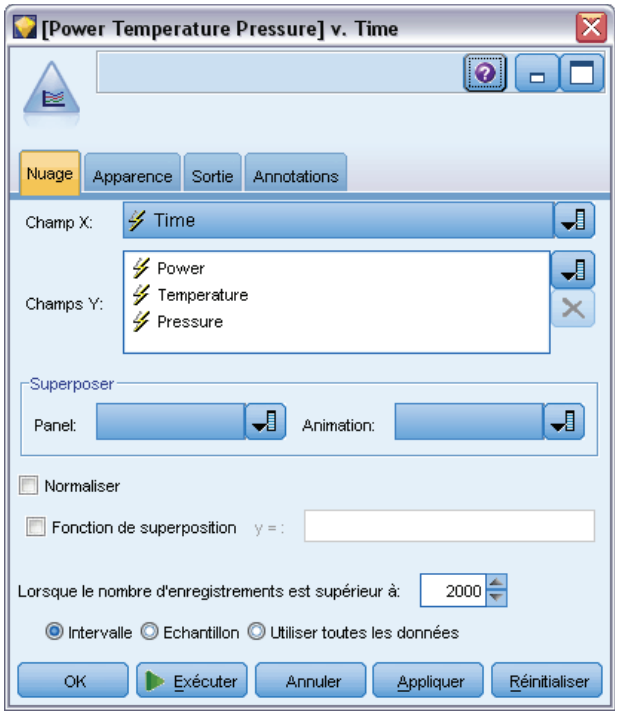


Noeud Courbes

Un graphique Courbes est un type de graphique particulier qui affiche plusieurs champs *Y* pour un seul champ *X*. Les champs *Y* sont représentés par des lignes colorées. Chacun équivaut à un nœud Nuage dont le style est défini sur Ligne et le mode X sur Trier. Les graphiques Courbes sont utiles lorsque vous avez des données de séquence temporelle et que vous souhaitez explorer les variations de plusieurs variables dans le temps.

Onglet Nuage de courbes

Figure 5-58
Paramètres de l'onglet Nuage pour un noeud Courbes



Champ X. Dans la liste, sélectionnez le champ à afficher sur l'axe *x* horizontal.

Champs Y. Sélectionnez dans la liste les champs à afficher sur l'intervalle des valeurs de champ *X*. Utilisez le sélecteur de champs pour sélectionner plusieurs champs. Cliquez sur le bouton de suppression pour supprimer des champs de la liste.

Superposer. Il existe plusieurs méthodes pour mettre en évidence les catégories des valeurs de données. Par exemple, vous pouvez utiliser une superposition animée afin d'afficher plusieurs nuages pour chaque valeur des données. C'est utile pour les ensembles contenant plus de 10 catégories. Lors d'une utilisation avec des ensembles contenant plus de 15 catégories, vous remarquerez peut-être une baisse des performances. Pour plus d'informations, reportez-vous à la section Apparences, superpositions, panneaux et animation sur p. 247.

Normaliser. Sélectionnez cette option pour mettre toutes les valeurs *Y* à l'échelle sur l'intervalle 0–1 afin de les afficher sur le graphique. La fonction de normalisation vous permet d'explorer les relations existant entre les lignes, relations qui risqueraient sinon d'être occultées en raison des différences au niveau de l'intervalle de valeurs de chaque série ; il est recommandé de l'utiliser lorsque vous représentez plusieurs lignes sur le même graphique ou lorsque vous comparez des graphiques dans des panneaux mitoyens. (Il est inutile d'appliquer une normalisation lorsque toutes les valeurs de données sont comprises dans un même intervalle.)

Figure 5-59
Graphique Courbes standard indiquant les variations de la centrale électrique dans le temps (sans normalisation, il est impossible de visualiser le nuage concernant la pression)

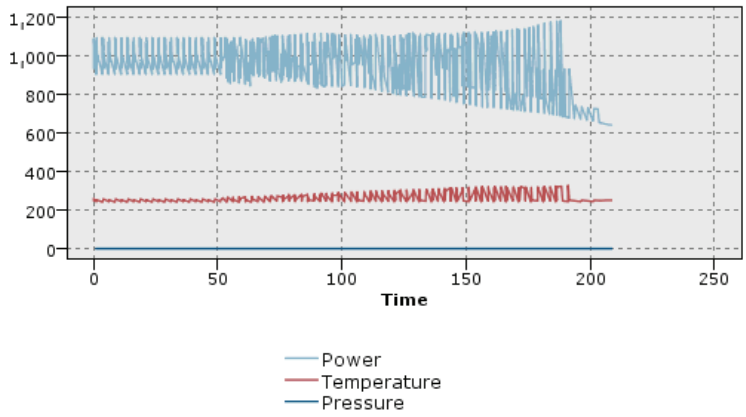
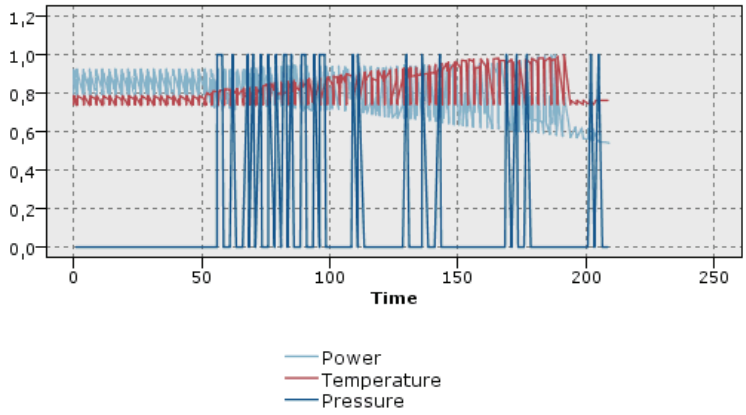


Figure 5-60
Graphique Courbes normalisé montrant le nuage de la pression



Fonction de superposition. Sélectionnez cette option pour indiquer la fonction connue à comparer aux valeurs réelles. Par exemple, pour comparer des valeurs réelles à des valeurs prédites, vous pouvez représenter la fonction y = x sous la forme d'une superposition. Indiquez une fonction y = dans la zone de texte. La fonction par défaut est y = x, mais vous pouvez définir toutes sortes de fonctions, telles qu'une fonction quadratique ou une expression arbitraire en termes de x.

Remarque : les fonctions de superposition ne sont pas disponibles pour les panneaux ou les graphiques animés.

Lorsque le nombre d'enregistrements est supérieur à. Spécifiez la méthode de représentation des ensembles de données volumineux. Vous pouvez spécifier la taille maximale des ensembles de données ou utiliser les 2 000 points par défaut. Lorsque vous sélectionnez les options Intervalle ou Echantillon, les performances des ensembles de données volumineux sont optimisées. Vous pouvez

également choisir de représenter tous les points de données en sélectionnant Utiliser toutes les données, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

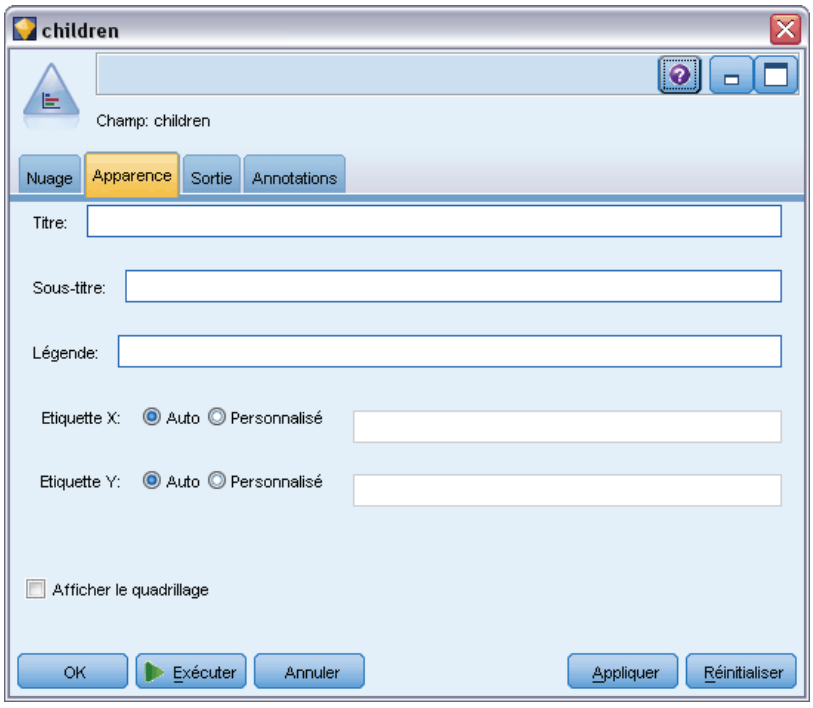
Remarque : lorsque le mode X est paramétré sur Superposer ou sur Selon lecture, ces options sont désactivées et seuls les *n* premiers enregistrements sont utilisés.

- **Intervalle.** Sélectionnez cette option pour permettre la création d'intervalles lorsque l'ensemble de données contient plus d'enregistrements que le nombre spécifié. La création d'intervalles applique une fine grille au graphique avant que soit effectué le traçage réel et compte le nombre de connexions apparaissant dans chacune des cellules de la grille. Dans le graphique final, une connexion est représentée dans chaque cellule, au niveau du centroïde de l'intervalle (moyenne de tous les emplacements de connexion de l'intervalle).
- **Exemple :** Sélectionnez cette option pour échantillonner de façon aléatoire les données dans le nombre d'enregistrements spécifié.

Onglet Apparence de courbes

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-61
Paramètres de l'onglet Apparence pour la plupart des noeuds Graphiques



Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

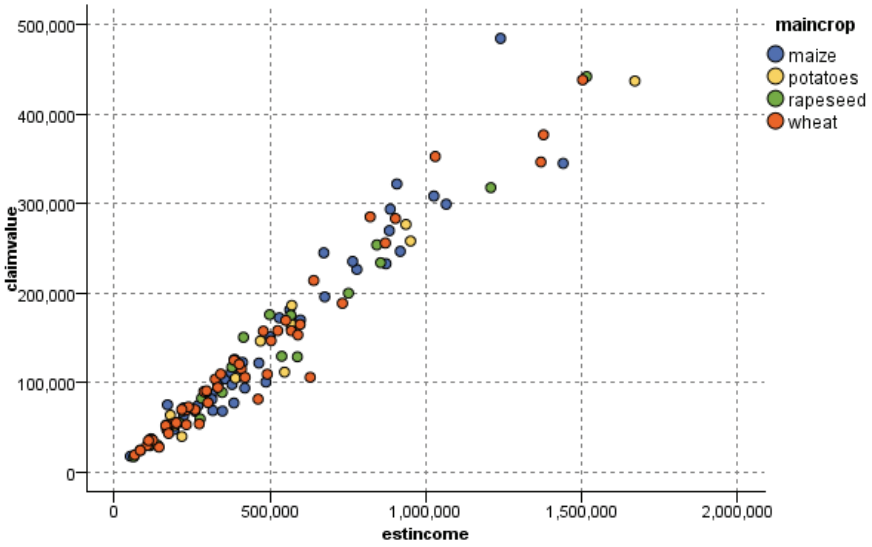
Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un graphique Courbes

Les graphiques Nuage et Courbes sont principalement basés sur la comparaison des valeurs *X* par rapport aux valeurs *Y*. Par exemple, si vous étudiez une fraude potentielle en matière de demande de subventions agricoles (illustrée dans le fichier *fraud.str* du dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler), vous pouvez comparer, par le biais d'un réseau de neurones, le revenu réclamé dans la demande à son estimation. L'utilisation d'une superposition, telle que le type de culture, permettra de démontrer s'il existe un lien entre la demande (valeur ou nombre) et le type de culture.

Figure 5-62
Nuage représentant la relation entre le revenu estimé et la valeur de la demande, le type de culture principale servant de superposition

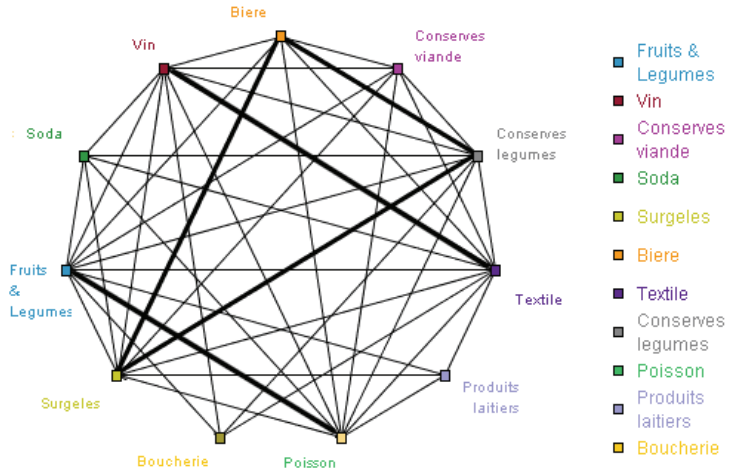


Les graphiques Nuage, Courbes et Evaluation étant des illustrations en deux dimensions de la comparaison entre *Y* et *X*, il est facile d'interagir avec ceux-ci en définissant des zones, en marquant un élément, ou en traçant des bandes. Vous pouvez également générer des noeuds pour les données représentées par ces zones, bandes, ou éléments. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360.

Noeud Relations

Les noeuds Relations montrent la force des relations existant entre les valeurs de plusieurs champs symboliques. Le graphique affiche les connexions à l'aide de divers types de ligne indiquant la force de la connexion. Par exemple, vous pouvez utiliser un noeud Relations pour explorer la relation qui existe entre différents articles achetés sur un site de commerce électronique ou dans un magasin classique de vente au détail.

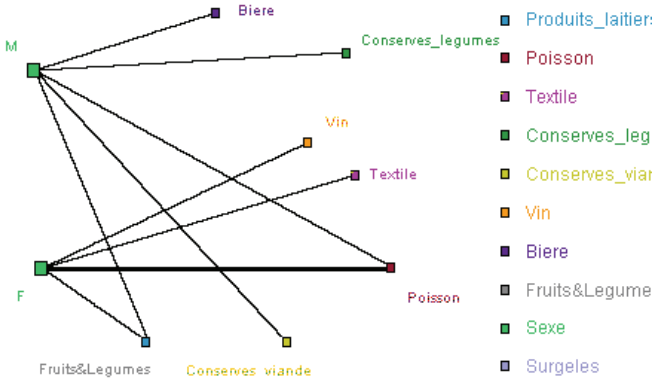
Figure 5-63
Graphique Relations montrant les relations entre les articles d'épicerie achetés



Relations orientées

Les noeuds Relations orientées sont semblables aux noeuds Relations car ils montrent la force des relations entre des champs symboliques. Cependant, les graphiques Relations orientées affichent uniquement les connexions d'un ou de plusieurs champs A partir de vers un seul champ Vers. Les connexions sont unidirectionnelles car ce sont des connexions unilatérales.

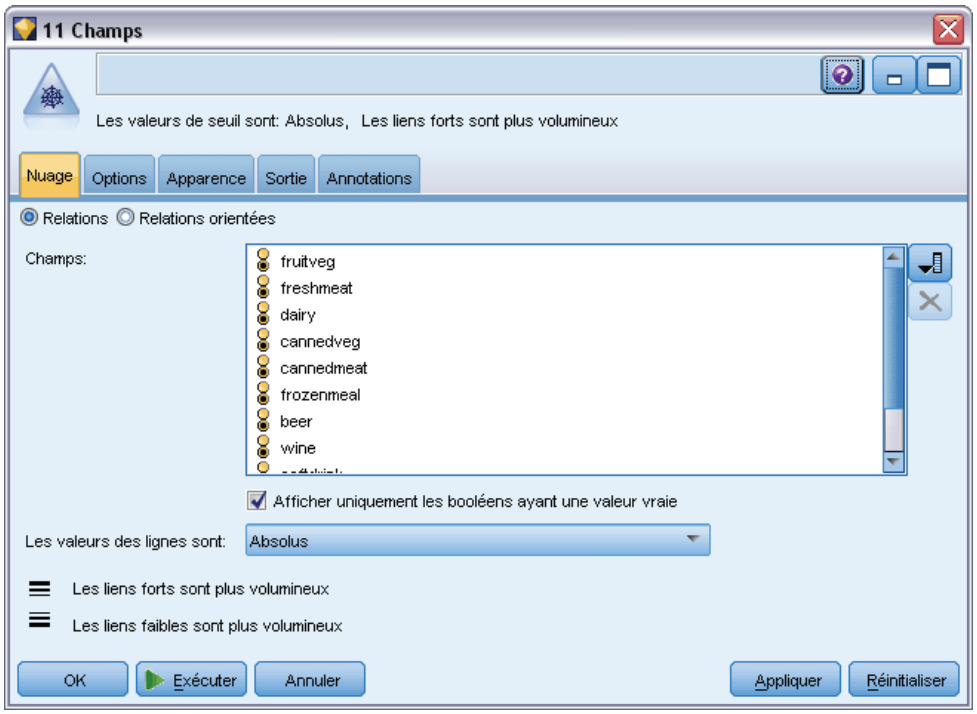
Figure 5-64
Graphique Relations orientées montrant la relation entre les articles d'épicerie achetés et le sexe de l'individu



Comme pour les noeuds Relations, le graphique affiche les connexions à l'aide de divers types de ligne indiquant la force de la connexion. Par exemple, vous pouvez utiliser un noeud Relations orientées pour explorer la relation existant entre le sexe de l'individu et la propension à acheter certains articles.

Onglet Graphique relations

Figure 5-65
Paramètres de l'onglet Nuage pour un noeud Relations



Relations. Sélectionnez cette option pour créer un graphique Relations illustrant la force des relations entre tous les champs spécifiés.

Relations orientées. Sélectionnez cette option pour créer un graphique Relations orientées illustrant la force des relations entre plusieurs champs et les valeurs d'un champ, comme le sexe d'un individu ou la religion. Lorsque cette option est sélectionnée, Champ Vers est activé et la commande Champs plus bas est renommée Champs A partir de pour plus de clarté.

Figure 5-66
Options des relations orientées

Champ Vers (relations orientées uniquement). Sélectionnez un champ booléen ou un champ nominal utilisé pour une relation orientée. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés.

Champs/Champs A partir de. Sélectionnez les champs permettant de créer le graphique Relations. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés. Utilisez le sélecteur de champs pour sélectionner plusieurs champs ou sélectionner les champs par type.

Remarque : dans les relations orientées, cette commande sert à sélectionner les champs A partir de.

Afficher uniquement les booléens ayant une valeur vraie. Sélectionnez cette option pour afficher uniquement les booléens ayant une valeur true (vrai) pour un champ booléen. Cette option simplifie l'affichage des relations et est souvent utilisée avec les données pour lesquelles l'occurrence des valeurs positives est particulièrement élevée.

Les valeurs des lignes sont. Sélectionnez le type de seuil dans la liste déroulante.

- L'option Absolus définit les seuils en fonction du nombre d'enregistrements contenant chaque paire de valeurs.
- L'option Pourcentages globaux indique le nombre absolu d'observations représentées par le lien sous la forme d'une proportion de toutes les occurrences de chaque paire de valeurs représentée dans le graphique Relations.
- Les options Pourcentages de la plus petite valeur/du plus petit champ et Pourcentages de la plus grande valeur/du plus grand champ indiquent le champ/la valeur à utiliser pour l'évaluation des pourcentages. Supposons, par exemple, que 100 enregistrements comportent la valeur *drugY* dans le champ *Médicament* et que seuls 10 enregistrements comportent la valeur *FAIBLE* dans le champ *BP*. Si 7 enregistrements comportent les deux valeurs *drugY* et *FAIBLE*, ce pourcentage est égal à 70 % ou à 7 %, selon le champ référencé, le plus petit (*BP*) ou le plus grand (*Drug*).

Remarque : avec les graphiques Relations orientées, les troisième et quatrième options mentionnées ci-dessus ne sont pas disponibles. En revanche, vous pouvez sélectionner Pourcentage de valeur/champ Vers et Pourcentage de valeur/champ A partir de.

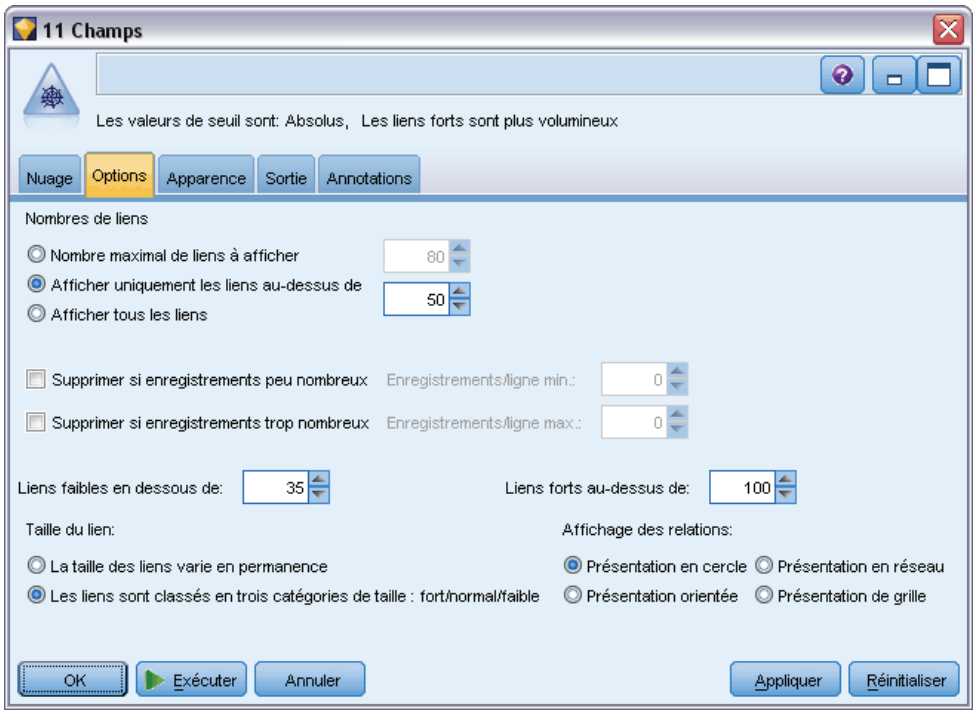
Les liens forts sont plus volumineux.Sélectionnée par défaut, cette option correspond à la méthode standard d'affichage des liens entre les champs.

Les liens faibles sont plus volumineux.Sélectionnez cette option pour inverser la signification des liens représentés par des lignes en gras. Cette option est fréquemment utilisée pour détecter des fraudes ou examiner des valeurs éloignées.

Onglet Options de relations

L'onglet Options des noeuds Relations contient plusieurs options supplémentaires permettant de personnaliser le graphique de sortie.

Figure 5-67
Paramètres de l'onglet Options du noeud Relations



Nombres de liens. Les options suivantes permettent de déterminer le nombre de liens affichés dans le graphique de sortie. Certaines de ces options, telles que Liens faibles au-dessus de et Liens forts au-dessus de, sont également disponibles dans la fenêtre du graphique de sortie. Dans le graphique final, vous pouvez également utiliser un curseur de défilement pour rectifier le nombre de liens affichés.

- **Nombre maximal de liens à afficher.** Choisissez un chiffre indiquant le nombre maximal de liens à afficher dans le graphique de sortie. Utilisez les flèches pour rectifier la valeur.
- **Afficher uniquement les liens au-dessus de.** Choisissez un chiffre indiquant la valeur minimale pour laquelle afficher une connexion dans le graphique Relations. Utilisez les flèches pour rectifier la valeur.
- **Afficher tous les liens.** Sélectionnez cette option pour afficher tous les liens sans tenir compte des valeurs minimale ou maximale. L'activation de cette option peut accroître le temps de traitement si le nombre de champs est important.

Supprimer si enregistrements peu nombreux. Sélectionnez cette option pour ignorer les connexions qui ne comportent que peu d'enregistrements. Définissez le seuil de cette option en entrant un nombre dans le champ Enregistrements/ligne min..

Supprimer si enregistrements trop nombreux. Sélectionnez cette option pour ignorer les connexions comportant un nombre élevé d'enregistrements. Entrez un nombre dans le champ Enregistrements/ligne max..

Liens faibles en dessous de. Choisissez un chiffre indiquant le seuil entre les connexions faibles (lignes en pointillé) et les connexions standard (lignes normales). Toutes les connexions au-dessous de cette valeur sont considérées comme faibles.

Liens forts au-dessus de. Choisissez un chiffre indiquant le seuil entre les connexions fortes (lignes en gras) et les connexions standard (lignes normales). Toutes les connexions au-dessus de cette valeur sont considérées comme fortes.

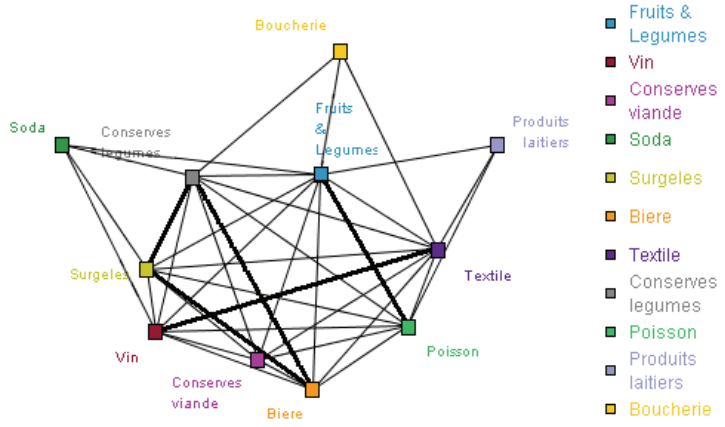
Taille du lien. Choisissez les options permettant de déterminer la taille des liens :

- **La taille des liens varie en permanence.** Sélectionnez cette option pour afficher une amplitude de tailles de lien reflétant la variation des forces de connexion en fonction des valeurs de données réelles.
- **Les liens sont classés en trois catégories de taille : fort/normal/faible.** Sélectionnez cette option pour afficher trois forces de connexion : forte, normale et faible. Les points de césure de ces catégories peuvent être définis grâce aux options ci-avant, ainsi que dans le graphique final.

Affichage des relations. Sélectionnez le type d'affichage des relations :

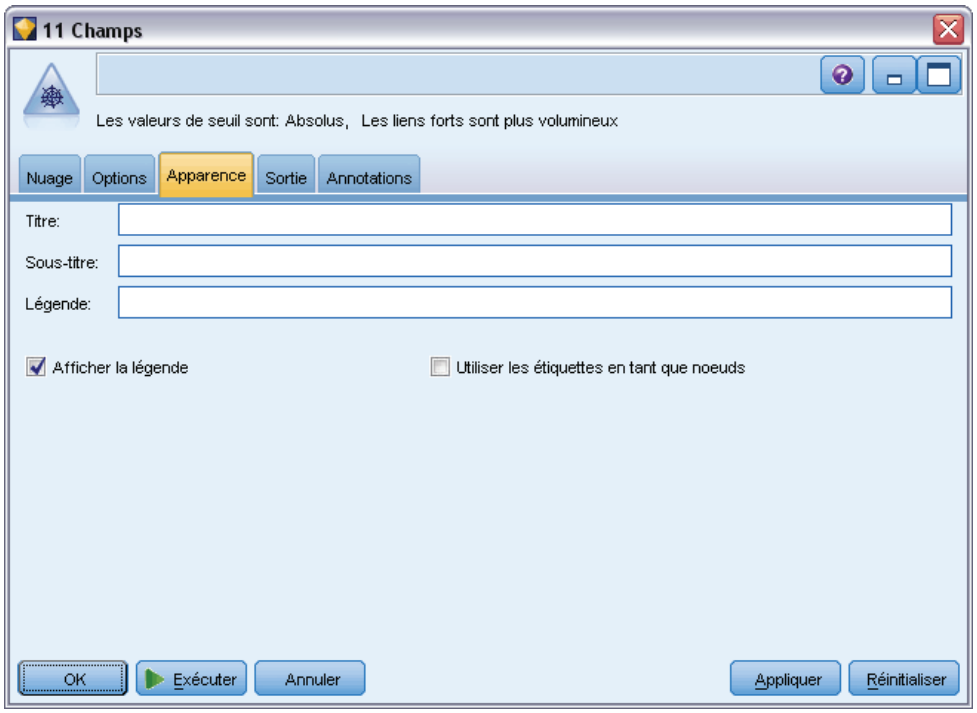
- **Présentation en cercle.** Sélectionnez cette option pour utiliser l'affichage standard des relations.
- **Présentation en réseau.** Sélectionnez cette option pour utiliser un algorithme afin de regrouper les liens les plus forts. L'objectif est de mettre en évidence les liens forts en utilisant la différenciation spatiale et l'épaisseur des lignes.
- **Présentation orientée.** Sélectionnez cette option pour créer un affichage des relations orientées utilisant la sélection Champ Vers de l'onglet Nuage comme cible de la direction.
- **Présentation de grille.** Sélectionnez cette option pour créer un affichage des relations présenté sur une grille régulière.

Figure 5-68
Graphique Relations montrant des connexions fortes entre les surgelés et les conserves de légumes d'une part, et les autres articles d'épicerie d'autre part



Onglet Apparence relations

Figure 5-69
Paramètres de l'onglet Apparence pour un noeud Relations



Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

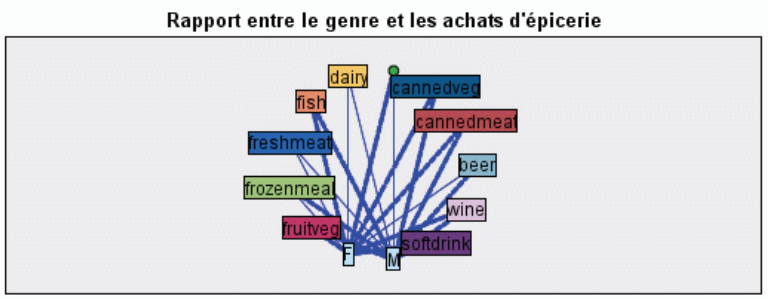
Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Afficher la légende. Vous pouvez spécifier si la légende apparaît. Masquer la légende peut améliorer l'apparence des nuages comportant de nombreux champs.

Utiliser les étiquettes en tant que noeuds. Vous pouvez insérer le texte de l'étiquette dans chaque noeud au lieu d'afficher les étiquettes côte à côte. Pour les nuages dotés de peu de champs, cela risque de rendre le graphique plus lisible.

Figure 5-70
Graphique Relations affichant les étiquettes en tant que noeuds



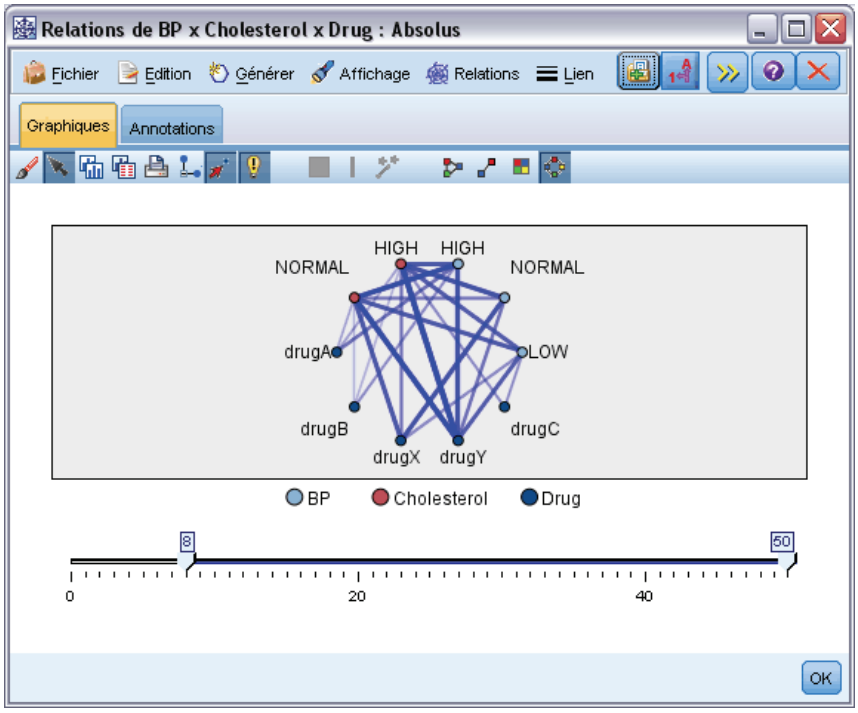
Utilisation d'un graphique Relations

Les noeuds Relations sont utilisés pour montrer la force des liens entre les valeurs de plusieurs champs symboliques. Les connexions sont affichées dans un graphique composé de différents types de ligne servant à indiquer la force croissante des connexions. Vous pouvez, par exemple, utiliser un nœud Relations pour explorer le lien entre les niveaux de cholestérol et de tension artérielle, et le médicament le plus efficace pour traiter la maladie d'un patient.

- Les connexions fortes sont représentées à l'aide de lignes en gras. Ceci indique que les deux valeurs sont fortement liées et nécessitent une attention particulière.
- Les connexions moyennes sont représentées par des lignes d'épaisseur normale.
- Les connexions faibles sont représentées par des lignes en pointillé.
- Si aucune ligne n'apparaît entre deux valeurs, cela signifie soit que les deux valeurs n'apparaissent jamais dans le même enregistrement, soit que le nombre d'enregistrements contenant cette combinaison est inférieur au seuil défini dans la boîte de dialogue du nœud Relations.

Une fois que vous avez créé un nœud Relations, vous disposez de plusieurs options pour ajuster l'affichage du graphique et générer des noeuds pour une analyse plus approfondie.

Figure 5-71
Graphique Relations indiquant un certain nombre de relations fortes, comme celle entre la tension artérielle normale et le médicament MédX, ou entre le taux de cholestérol élevé et MédY



Pour les noeuds Relations et Relations orientées, vous pouvez :

- Modifier la présentation de l'affichage des relations.
- Masquer des points pour simplifier l'affichage.
- Modifier les seuils qui gèrent les styles de ligne.
- Mettre en surbrillance des lignes entre certaines valeurs pour indiquer qu'il s'agit d'une relation « sélectionnée ».
- Générer un nœud Sélectionner pour un ou plusieurs enregistrements « sélectionnés », ou un nœud booléen Calculer associé à une ou plusieurs relations du graphique Relations

Pour ajuster des points

- **Déplacez** les points en cliquant dessus à l'aide de la souris et en les faisant glisser jusqu'à l'emplacement voulu. Le graphique sera redessiné pour faire apparaître le nouvel emplacement.
- **Masquez** les points en cliquant dessus à l'aide du bouton droit de la souris et en sélectionnant Masquer ou Masquer et redessiner dans le menu contextuel. L'option Masquer masque uniquement le point sélectionné et toute ligne associée à ce point. Masquer et redessiner

redessine le graphique, en tenant compte de vos modifications. Toutes les modifications manuelles sont annulées.

- **Affichez** tous les points masqués en sélectionnant Tout afficher ou Tout afficher et redessiner dans le menu Relations de la fenêtre du graphique. Sélectionnez Tout afficher et redessiner pour redessiner le graphique et effectuer les ajustements nécessaires pour inclure tous les points précédemment masqués, ainsi que leurs connexions.

Pour sélectionner (ou mettre en évidence) des lignes

Les lignes sélectionnés sont surlignées en rouge.

- Pour sélectionner une seule ligne, cliquez dessus avec le bouton gauche.
- Pour sélectionnez plusieurs lignes, effectuez l'une des actions suivantes :
  - A l'aide du curseur, dessinez un cercle autour des points des lignes que vous souhaitez sélectionnez.
  - Maintenez la touche CTRL enfoncée et cliquez sur les lignes à sélectionner avec le bouton gauche.

Vous pouvez désélectionner toutes les lignes sélectionnées en cliquant sur l'arrière-plan du graphique, ou en choisissant Effacer la sélection dans le menu Relations de la fenêtre du graphique.

Pour visualiser le graphique Relations à l'aide d'une autre présentation

- Dans le menu Relations, choisissez Présentation en cercle, Présentation en réseau, Présentation orientée ou Présentation de grille pour modifier la présentation du graphique.

Activation ou désactivation du curseur des liens

- Dans le menu Affichage, choisissez Curseur des liens.

Pour sélectionner les enregistrements d'une relation unique ou leur ajouter un booléen

- Cliquez avec le bouton droit de la souris sur la ligne représentant la relation voulue.
- Dans le menu contextuel, choisissez Générer le nœud Sélectionner pour le lien ou Générer le nœud Calculer pour le lien.

  Un nœud Sélectionner ou Calculer est automatiquement ajouté dans l'espace de travail du flux avec les options et les conditions appropriées définies :

  - Le nœud Sélectionner sélectionne tous les enregistrements de la relation donnée.
  - Le noeud Calculer génère un booléen qui indique si la relation sélectionnée est valide pour tous les enregistrements de l'ensemble de données. Le nom du champ booléen correspond à l'association (à l'aide d'un trait de soulignement) des deux valeurs constituant la relation, comme *FAIBLE_MédC* ou *MédC_FAIBLE*.

Pour sélectionner les enregistrements d'un groupe de relations ou leur ajouter un booléen

- Sélectionnez dans le graphique Relations les lignes représentant les relations voulues.

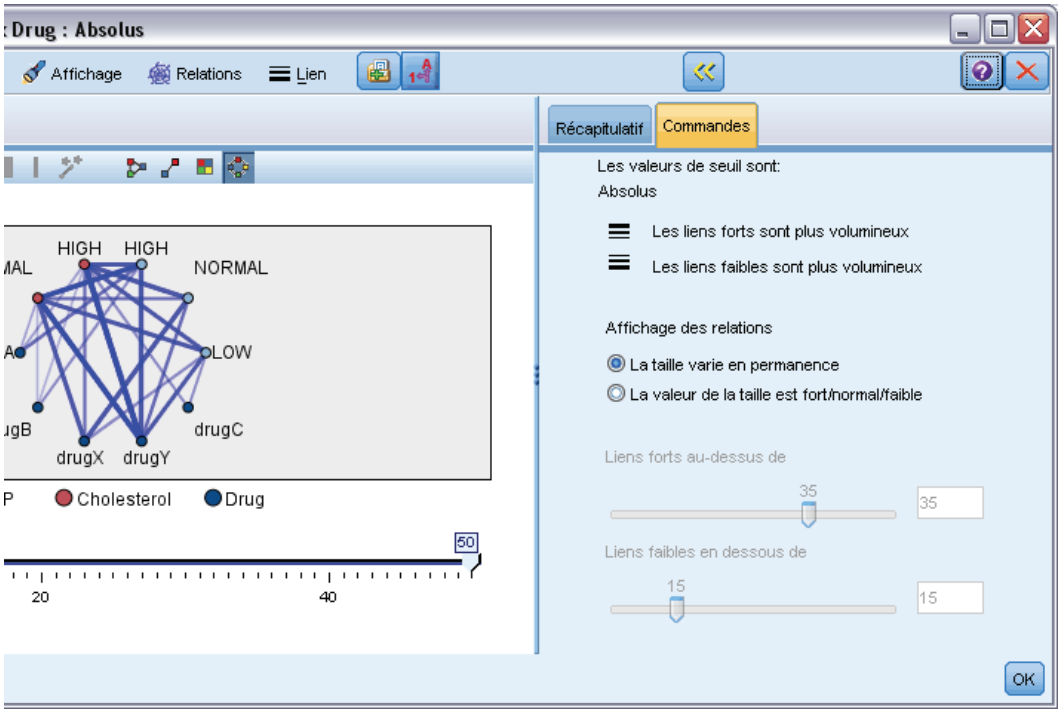
- Dans le menu Générer de la fenêtre du graphique, sélectionnez Noeud Sélectionner (Et), Noeud Sélectionner (Ou), Noeud Calculer (Et) ou Noeud Calculer (Ou).
  - Les noeuds « ou » donnent la disjonction des conditions. Autrement dit, le noeud est appliqué aux enregistrements pour lesquels l'une des relations sélectionnées est valide.
  - Les noeuds « et » donnent la conjonction des conditions. Autrement dit, le noeud est appliqué uniquement aux enregistrements pour lesquels toutes les relations sélectionnées sont valides. Une erreur se produit si certaines des relations sélectionnées s'excluent mutuellement.

  Une fois votre sélection effectuée, un noeud Sélectionner ou Calculer est automatiquement ajouté dans l'espace de travail de flux avec les options et les conditions appropriées définies.

Ajustement des seuils des graphiques Relations

Une fois le graphique Relations créé, vous pouvez ajuster les seuils qui gèrent les styles des lignes à l'aide du curseur de la barre d'outils pour modifier la ligne visible minimale. Vous pouvez également afficher d'autres options de seuil en cliquant sur le bouton de la barre d'outils représentant une double-flèche jaune, afin d'agrandir la fenêtre du graphique Relations. Cliquez ensuite sur l'onglet Commandes pour afficher les options supplémentaires.

Figure 5-72
Fenêtre agrandie comportant les options d'affichage et de seuil



Les valeurs de seuil sont. Affiche le type de seuil sélectionné lors de la création dans la boîte de dialogue du noeud Relations.

Les liens forts sont plus volumineux. Sélectionnée par défaut, cette option correspond à la méthode standard d'affichage des liens entre les champs.

Les liens faibles sont plus volumineux. Sélectionnez cette option pour inverser la signification des liens représentés par des lignes en gras. Cette option est fréquemment utilisée pour détecter des fraudes ou examiner des valeurs éloignées.

Affichage des relations. Choisissez les options permettant de déterminer la taille des liens du graphique de sortie :

- **La taille varie en permanence.** Sélectionnez cette option pour afficher une amplitude de tailles de lien reflétant la variation des forces de connexion en fonction des valeurs de données réelles.
- **La valeur de la taille est fort/normal/faible.** Sélectionnez cette option pour afficher trois forces de connexion : forte, normale et faible. Les points de césure de ces catégories peuvent être définis grâce aux options ci-avant, ainsi que dans le graphique final.

Liens forts au-dessus de. Choisissez un chiffre indiquant le seuil entre les connexions fortes (lignes en gras) et les connexions standard (lignes normales). Toutes les connexions au-dessus de cette valeur sont considérées comme fortes. Utilisez le curseur pour rectifier la valeur ou saisissez un chiffre dans le champ.

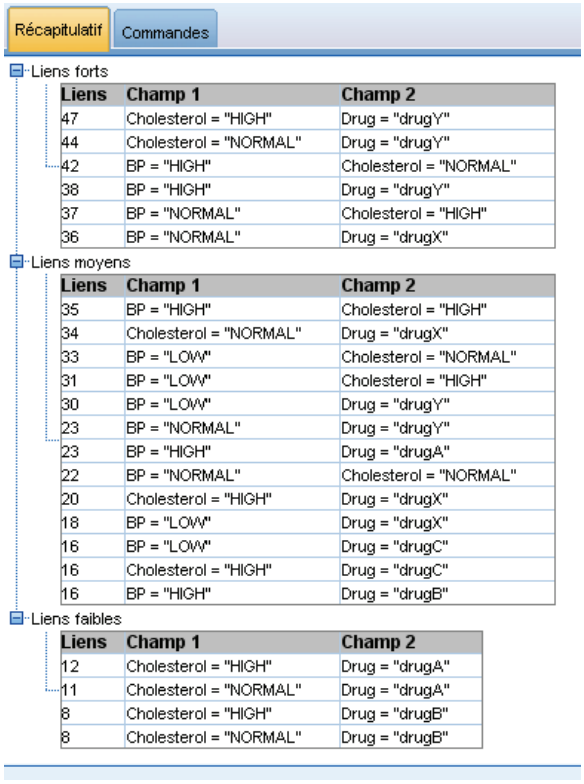
Liens faibles en dessous de. Choisissez un chiffre indiquant le seuil entre les connexions faibles (lignes en pointillé) et les connexions standard (lignes normales). Toutes les connexions au-dessous de cette valeur sont considérées comme faibles. Utilisez le curseur pour rectifier la valeur ou saisissez un chiffre dans le champ.

Après avoir ajusté les seuils du graphique Relations, vous pouvez réorganiser ou redessiner l'affichage des relations en utilisant les nouvelles valeurs des seuils à travers le menu Relations situé sur la barre d'outils du graphique Relations. Une fois que vous avez trouvé les paramètres révélant les motifs les plus significatifs, vous pouvez mettre à jour les paramètres d'origine du noeud Relations (également appelé noeud Relations parent) en sélectionnant Mettre à jour le noeud parent dans le menu Relations de la fenêtre du graphique.

Création d'un récapitulatif des relations

Vous pouvez créer un récapitulatif des relations répertoriant les liens forts, moyens et faibles en cliquant sur le bouton de la barre d'outils représentant une double-flèche jaune, afin d'agrandir la fenêtre du graphique Relations. Cliquez ensuite sur l'onglet Récapitulatif pour afficher les tableaux de chaque type de lien. Vous pouvez agrandir ou réduire les tableaux en utilisant le bouton bascule correspondant.

Figure 5-73
Récapitulatif de relations répertoriant les connexions entre la tension artérielle, le cholestérol et le type de médicament


Récapitulatif | Commandes

Liens forts

Liens	Champ 1	Champ 2
47	Cholesterol = "HIGH"	Drug = "drugY"
44	Cholesterol = "NORMAL"	Drug = "drugY"
42	BP = "HIGH"	Cholesterol = "NORMAL"
38	BP = "HIGH"	Drug = "drugY"
37	BP = "NORMAL"	Cholesterol = "HIGH"
36	BP = "NORMAL"	Drug = "drugX"

Liens moyens

Liens	Champ 1	Champ 2
35	BP = "HIGH"	Cholesterol = "HIGH"
34	Cholesterol = "NORMAL"	Drug = "drugX"
33	BP = "LOW"	Cholesterol = "NORMAL"
31	BP = "LOW"	Cholesterol = "HIGH"
30	BP = "LOW"	Drug = "drugY"
23	BP = "NORMAL"	Drug = "drugY"
23	BP = "HIGH"	Drug = "drugA"
22	BP = "NORMAL"	Cholesterol = "NORMAL"
20	Cholesterol = "HIGH"	Drug = "drugX"
18	BP = "LOW"	Drug = "drugX"
16	BP = "LOW"	Drug = "drugC"
16	Cholesterol = "HIGH"	Drug = "drugC"
16	BP = "HIGH"	Drug = "drugB"

Liens faibles

Liens	Champ 1	Champ 2
12	Cholesterol = "HIGH"	Drug = "drugA"
11	Cholesterol = "NORMAL"	Drug = "drugA"
8	Cholesterol = "HIGH"	Drug = "drugB"
8	Cholesterol = "NORMAL"	Drug = "drugB"

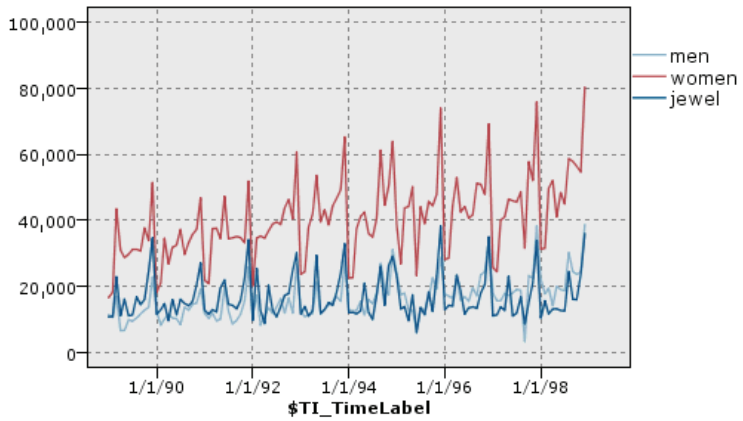
Pour imprimer le récapitulatif, choisissez l'option suivante dans le menu de la fenêtre du graphique Relations :

Fichier > Imprimer le récapitulatif

Noeud Tracé horaire

Les noeuds Tracé horaire vous permettent de visualiser la représentation d'une ou de plusieurs séries temporelles au fil du temps. Les séries représentées doivent contenir des valeurs numériques et sont supposées avoir lieu sur une durée au sein de laquelle les périodes sont uniformes. Utilisez un noeud Intervalle de temps avant un noeud Tracé horaire pour créer un champ *TimeLabel*, lequel est utilisé par défaut pour désigner l'axe *x* des graphiques. Pour plus d'informations, reportez-vous à la section Noeud Intervalles de temps dans le chapitre 4 sur p. 219.

Figure 5-74
Représentation des ventes de vêtements et de bijoux homme et femme dans le temps

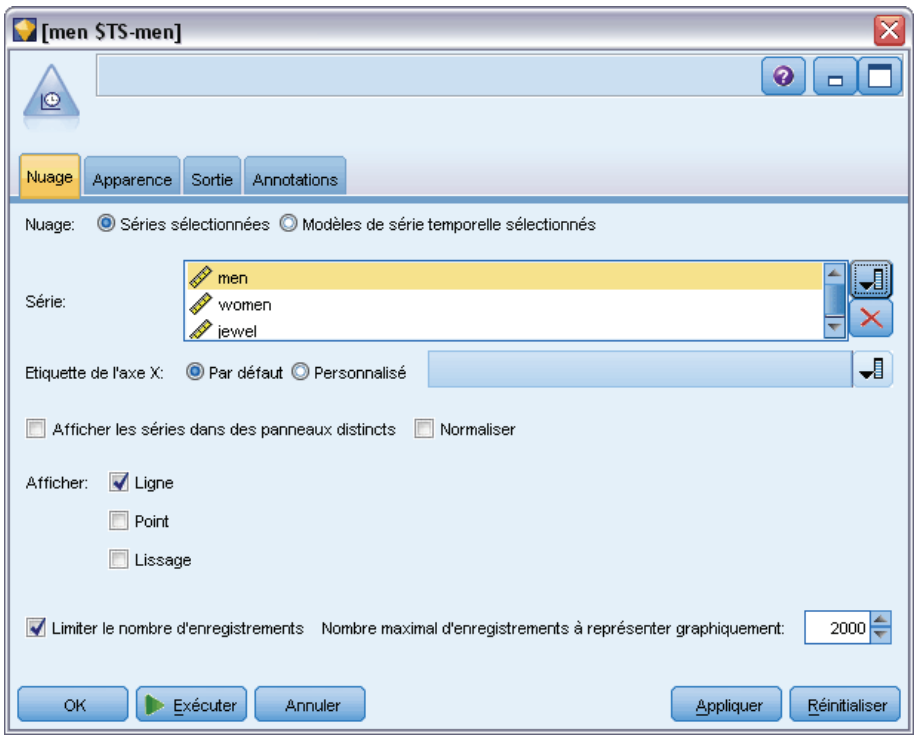


Création d'interventions et d'événements

Vous pouvez créer des champs d'événement et d'intervention à partir du Tracé horaire en générant un noeud Calculer (booléen ou nominal) à partir des menus contextuels. Par exemple, vous pourriez créer un champ d'événements en cas de grève des chemins de fer, où les conditions de circulation sont True (vrai) si l'événement est survenu et False (faux) dans le cas contraire. Pour un champ d'intervention, une augmentation de prix par exemple, vous pouvez utiliser un nœud de calcul pour identifier la date de l'augmentation, avec 0 pour l'ancien prix et 1 pour le nouveau. Pour plus d'informations, reportez-vous à la section Noeud Calculer dans le chapitre 4 sur p. 166.

Onglet Tracé horaire

Figure 5-75
Paramètres de l'onglet Nuage pour un noeud Tracé horaire



Nuage. Permet de choisir comment tracer les séries temporelles.

- **Séries sélectionnées.** Trace des valeurs pour les séries temporelles sélectionnées. Si vous sélectionnez cette option lors du tracé des intervalles de confiance, désélectionnez la case Normaliser.
- **Modèles de série temporelle sélectionnés**. Utilisée en association avec un modèle de séries temporelles, cette option trace tous les champs liés (valeurs réelles et prédites, et intervalles de confiance) pour une ou plusieurs séries temporelles sélectionnées. Cette option désactive d'autres options de la boîte de dialogue. Il s'agit de l'option recommandée pour le tracé d'intervalles de confiance.

Séries. Sélectionnez un ou plusieurs champs contenant des séries temporelles à représenter. Il doit s'agir de données numériques.

Etiquette de l'axe X. Choisissez l'étiquette par défaut ou un champ unique à utiliser en tant qu'étiquette de l'axe *x* dans les graphiques Nuage. Si vous choisissez Par défaut, le système utilise le champ TimeLabel créé à partir d'un noeud Intervalles de temps en amont ou d'entiers séquentiels s'il n'existe pas de noeud Intervalles de temps. Pour plus d'informations, reportez-vous à la section Noeud Intervalles de temps dans le chapitre 4 sur p. 219.

Afficher les séries dans des panneaux distincts. Indique si chaque série apparaît dans un panneau distinct. Si vous ne choisissez pas cette option, toutes les séries temporelles sont représentées sur le même graphique et les lissages ne sont pas disponibles. Dans le cas d'une représentation sur un même graphique, chaque série arbore une couleur différente.

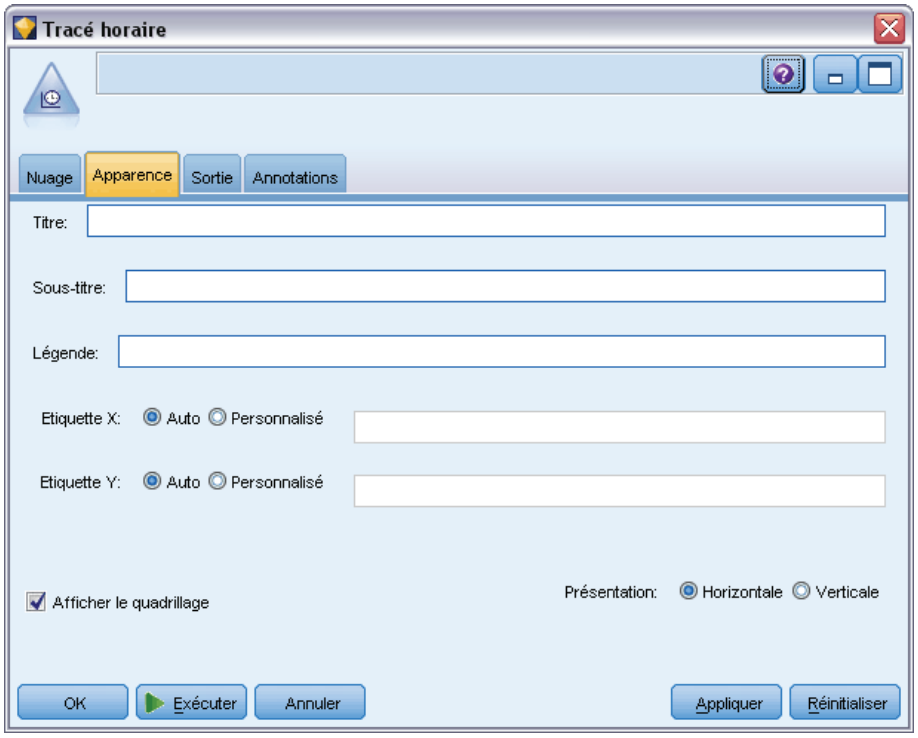
Normaliser. Sélectionnez cette option pour mettre toutes les valeurs *Y* à l'échelle sur l'intervalle 0–1 afin de les afficher sur le graphique. La fonction de normalisation vous permet d'explorer les relations existant entre les lignes, relations qui risqueraient sinon d'être occultées en raison des différences au niveau de l'intervalle de valeurs de chaque série ; il est recommandé de l'utiliser lorsque vous représentez plusieurs lignes sur le même graphique ou lorsque vous comparez des graphiques dans des panneaux mitoyens. (Il est inutile d'appliquer une normalisation lorsque toutes les valeurs de données sont comprises dans un même intervalle.)

Afficher : Sélectionnez un ou plusieurs éléments à afficher dans votre graphique Nuage. Vous avez le choix entre des lignes, des points et des lissages (LOESS). Les lissages sont disponibles uniquement si vous affichez les séries dans des panneaux distincts. Par défaut, l'élément ligne est sélectionné. Veillez à sélectionner au moins un élément de graphique Nuage avant d'exécuter le noeud Graphiques, sinon le système renvoie une erreur indiquant que vous n'avez sélectionné aucun élément à représenter.

Limiter le nombre d'enregistrements. Sélectionnez cette option si vous souhaitez limiter le nombre d'enregistrements représentés. Spécifiez le nombre d'enregistrements, lus à partir du début de votre fichier de données, qui seront représentés dans l'option Nombre maximal d'enregistrements à représenter graphiquement. Ce nombre est défini sur 2 000 par défaut. Pour représenter les *n* derniers enregistrements de votre fichier, vous pouvez utiliser un noeud Trier avant ce noeud pour organiser les enregistrements dans l'ordre temporel décroissant.

Onglet Apparence du tracé horaire

Figure 5-76
Paramètres de l'onglet Apparence du noeud Tracé horaire



Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

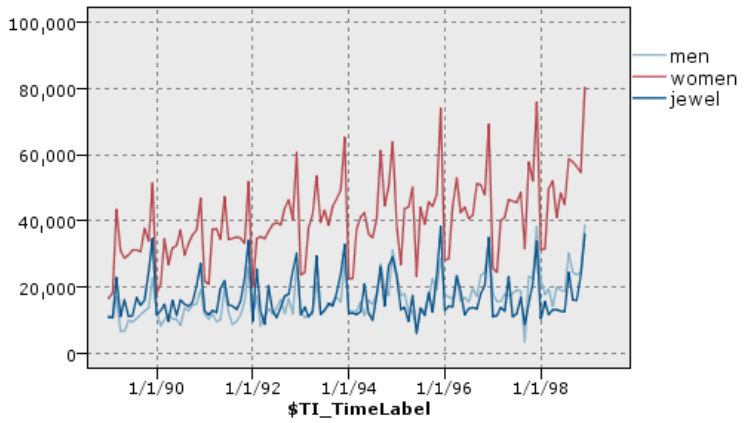
Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Présentation. Pour les tracés horaires uniquement, vous pouvez indiquer si les valeurs temporelles doivent être représentées sur un axe horizontal ou sur un axe vertical.

Utilisation d'un graphique Tracé horaire

Une fois que vous avez créé un graphique Tracé horaire, vous disposez de plusieurs options pour ajuster l'affichage du graphique et générer des noeuds pour une analyse plus approfondie. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360.

Figure 5-77
Représentation des ventes de vêtements et de bijoux homme et femme dans le temps



Après avoir créé un graphique Tracé horaire, défini des bandes et examiné les résultats, vous pouvez utiliser les options du menu Générer et du menu contextuel pour créer des noeuds Sélectionner ou Calculer. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Noeud Evaluation

Le noeud Evaluation permet d'évaluer et de comparer facilement des modèles prédictifs afin de choisir celui le mieux adapté à l'application. Les graphiques Evaluation montrent l'aptitude des modèles à prédire des résultats spécifiques. Ils trient les enregistrements en fonction de la valeur prédite et de la confiance dans cette prévision, divisent les enregistrements en groupes de taille égale (**quantiles**), puis reportent la valeur du critère traité pour chaque quantile, du plus élevé au plus faible. Les divers modèles apparaissent sous forme de lignes dans le graphique.

Les résultats sont traités grâce à la définition d'une valeur ou d'une amplitude de valeurs spécifique en tant que **correspondance**. Les correspondances indiquent généralement une réussite (telle qu'une vente conclue avec un client) ou un événement intéressant (tel qu'un diagnostic médical spécifique). Vous pouvez définir des critères de correspondance dans l'onglet Options de la boîte de dialogue. Vous pouvez également utiliser les critères de correspondance par défaut suivants :

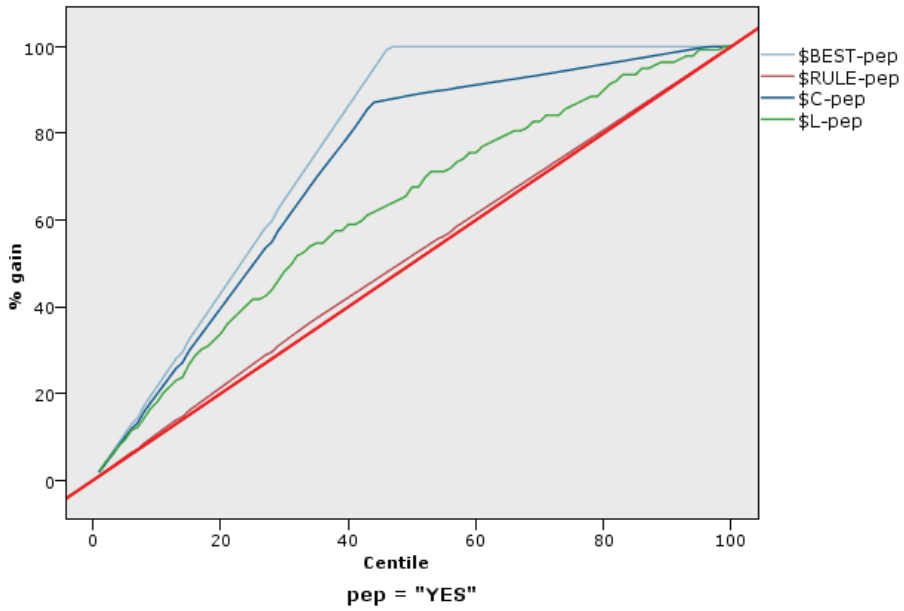
- Les champs de sortie **booléens** sont simples ; les correspondances renvoient à des valeurs *vraies*.
- En ce qui concerne les champs de sortie **nominaux**, c'est la première valeur de l'ensemble qui définit une correspondance.
- Pour les champs de sortie **continus**, les correspondances sont les valeurs supérieures à la moitié de l'intervalle du champ.

Il existe cinq types de graphique Evaluation, chacun mettant en valeur un critère d'évaluation différent :

Graphiques de gains

Les gains sont définis comme la proportion du nombre total de correspondances représentée dans chaque quantile. Les gains sont calculés de la façon suivante : (nombre de correspondances dans le quantile / nombre total de correspondances) × 100 %.

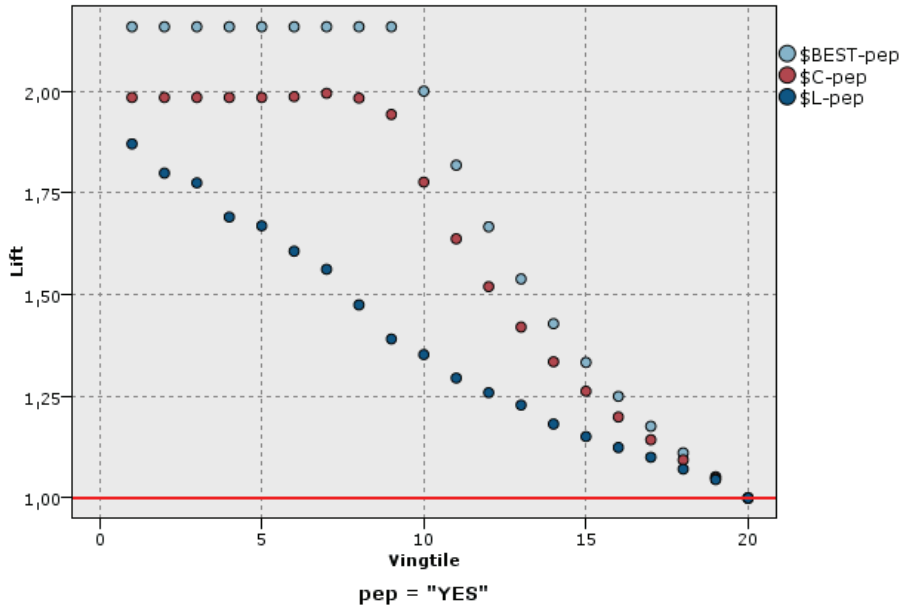
Figure 5-78
Graphique de gains (cumulatif) avec affichage de la ligne de référence, de la meilleure ligne et de la règle de marché



Graphiques Lift

Ces graphiques comparent le pourcentage d'enregistrements dans chaque quantile qui se sont traduits par des correspondances et le pourcentage total de correspondances dans les données d'apprentissage. Le calcul s'effectue de la façon suivante : (correspondances dans le quantile / enregistrements dans le quantile) / (nombre total de correspondances / nombre total d'enregistrements).

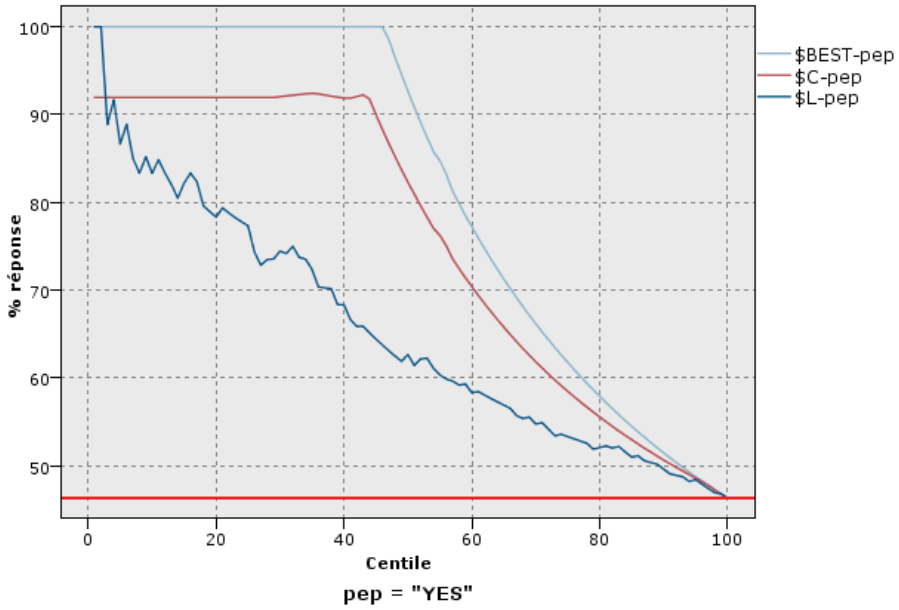
Figure 5-79
Graphique Lift (cumulatif) utilisant des points et la meilleure ligne



Graphiques de réponses

La réponse correspond tout simplement au pourcentage d'enregistrements dans le quantile qui sont des correspondances. La réponse se calcule de la façon suivante : (nombre de correspondances dans le quantile / enregistrements dans le quantile) × 100 %.

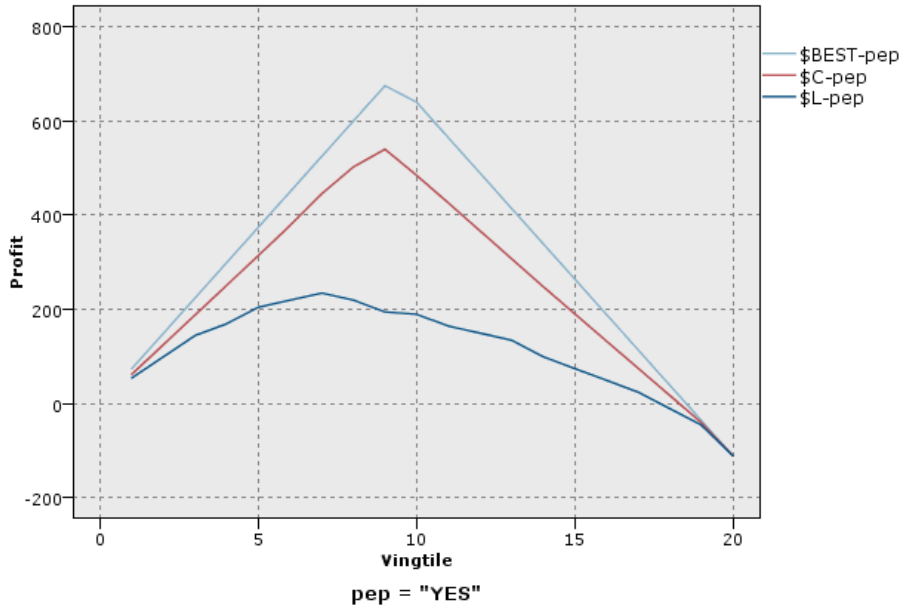
Figure 5-80
Graphique de réponses (cumulatif) affichant la meilleure ligne



Graphiques de profits

Le profit est égal au **revenu** de chaque enregistrement moins le **coût** de l'enregistrement. Les profits d'un quantile correspondent à la somme des profits de tous ses enregistrements. Les revenus sont supposés ne s'appliquer qu'aux correspondances, mais les coûts s'appliquent à tous les enregistrements. Les profits et les coûts peuvent être fixes ou peuvent être déterminés par les champs des données. Les profits sont calculés de la façon suivante : (somme des revenus des enregistrements du quantile − somme des coûts des enregistrements du quantile).

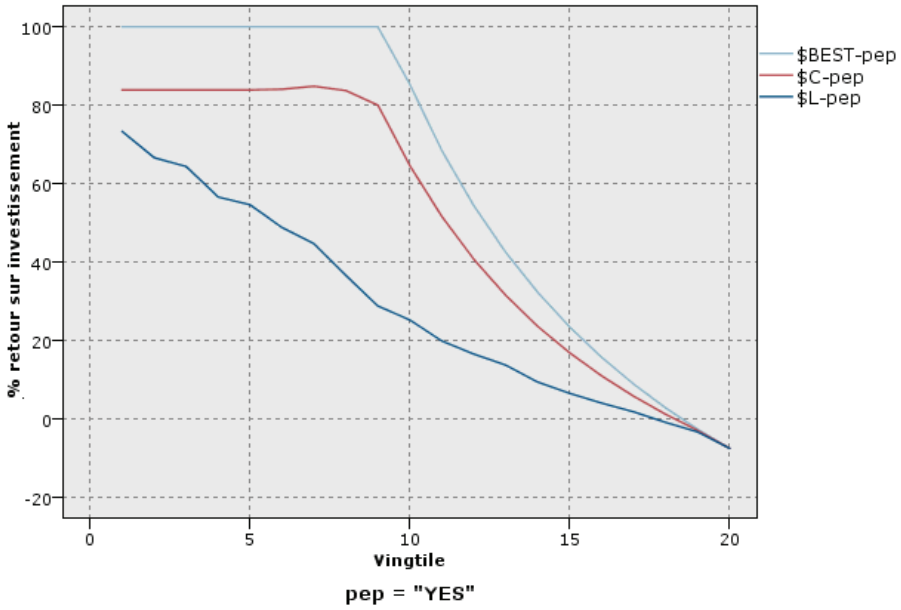
Figure 5-81
Graphique de profits (cumulatif) affichant la meilleure ligne



Graphiques de retour sur investissement

Le retour sur investissement est semblable au profit dans le sens où il s'agit de définir des revenus et des coûts. Le retour sur investissement compare les profits du quantile à ses coûts. Le retour sur investissement se calcule de la façon suivante : (profits du quantile / coûts du quantile) × 100 %.

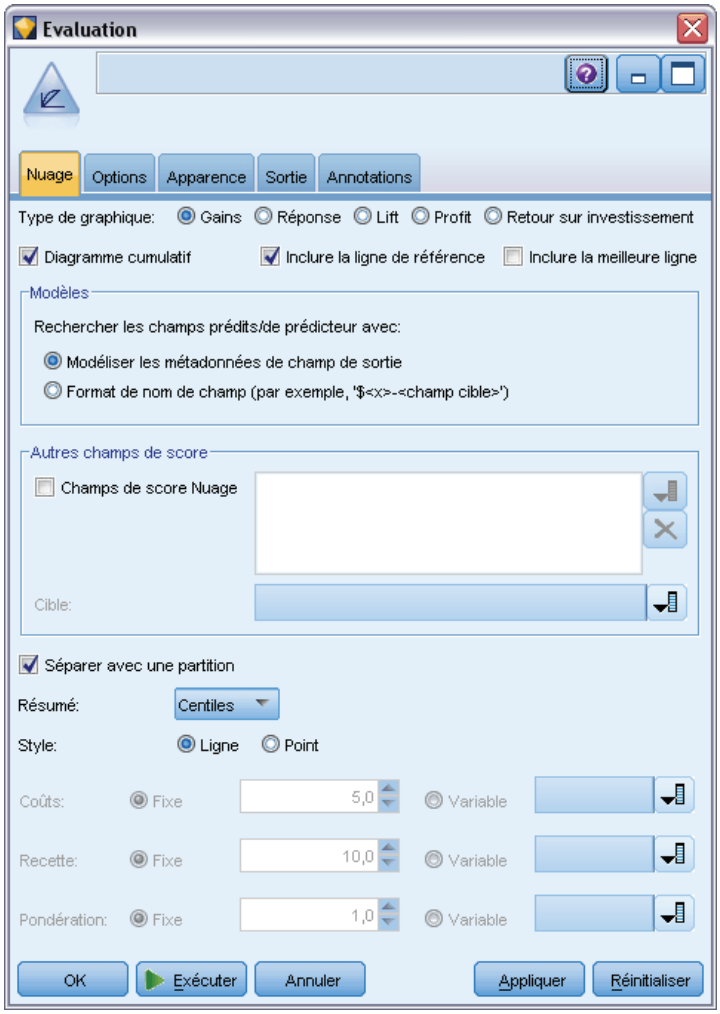
Figure 5-82
Graphique de retour sur investissement (cumulatif) affichant la meilleure ligne



Les graphiques Evaluation peuvent également être cumulatifs. Ainsi, chaque point est égal à la valeur du quantile correspondant, plus celle de tous les quantiles supérieurs. Les graphiques cumulatifs soulignent mieux la performance globale des modèles, alors que les graphiques non cumulatifs permettent de mettre en valeur les zones problématiques des modèles.

Onglet Nuage d'évaluation

Figure 5-83
Paramètres de l'onglet Nuage pour un noeud Evaluation



Type de graphique. Sélectionnez l'un des types suivants : Gains, Réponse, Lift, Profit ou Retour sur investissement.

Diagramme cumulatif. Cochez cette case pour créer un graphique cumulatif. Dans les graphiques cumulatifs, les valeurs reportées correspondent à celles de chaque quantile, plus celles de tous les quantiles supérieurs.

Inclure la ligne de référence. Sélectionnez cette option pour inclure une ligne de référence dans le graphique, qui indique une proportion de correspondances parfaitement aléatoire où la confiance devient inutile. (Cette option n'est pas disponible pour les graphiques de profits et de retour sur investissement.)

Inclure la meilleure ligne. Sélectionnez cette option pour inclure une meilleure ligne dans le graphique, qui indique une confiance parfaite (où les correspondances équivalent à 100 % des observations).

Rechercher les champs prédits/prédicteurs avec. Sélectionnez soit Modéliser les métadonnées de champ de sortie pour rechercher les champs prédits dans le graphique en utilisant leurs métadonnées, soit Format de nom de champ pour les rechercher par nom.

Champs de score Nuage. Cochez cette case pour activer le sélecteur de champs de score. Puis sélectionnez un ou plusieurs champs de score à intervalles ou continus. Ces champs ne sont pas strictement des modèles prédictifs mais peuvent être utiles pour classer des enregistrements selon leur propension à être une correspondance. Le noeud Evaluation peut comparer toute combinaison d'un ou de plusieurs champs de score avec un ou plusieurs modèles prédictifs. Un exemple typique peut être de comparer plusieurs champs RFM avec votre meilleur modèle prédictif.

Cible. Sélectionnez le champ cible à l'aide du sélecteur de champ. Choisissez tout champ nominal ou booléen instancié comportant deux valeurs ou plus.

Remarque : Ce champ cible s'applique seulement aux champs de score (les modèles prédictifs déterminent leurs propres cibles) et est ignoré si un critère de correspondance personnalisé est défini sur l'onglet Options.

Diviser par partition. Si un champ de partition permet de diviser des enregistrements en échantillons d'apprentissage, de test et de validation, sélectionnez cette option pour afficher un graphique Evaluation distinct pour chaque partition. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 sur p. 207.

Remarque : lorsque vous divisez des enregistrements par partition, ceux dont le champ de partition contient des valeurs nulles sont exclus de l'évaluation. Ce problème ne se pose jamais si un noeud Partitionner est utilisé, car ce type de noeud ne génère aucune valeur nulle.

Diagramme : Dans la liste déroulante, sélectionnez la taille des quantiles à représenter sur le graphique. Les options disponibles sont Quartiles, Quintiles, Déciles, Vingtiles, Centiles et 1000-tiles.

Style. Sélectionnez Ligne ou Point.

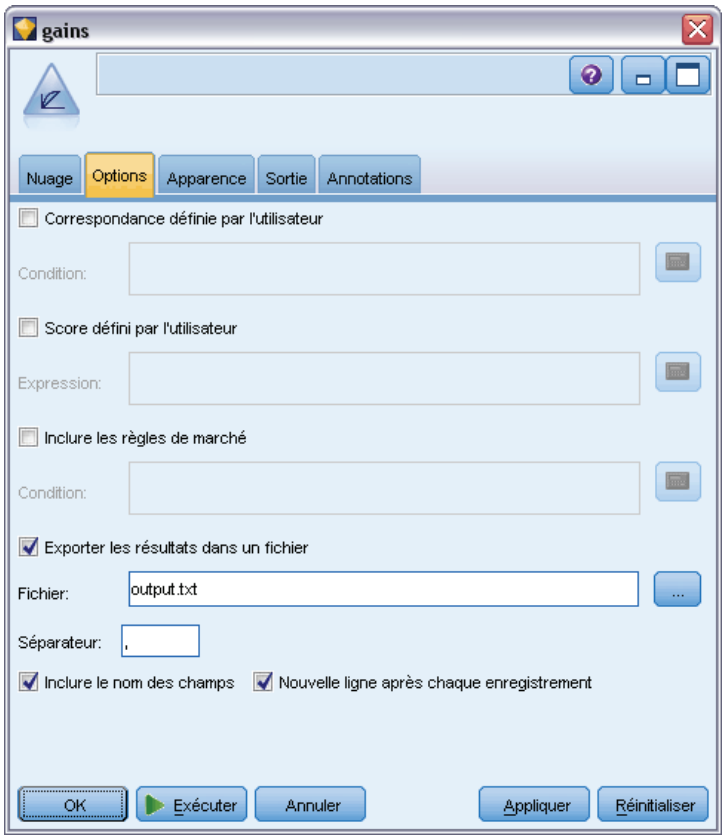
Tableau des profits et retour sur investissement. Pour les graphiques de profits et de retour sur investissement, vous pouvez en outre préciser les coûts, le revenu et la pondération.

- **Coûts.** Précisez le coût associé à chaque enregistrement. Vous pouvez sélectionner Fixe ou Variable. Pour les coûts fixes, précisez la valeur du coût. Pour les coûts variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de coût.
- **Recette.** Précisez le revenu associé à chaque enregistrement représentant une correspondance. Vous pouvez sélectionner Fixe ou Variable. Pour les revenus fixes, précisez la valeur du revenu. Pour les revenus variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de revenu.
- **Pondération.** Si les enregistrements de vos données représentent plusieurs unités, vous pouvez utiliser les pondérations d'effectif pour ajuster les résultats. Indiquez la pondération associée à chaque enregistrement, à l'aide des options Fixe ou Variable. Pour une pondération fixe, précisez la valeur de la pondération (nombre d'unités par enregistrement). Pour une pondération variable, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de pondération.

Onglet Options d'évaluation

L'onglet Options des graphiques Evaluation permet de définir facilement les correspondances, les règles de marché et les critères d'évaluation affichés dans les graphiques. Vous pouvez également définir des options pour exporter les résultats de l'évaluation du modèle.

Figure 5-84
Paramètres de l'onglet Options pour un noeud Evaluation



Correspondance définie par l'utilisateur. Sélectionnez cette option pour spécifier la condition personnalisée utilisée pour indiquer une correspondance. Cette option permet de définir les résultats qui vous intéressent au lieu de les déduire du type de champ cible et de l'ordre des valeurs.

- **Condition.** Lorsque l'option Correspondance définie par l'utilisateur est sélectionnée, vous devez indiquer l'expression CLEM de la condition de correspondance. Par exemple, @TARGET = "YES" est une condition valide qui indique que la valeur *Oui* du champ cible sera considérée comme une correspondance lors de l'évaluation. La condition indiquée sera utilisée pour tous les champs cible. Pour créer une condition, entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition. Si les données sont instanciées, vous pouvez insérer des valeurs directement à partir du Générateur de formules.

Score défini par l'utilisateur. Sélectionnez cette option pour indiquer une condition servant à évaluer les observations avant de les affecter à des quantiles. Le score par défaut est calculé à partir de la valeur prédite et de la confiance. Utilisez le champ Expression pour créer une expression d'évaluation personnalisée.

- **Expression.** Indiquez l'expression CLEM utilisée pour l'évaluation. Par exemple, si une sortie numérique de l'intervalle 0–1 est triée afin que les valeurs inférieures soient meilleures que les valeurs supérieures, vous pouvez définir une correspondance supérieure (@TARGET < 0.5), ainsi que le score associé (1 • @PREDICTED). L'expression du score doit correspondre à une valeur numérique. Pour créer une condition, entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition.

Inclure les règles de marché. Sélectionnez cette option pour indiquer une condition de règle reflétant les critères intéressants. Par exemple, vous voudrez peut-être afficher une règle pour tous les cas où mortgage = "Y" and income >= 33000. Les règles de marché apparaissent sur le graphique et sont appelées *Règle* dans la clé.

- **Condition.** Indiquez l'expression CLEM utilisée pour définir une règle de marché dans le graphique de sortie. Entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition. Si les données sont instanciées, vous pouvez insérer des valeurs directement à partir du Générateur de formules.

Exporter les résultats dans un fichier. Sélectionnez cette option pour exporter les résultats de l'évaluation du modèle dans un fichier texte délimité. Vous pouvez lire ce fichier pour réaliser des analyses spécifiques des valeurs calculées. Pour l'exportation, définissez les options suivantes :

- **Nom du fichier.** Entrez le nom du fichier de sortie. Utilisez le bouton ... pour accéder au dossier voulu.
- **Séparateur.** Entrez le caractère, tel qu'une virgule ou un espace, à utiliser comme séparateur de champ.

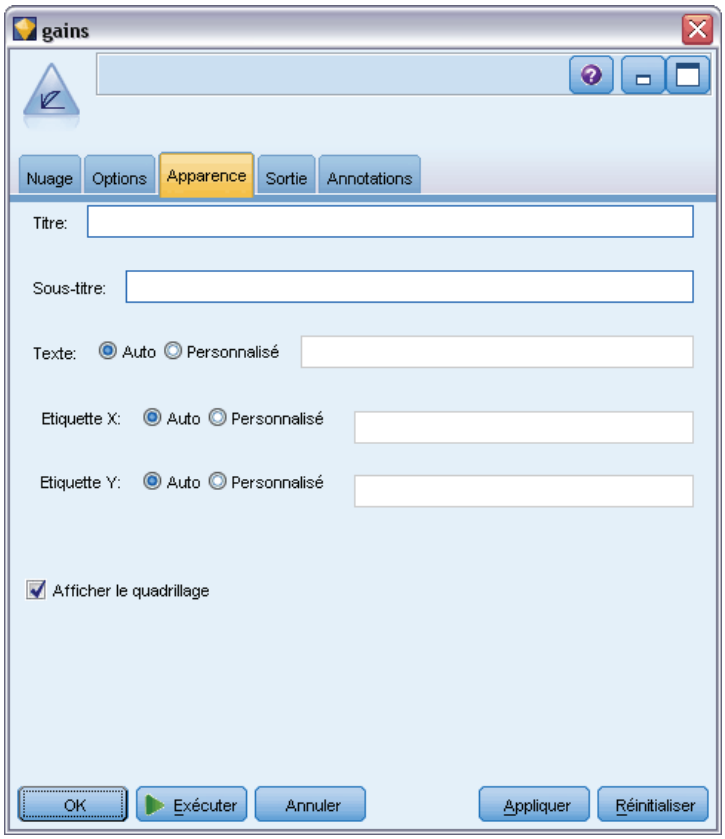
Inclure les noms des champs. Sélectionnez cette option pour inclure les noms de champ sur la première ligne du fichier de sortie.

Nouvelle ligne après chaque enregistrement. Sélectionnez cette option pour commencer chaque enregistrement sur une nouvelle ligne.

Onglet Apparence de l'évaluation

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Figure 5-85
Paramètres de l'onglet Apparence pour un noeud Evaluation



Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous'titre du graphique.

Texte : Vous pouvez soit accepter l'étiquette de texte générée automatiquement, soit sélectionner Personnalisé pour indiquer une étiquette.

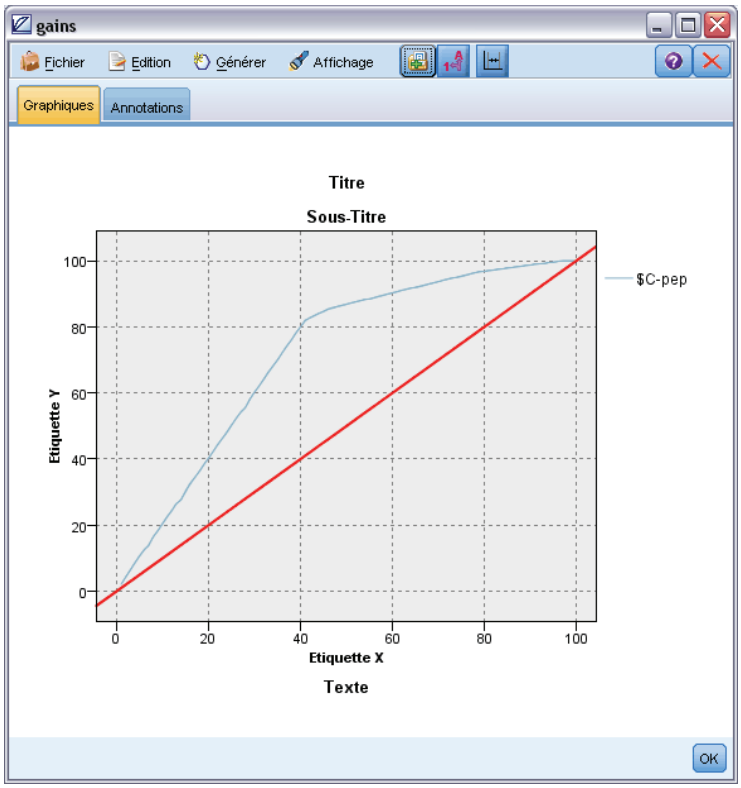
Etiquette X. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *x* (horizontal), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Etiquette Y. Vous pouvez soit accepter l'étiquette générée automatiquement pour l'axe *y* (vertical), soit sélectionner Personnalisé pour indiquer une étiquette personnalisée.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

L'exemple suivant montrent où sont placées sur le graphique les options d'apparence.

Figure 5-86
Position des options d'apparence du graphique sur un graphique d'évaluation



Lecture des résultats d'une évaluation de modèle

L'interprétation d'un graphique Evaluation dépend dans une certaine mesure du type du graphique, mais il existe cependant des caractéristiques communes à tous les graphiques Evaluation. Sur les graphiques cumulatifs, les lignes les plus hautes indiquent les modèles mieux adaptés, tout particulièrement sur la gauche du graphique. Souvent, lors de la comparaison de plusieurs modèles, les lignes se croisent, indiquant qu'un modèle est meilleur sur une partie du graphique et un autre modèle sur une autre partie. Dans ce cas, vous devez prendre en considération la portion de l'échantillon qui vous intéresse (ce qui revient à définir un point sur l'axe *x*) lors du choix du modèle.

La plupart des graphiques non cumulatifs sont très similaires. Dans les modèles satisfaisants, les graphiques non cumulatifs sont hauts sur la gauche et bas sur la droite du graphique. (Si un graphique non cumulatif affiche un motif en dents de scie, vous pouvez le rendre plus régulier en réduisant le nombre de quantiles à reporter et en réexécutant le graphique.) La présence de lignes basses sur la gauche du graphique ou de lignes hautes sur la droite indiquent parfois des zones où les prévisions du modèle sont médiocres. Une ligne droite sur l'ensemble du graphique indique que le modèle ne fournit aucune information.

Graphiques de gains. Les graphiques de gains cumulatifs commencent toujours à 0 % sur la gauche et finissent toujours à 100 % sur la droite. Les diagrammes de gains des bons modèles présentent une hausse rapide en direction de la valeur 100 %, puis se stabilisent. Un modèle ne fournissant

aucune information suit une trajectoire en diagonale du coin inférieur gauche au coin supérieur droit (affiché sur le graphique si l'option Inclure la ligne de référence est sélectionnée).

Graphiques Lift. Les graphiques Lift cumulatifs commencent au-dessus de 1,0 à gauche, puis baissent progressivement jusqu'à atteindre 1,0 à droite. Le bord droit du graphique représente l'intégralité de l'ensemble de données, donc le rapport entre les correspondances des quantiles cumulatifs et les correspondances des données est égal à 1,0. Dans les modèles satisfaisants, le graphique Lift commence bien au-dessus de 1,0 sur la gauche, reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse rapidement vers 1,0 sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour de 1 sur la totalité du graphique. (Si l'option Inclure la ligne de référence est sélectionnée, une ligne de référence horizontale correspondant à la valeur 1 figure sur le graphique.)

Graphiques de réponses. Les graphiques de réponses cumulatifs sont semblables aux graphiques Lift, à l'exception de la mise à l'échelle. Les graphiques de réponses commencent autour de 100 %, puis baissent progressivement jusqu'à atteindre le taux de réponse global (nombre total de correspondances / nombre total d'enregistrements), à droite. Dans les modèles satisfaisants, la ligne commence autour ou à 100 % (sur la gauche), reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse rapidement vers le taux de réponse global sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour du taux de réponse global sur la totalité du graphique. (Si l'option Inclure la ligne de référence est sélectionnée, une ligne de référence horizontale correspondant au taux de réponse global figure sur le graphique.)

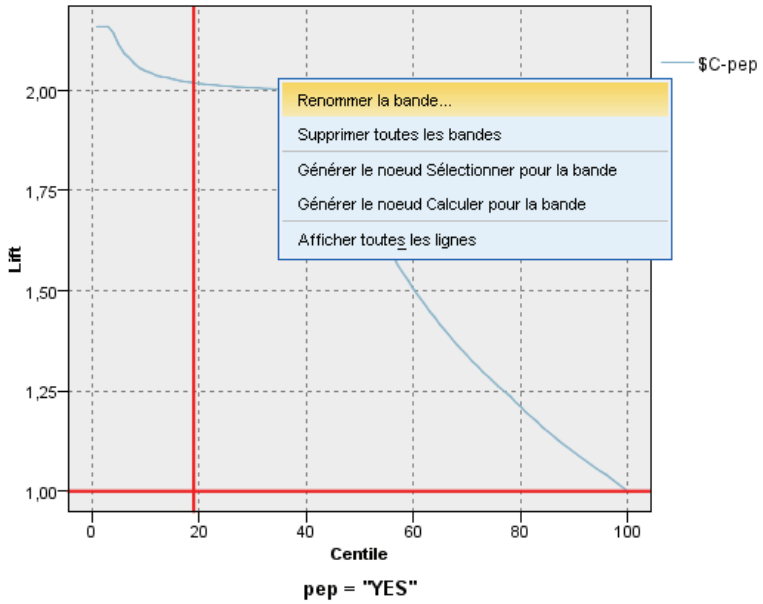
Graphiques de profits. Les graphiques de profits cumulatifs montrent la somme des profits à mesure que vous augmentez la taille de l'échantillon sélectionné (de gauche à droite). Les graphiques de profits commencent généralement autour de 0, augmentent régulièrement à mesure que vous avancez vers la droite jusqu'à atteindre un pic ou un plateau au centre du graphique, puis baissent vers le bord droit du graphique. Dans les modèles satisfaisants, les profits affichent un pic bien défini au centre du graphique. Si le modèle ne fournit aucune information, la ligne est relativement droite et peut augmenter, diminuer ou se stabiliser en fonction de la structure coût/revenu utilisée.

Graphiques de retour sur investissement. Les graphiques de retour sur investissement cumulatifs sont semblables aux graphiques de réponses et aux graphiques Lift, à l'exception de la mise à l'échelle. Les graphiques de retour sur investissement commencent généralement au-dessus de 0 %, puis baissent progressivement jusqu'à atteindre le retour sur investissement global de l'intégralité de l'ensemble de données (qui peut être un nombre négatif). Dans les modèles satisfaisants, la ligne commence bien au-dessus de 0 %, reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse assez rapidement vers le retour sur investissement global sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour de la valeur du retour sur investissement global.

Utilisation d'un graphique Evaluation

Comme dans les graphiques Histogramme et Résumé, vous pouvez utiliser la souris pour explorer les graphiques Evaluation. L'axe *x* représente les scores des modèles dans les quantiles indiqués (vingtiles ou déciles).

Figure 5-87
Utilisation d'un graphique Evaluation



Vous pouvez partitionner l'axe *x* en bandes (comme pour un graphique Histogramme) en utilisant l'icône de fractionnement pour afficher les options permettant de fractionner automatiquement l'axe en bandes égales. Pour plus d'informations, reportez-vous à la section Exploration de graphiques sur p. 360. Vous pouvez éditer manuellement les limites des bandes en sélectionnant Bandes graphiques dans le menu Edition.

Après avoir créé un graphique Evaluation, défini des bandes et examiné les résultats, vous pouvez utiliser les options du menu Générer et du menu contextuel pour créer automatiquement des noeuds basés sur les sélections du graphique. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Lorsque vous générez des noeuds à partir d'un graphique Evaluation, vous êtes invité à sélectionner un modèle parmi tous ceux disponibles dans le graphique.

Figure 5-88
Sélection d'un modèle pour la génération du noeud

Sélectionnez un modèle et cliquez sur OK pour générer le nouveau noeud dans l'espace de travail de flux.

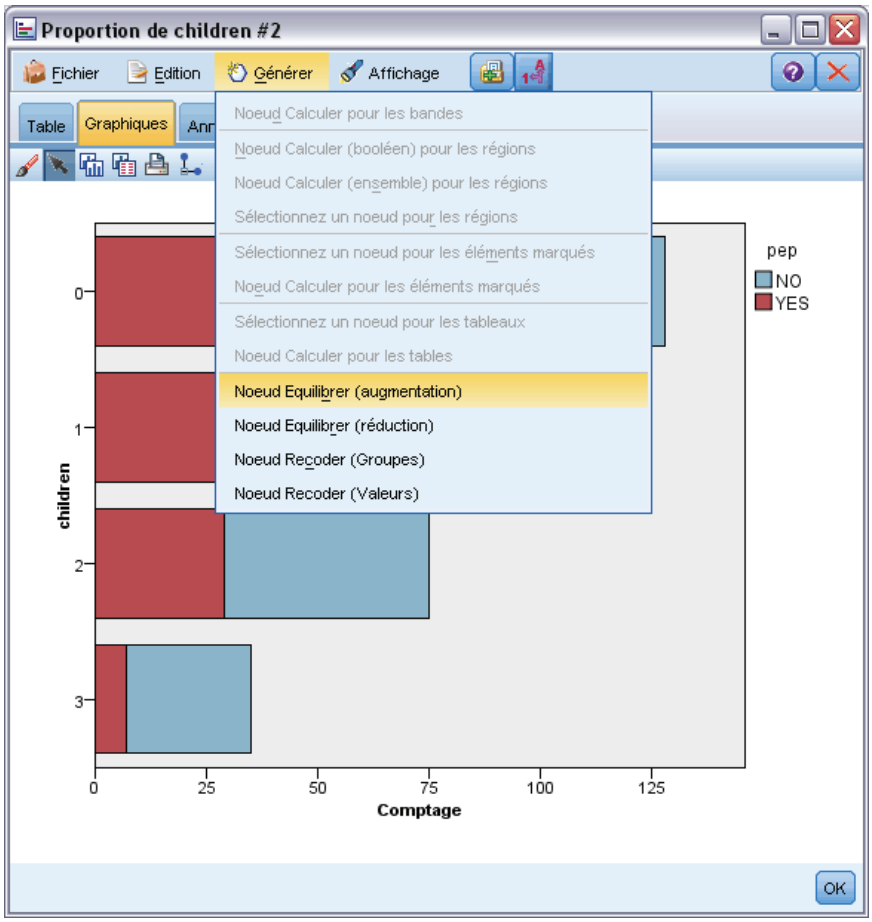
Exploration de graphiques

Tandis que le mode d'édition vous permet de modifier la mise en forme et l'aspect du graphique, le mode d'interaction vous permet d'explorer de manière analytique les données et les valeurs représentées par le graphique. Le principal objectif de l'exploration est d'analyser les données puis d'identifier les valeurs à l'aide de bandes, de zones et de marquages pour créer des noeuds Sélectionner, Calculer ou Equilibrer. Pour sélectionner ce mode, sélectionnez Affichage > Mode d'interaction dans les menus (ou cliquez sur l'icône de la barre d'outils).

Bien que certains graphiques puissent utiliser tous les outils d'exploration, d'autres n'en acceptent qu'un seul. Le mode d'interaction comprend :

- La définition et la modification de bandes, qui permettent de fractionner les valeurs le long d'un axe d'échelle *x*. Pour plus d'informations, reportez-vous à la section Utilisation de bandes sur p. 361.
- La définition et l'édition de zones, qui permettent d'identifier un groupe de valeurs dans une zone rectangulaire. Pour plus d'informations, reportez-vous à la section Présentation des zones sur p. 365.
- Le marquage ou l'annulation du marquage d'éléments pour choisir vous-même les valeurs à utiliser pour créer un noeud Sélectionner ou Calculer. Pour plus d'informations, reportez-vous à la section Présentation des éléments marqués sur p. 368.
- La création de noeuds à l'aide des valeurs identifiées par les bandes, les zones, les éléments marqués et les liens de relations à utiliser dans votre flux. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Figure 5-89
Graphique avec le menu Générer affichant



Utilisation de bandes

Dans tout graphique doté d'un champ d'échelle sur l'axe *x*, vous pouvez dessiner des lignes de bande verticales pour fractionner l'intervalle de valeurs sur l'axe *x*. Si un graphique contient plusieurs panneaux, une ligne de bande dessinée sur un panneau est également représentée sur les autres panneaux.

Certains graphiques n'acceptent pas les bandes. Voici certains des graphiques pouvant contenir des bandes : les histogrammes, les diagrammes à barres et proportion, les graphiques nuages (linéaires, de dispersion, horaires, etc.), résumés, et les graphiques d'évaluation. Dans les graphiques divisés en panneaux, les bandes apparaissent sur tous les panneaux. Et dans certains cas d'une matrice SPLOM, vous verrez s'afficher une ligne de bande horizontale du fait que l'axe sur lequel a été dessinée la bande de champ/variable a été inversé.

Figure 5-90
Graphique à trois bandes

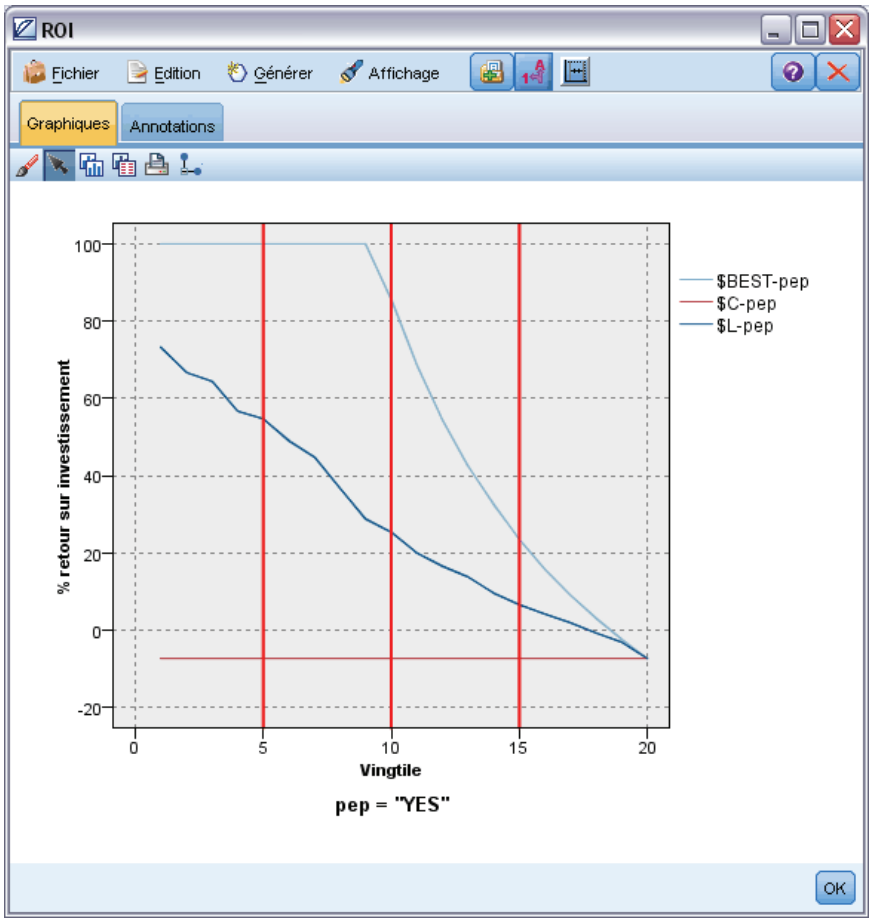
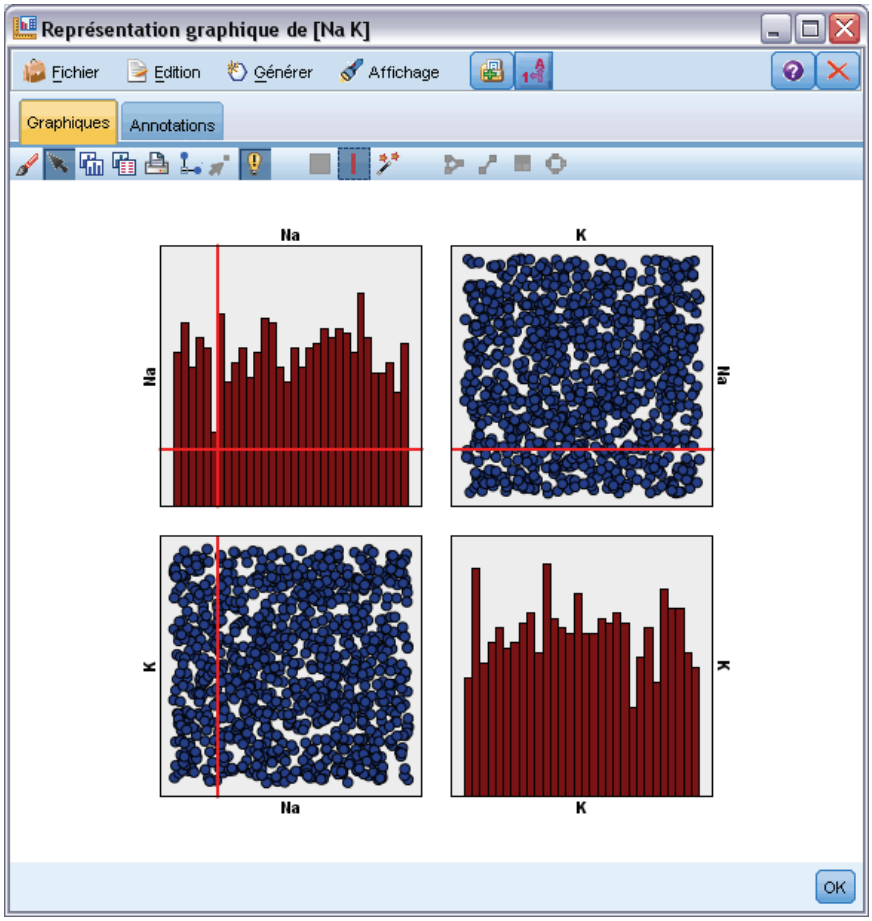


Figure 5-91
SPLOM avec bandes



Définition des bandes

Dans un graphique sans bande, l'ajout d'une ligne de bande fractionne le graphique en deux bandes. La valeur de la ligne de bande représente le point de départ, également appelé limite inférieure, de la deuxième bande lors de la lecture du graphique de gauche à droite. De la même manière, dans un graphique à deux bandes, l'ajout d'une ligne de bande fractionne l'une de ces bandes en deux, créant ainsi une troisième bande. Par défaut, les bandes sont nommées *bandN*, où *N* correspond au nombre de bandes, de gauche à droite, sur l'axe *x*.

Une fois que vous avez défini une bande, vous pouvez utiliser la fonction glisser-déposer pour la repositionner sur l'axe *x*. Vous pouvez accéder à d'autres raccourcis en cliquant avec le bouton droit à l'intérieur de la bande pour des tâches telles que renommer, supprimer ou créer des noeuds pour cette bande en particulier.

Pour définir des bandes :

▶ Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.

▶ Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une bande.

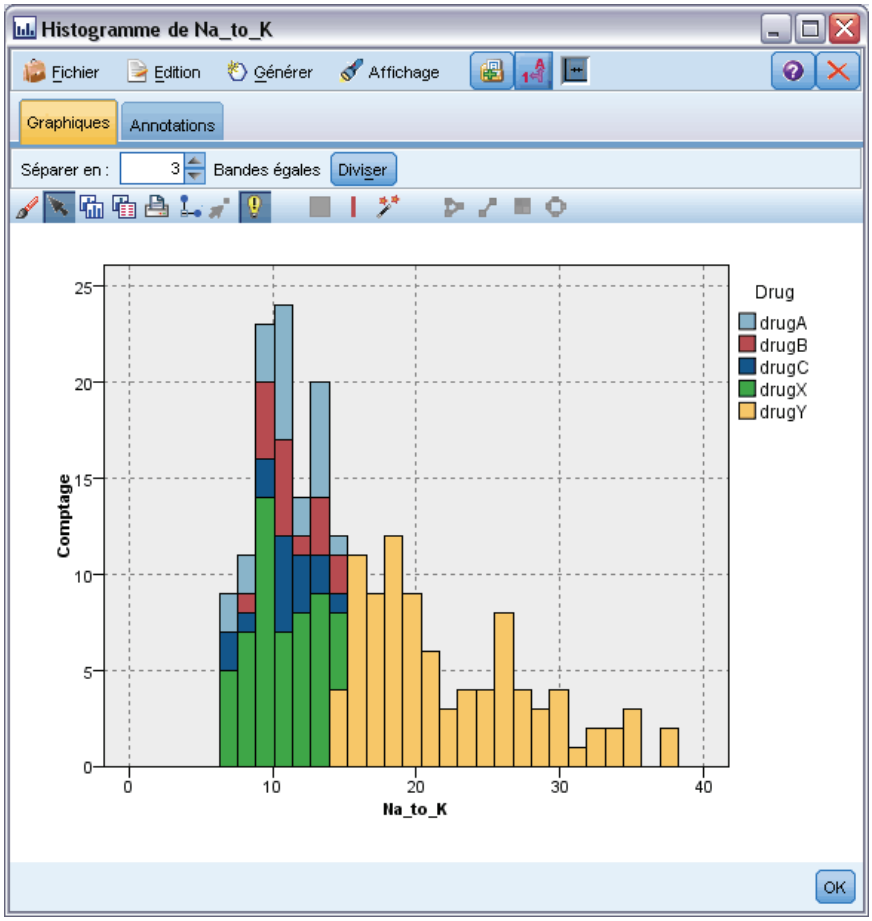
Figure 5-92
Bouton de barre d'outils Dessiner des bandes

▶ Dans un graphique qui accepte les bandes, cliquez sur le point de valeur de l'axe *x* au niveau duquel vous voulez définir une ligne de bande.

Remarque : Vous pouvez également cliquer sur l'icône de la barre d'outils Diviser le graphique en bandes et saisir le nombre de bandes égales souhaitées, puis cliquer sur Fractionner.

Figure 5-93
Icône de fractionnement utilisée pour développer la barre d'outils et afficher les options permettant de fractionner l'axe en bandes

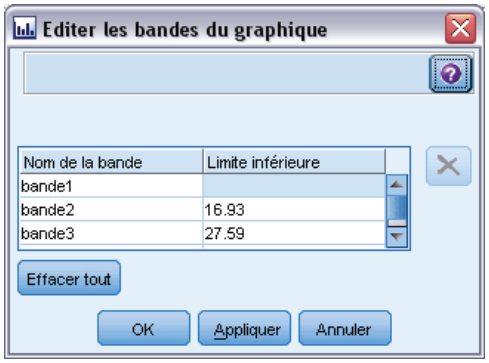
Figure 5-94
Barre d'outils Création de bandes égales avec bandes activées



Modification, changement de nom, et suppression de bandes

Vous pouvez modifier les propriétés des bandes existantes dans la boîte de dialogue Modifier les bandes du graphique ou via les menus contextuels dans le graphique lui-même.

Figure 5-95
Boîte de dialogue Modifier les bandes du graphique



Pour modifier des bandes :

- Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.
- Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une bande.
- Dans les menus, choisissez Edition > Bandes du graphique. La boîte de dialogue Modifier les bandes du graphique s'ouvre.
- Si vous avez plusieurs champs dans votre graphique (graphiques SPLOM par exemple), vous pouvez sélectionner le champ souhaité dans la liste déroulante.
- Ajoutez une nouvelle bande en saisissant un nom et une limite inférieure. Appuyez sur la touche Entrée pour commencer une nouvelle ligne.
- Modifiez la frontière d'une bande en ajustant la valeur de la Limite inférieure.
- Renommez une bande en saisissant un nouveau nom de bande.
- Supprimez une bande en sélectionnant la ligne dans le tableau, puis en cliquant sur le bouton de suppression.
- Cliquez sur OK pour appliquer vos modifications et fermer la boîte de dialogue.

  Remarque : Vous pouvez également supprimer et renommer les bandes directement dans le graphique en cliquant avec le bouton droit de la souris sur la ligne de la bande et en choisissant l'option souhaitée dans les menus contextuels.

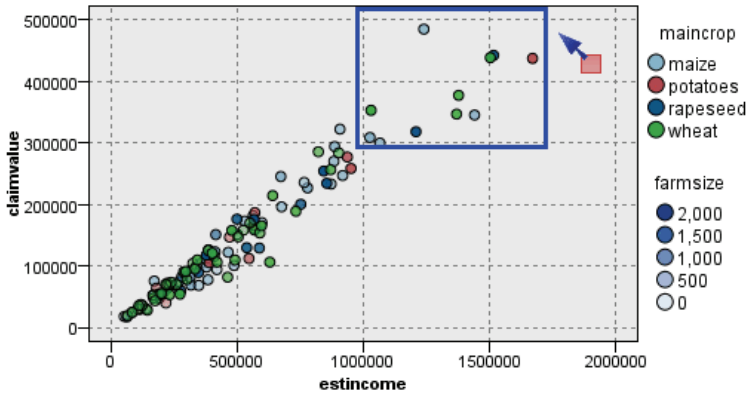
Présentation des zones

Dans tout graphique contenant deux axes d'échelle (ou d'intervalle), vous pouvez dessiner des zones pour regrouper des valeurs dans un rectangle, appelé zone. Une **zone** est une partie du graphique définie par ses valeurs *X* et *Y* minimales et maximales. Si un graphique contient

plusieurs panneaux, une zone dessinée sur un panneau est également représentée sur les autres panneaux.

Certains graphiques n'acceptent pas les zones. Voici certains des graphiques qui acceptent les zones : les graphiques nuages (linéaires, de dispersion, en bulles, horaires, etc.), les matrices SPLOM et les résumés. Ces zones sont dessinées dans un espace X,Y, et il est par conséquent impossible de les définir dans les graphiques en 1D, 3D ou animés. Dans les graphiques divisés en panneaux, les zones apparaissent sur tous les panneaux. Dans le cas d'une matrice de diagramme de dispersion (SPLOM), une zone correspondante apparaît dans les graphiques supérieurs correspondants, mais pas dans les graphiques en diagonale car ils n'affichent qu'un seul champ d'échelle.

Figure 5-96
Définition d'une zone dont les valeurs des demandes sont élevées



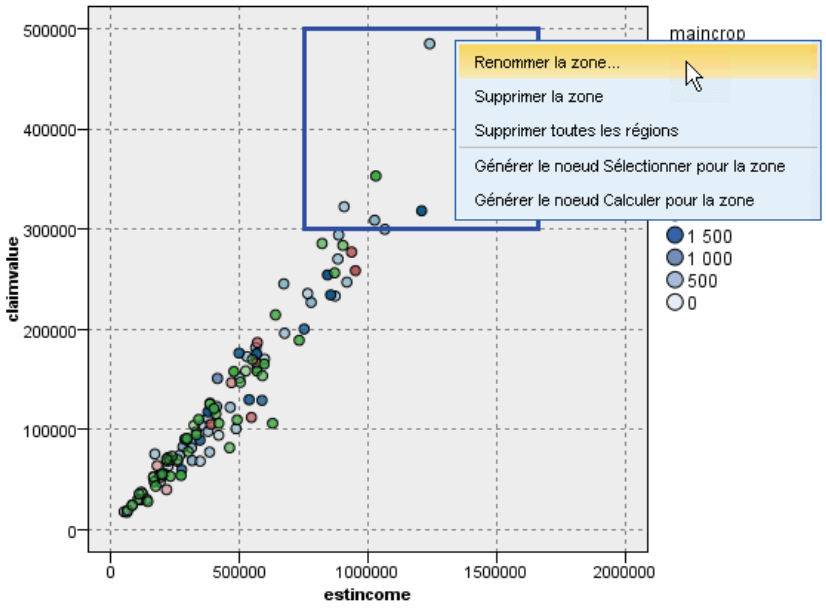
Définition des zones

Quelque soit l'endroit où vous définissez une zone, vous créez un groupement de valeurs. Par défaut, chaque nouvelle zone est appelée *Zone<N>*, où *N* correspond au nombre de zones déjà créées.

Une fois que vous avez défini une zone, vous pouvez cliquer avec le bouton droit de la souris sur la ligne de la zone pour accéder à certains raccourcis de base. Vous pouvez toutefois accéder à de nombreux autres raccourcis en cliquant avec le bouton droit à l'intérieur de la zone (et non sur la ligne) pour des tâches telles que le changement de nom, la suppression ou la création de noeuds Sélectionner et Calculer pour cette zone en particulier.

Vous pouvez sélectionner des sous-ensembles d'enregistrements en fonction de leur appartenance à une zone particulière ou à une zone parmi d'autres. Vous pouvez également incorporer à l'enregistrement des informations sur la zone en générant un noeud Calculer de sorte à ajouter un booléen aux enregistrements en fonction de leur appartenance à une zone. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Figure 5-97
Exploration de la zone dont les valeurs des demandes sont élevées



Pour définir des zones :

- Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.
- Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une zone.

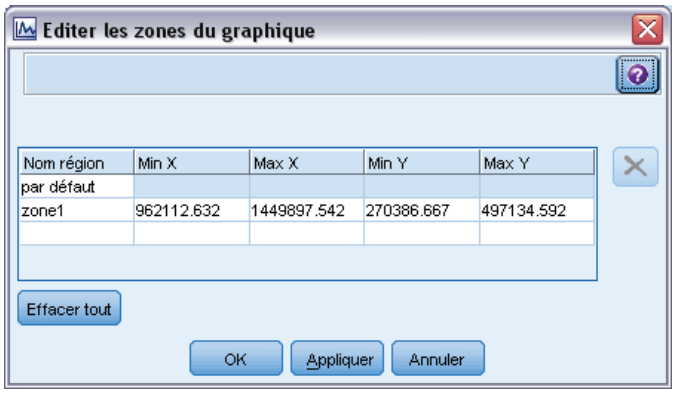
Figure 5-98
Bouton de barre d'outils Dessiner une zone

- Dans un graphique qui accepte les zones, cliquez et faites glisser votre souris pour dessiner la zone rectangulaire.

Modification, changement de nom, et suppression de zones

Vous pouvez modifier les propriétés des zones existantes dans la boîte de dialogue Modifier les zones du graphique ou via les menus contextuels dans le graphique lui-même.

Figure 5-99
Spécification des propriétés des zones définies



Pour modifier des zones :

- Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.
- Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une zone.
- Dans les menus, choisissez Edition > Zones du graphique. La boîte de dialogue Modifier les zones du graphique s'ouvre.
- Si vous avez plusieurs champs dans votre graphique (graphiques SPLOM par exemple), vous devez définir le champ de la zone dans les colonnes *Champ A* et*Champ B*.
- Pour ajouter une nouvelle zone sur une nouvelle ligne, saisissez un nom, sélectionnez les noms des champs (le cas échéant) et définissez les limites minimum et maximum pour chaque champ. Appuyez sur la touche Entrée pour commencer une nouvelle ligne.
- Modifiez les limites existantes de la zone en rectifiant les valeurs Minimum et Maximum de *A* et de *B*.
- Pour renommer une zone, modifiez son nom dans le tableau.
- Pour supprimer une zone, sélectionnez la ligne dans le tableau puis cliquez sur le bouton de suppression.
- Cliquez sur OK pour appliquer vos modifications et fermer la boîte de dialogue.

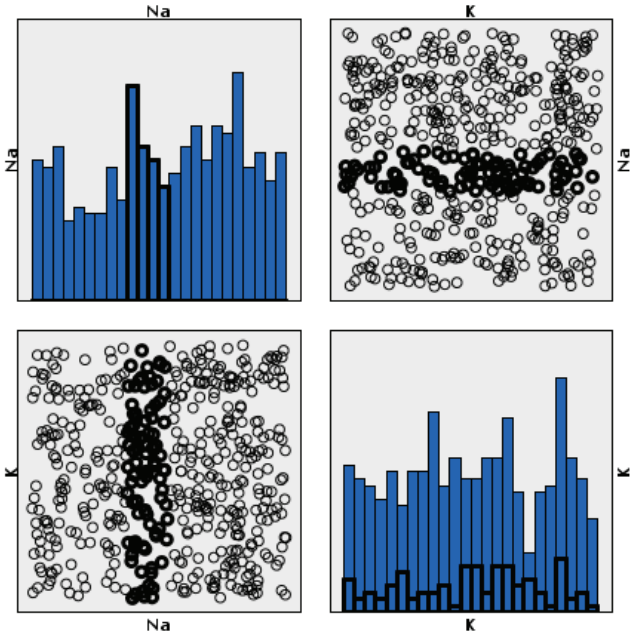
Remarque : Vous pouvez également supprimer et renommer les zones directement dans le graphique en cliquant avec le bouton droit de la souris sur la ligne de la zone et en choisissant l'option souhaitée dans les menus contextuels.

Présentation des éléments marqués

Vous pouvez marquer des éléments, comme des barres, des secteurs et des points de n'importe quel graphique. Il n'est pas possible de marquer les lignes, les zones et les aires dans les graphiques autres que Tracé horaire, Courbes et Évaluation, car dans ces cas, les lignes renvoient à des champs. Chaque fois que vous marquez un élément, vous mettez avant tout en évidence toutes les données représentées par cet élément. Dans tout graphique où la même observation

est représentée en plusieurs endroits (matrice SPLOM par exemple), le marquage est synonyme de brossage. Vous pouvez marquer des éléments figurant dans des graphiques, y compris dans des bandes et des zones. Chaque fois que vous marquez un élément, puis retournez au mode d'édition, le marquage reste visible.

Figure 5-100
Marquage d'éléments dans une matrice SPLOM



Vous pouvez marquer et annuler le marquage d'éléments en cliquant sur les éléments dans le graphique. Lorsque vous cliquez pour la première fois sur un élément pour le marquer, l'élément apparaît avec une couleur de bordure épaisse pour indiquer qu'il a été marqué. Si vous cliquez à nouveau sur l'élément, la bordure disparaît et l'élément n'est plus marqué. Pour marquer plusieurs éléments, vous pouvez maintenir enfoncée la touche Ctrl tout en cliquant sur les éléments, ou vous pouvez faire glisser la souris autour de chacun des éléments que vous souhaitez marquer à l'aide de la "baguette magique". Souvenez-vous que si vous cliquez sur une autre zone ou sur un autre élément tout en maintenant enfoncée la touche Ctrl, tous les éléments déjà marqués sont désélectionnés.

Vous pouvez générer des noeuds Sélectionner et Calculer à partir des éléments marqués dans votre graphique. Pour plus d'informations, reportez-vous à la section Génération de noeuds à partir de graphiques sur p. 370.

Pour marquer des éléments :

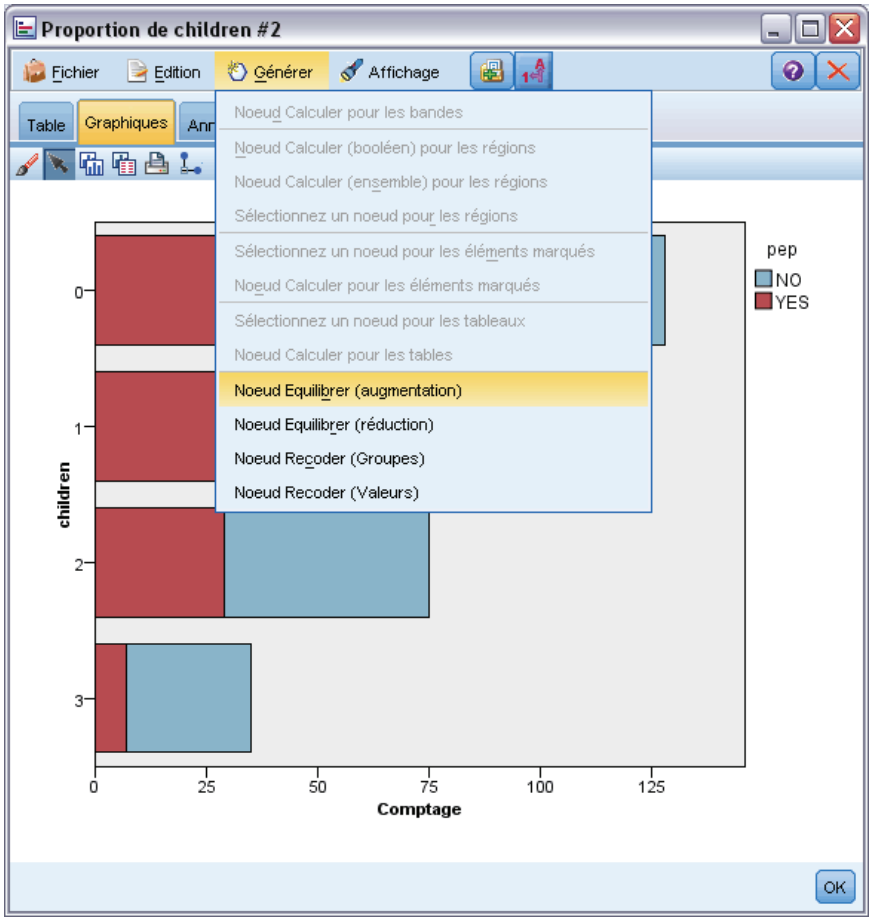
- Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.
- Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Marquer des éléments.
- Cliquez sur l'élément dont vous avez besoin ou cliquez et faites glisser votre souris pour dessiner une ligne autour de la région contenant plusieurs éléments.

Génération de noeuds à partir de graphiques

L'une des options les plus puissantes offertes par les graphiques IBM® SPSS® Modeler est la possibilité de générer des noeuds à partir d'un graphique ou d'une sélection dans le graphique. Vous pouvez ainsi, dans un graphique de tracé horaire, générer des noeuds Calculer et Sélectionner en fonction d'une sélection ou d'une zone de données, ce qui entraîne la définition de sous-ensembles de données. Par exemple, vous pouvez utiliser la puissance de cette fonction pour identifier et exclure les valeurs éloignées.

Chaque fois que vous dessinez une bande, vous pouvez également générer un noeud Calculer. Dans les graphiques à deux axes d'échelle, vous pouvez générer des noeuds Calculer ou Sélectionner à partir des zones dessinées dans votre graphique. Dans les graphiques contenant des éléments marqués, vous pouvez générer des noeuds Calculer, des noeuds Sélectionner, et dans certains cas des noeuds Filtrer à partir de ces éléments. La génération de noeuds Equilibrer est activée pour tout graphique représentant une distribution de nombres.

Figure 5-101
Graphique avec le menu Générer affichant



Chaque fois que vous générez un noeud, il est placé directement sur l'espace de travail de flux afin que vous puissiez le connecter à un flux existant. Les noeuds suivants peuvent être générés à partir de graphiques : Sélectionner, Calculer, Equilibrer, Filtrer et Recoder.

Noeuds Sélectionner

Les noeuds Sélectionner peuvent être générés pour tester l'inclusion d'enregistrements dans une zone et l'exclusion de tous les enregistrements non compris dans la zone, ou l'inverse dans le cas d'un traitement en aval.

- **Pour les bandes**. Vous pouvez générer un noeud Sélectionner qui inclut ou exclut les enregistrements compris dans cette bande. Le noeud Sélectionner pour les bandes est disponible uniquement via les menus contextuels car vous devez sélectionner la bande à utiliser dans le noeud Sélectionner.
- **Pour les zones**. Vous pouvez générer un noeud Sélectionner qui inclut ou exclut les enregistrements compris dans une zone.
- **Pour les éléments marqués**. Vous pouvez générer des noeuds Sélectionner pour capturer les enregistrements correspondant aux éléments marqués ou aux liens du graphique Relations.

Noeuds Calculer

Les noeuds Calculer peuvent être générés à partir de zones, de bandes et d'éléments marqués. Tous les graphiques peuvent produire des noeuds Calculer. Dans le cas des graphiques d'évaluation, une boîte de dialogue de sélection du modèle s'affiche. Dans le cas des graphiques Relations, le noeud Calculer ("Et") et le noeud Calculer ("Ou") sont possibles.

- **Pour les bandes**. Vous pouvez générer un noeud Calculer qui produit une catégorie pour chaque intervalle marqué sur l'axe, à l'aide des noms de bande répertoriés en tant que noms de catégorie dans la boîte de dialogue Modifier les bandes.
- **Pour les zones**. Vous pouvez générer un noeud Calculer (Calculer en tant que booléen) qui crée un champ booléen appelé *Dans_zone*, les booléens étant définis sur *T* pour les enregistrements inclus dans une zone et sur *F* pour les autres enregistrements. Vous pouvez également générer un noeud Calculer (Calculer en tant qu'ensemble) qui produit un ensemble avec une valeur pour chaque zone et un nouveau champ appelé *zone* pour chaque enregistrement, qui prend comme valeur le nom de la zone dans laquelle est compris l'enregistrement. Les enregistrements qui ne sont compris dans aucune zone reçoivent le nom de la zone par défaut. Les noms des valeur deviennent les noms des zone répertoriés dans la boîte de dialogue Modifier les zones.
- **Pour les éléments marqués**. Vous pouvez générer un noeud Calculer qui calcule un booléen dont la valeur est *True (vrai)* pour tous les éléments marqués et *False (faux)* pour tous les autres enregistrements.

Noeuds Equilibrer

Les noeuds Equilibrer peuvent être générés pour corriger des déséquilibres dans les données. Il est par exemple possible de réduire la fréquence des valeurs courantes (utilisez l'option de menu Noeud Equilibrer (réduire)) ou d'augmenter l'occurence des valeurs sous-représentées (utilisez l'option de menu Noeud Equilibrer (augmenter)). La génération de noeuds Equilibrer est activée pour tout graphique représentant une distribution de nombres, tel que Histogramme, Points, Résumé, Barre de nombres, Secteurs de nombres, et Courbes .

Noeuds Filtrer

Les noeuds Filtrer peuvent être générés pour renommer ou filtrer les champs en fonction de lignes ou de noeuds marqués dans le graphique. Dans le cas de graphiques d'évaluation, la ligne la plus appropriée ne génère pas de noeud Filtrer.

Noeuds Recoder

Les noeuds Recoder peuvent être générés pour recoder des valeurs. Cette option est utilisée pour les graphiques de distribution. Vous pouvez générer un noeud Recoder pour des **groupes**, afin de recoder des valeurs spécifiques d'un champ affiché en fonction de leur appartenance à un groupe (sélectionnez les groupes à l'aide de la combinaison Ctrl+clic sur l'onglet Tables). Vous pouvez également générer un noeud Recoder pour des **valeurs**, afin de recoder les données d'un ensemble de valeurs existant. Il peut s'agir par exemple de recoder les données d'un ensemble de valeurs standard afin de fusionner les données financières issues de différentes sociétés en vue de leur analyse.

Remarque : Si ces valeurs sont prédéfinies, vous pouvez les lire dans SPSS Modeler en tant que fichier plat et utiliser une distribution pour toutes les afficher. Ensuite, créez directement à partir du graphique un noeud (de valeurs) Recoder pour le champ. Cette opération place l'ensemble des valeurs cible dans la colonne (liste déroulante) *Nouvelles valeurs* du noeud Recoder.

Génération de noeuds à partir de graphiques

Vous pouvez utiliser le menu Générer de la fenêtre de sortie du graphique pour générer des noeuds. Le noeud généré est placé dans l'espace de travail de flux. Pour utiliser ce noeud, connectez-le à un flux existant.

Pour générer un noeud à partir d'un graphique :

- Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez Affichage > Mode d'interaction.
- Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Zone.
- Définissez les bandes, les zones ou tous les éléments marqués nécessaires pour générer votre noeud.
- Dans le menu Générer, choisissez le type de noeud que vous souhaitez produire. Seuls les noeuds possibles sont activés.

  Remarque : Vous pouvez également générer des noeuds directement à partir du graphique en cliquant avec le bouton droit de la souris et en choisissant l'option souhaitée dans les menus contextuels.

Modification des visualisations

Alors que le mode Exploration permet d'explorer les données et les valeurs représentées par la visualisation de manière analytique, le mode d'édition permet de modifier la présentation et l'apparence de la visualisation. Vous pouvez par exemple modifier les polices et les couleurs pour

respecter le guide de style de votre organisation. Pour sélectionner ce mode, sélectionnez Affichage > Mode d'édition dans les menus (ou cliquez sur l'icône de la barre d'outils).

En mode d'édition, il existe plusieurs barres d'outils permettant de modifier l'aspect de la présentation des visualisations. Si vous n'utilisez pas certaines barres d'outils, vous pouvez les masquer afin d'augmenter l'espace disponible dans la boîte de dialogue dans laquelle le graphique est affiché. Pour sélectionner ou désélectionner des barres d'outils, cliquez sur le nom de la barre d'outils souhaitée dans le menu Affichage.

Remarque : Pour ajouter des détails supplémentaires à vos visualisations, vous pouvez appliquer un titre, des notes de bas de page et des étiquettes d'axes. Pour plus d'informations, reportez-vous à la section Ajout de titres et de notes de bas de page sur p. 390.

Plusieurs options sont disponibles pour modifier une visualisation en **mode d'édition**. Vous pouvez :

- Editer le texte et le formater
- Modifier la couleur de remplissage, la transparence et le motif des cadres et des éléments graphiques.
- Changer la couleur et les tirets des bordures et des lignes
- Changer la forme et le rapport d'aspect des points, et faire pivoter les points
- Changer la taille des éléments graphiques (par exemple, les barres et les points)
- Ajuster l'espace entourant les éléments à l'aide des marges et via l'extension
- Spécifier le formatage des chiffres.
- Changer les paramètres d'axe et d'échelle
- Trier, exclure et réduire des catégories sur un axe catégoriel.
- Définir l'orientation des panels.
- Appliquer des transformations à un système de coordonnées.
- Modifier les statistiques, les types d'éléments graphiques et les modificateurs de collision.
- Changer la position de la légende
- Appliquer des feuilles de style de visualisation.

Les rubriques suivantes décrivent ces diverses tâches. Il est également recommandé de prendre connaissance des règles générales liées à l'édition de graphiques.

Basculer en mode d'édition

▶ A partir des menus, sélectionnez :
Affichage > Mode Edition

Règles générales de modification des visualisations

Mode Edition

Toutes les modifications sont effectuées en mode d'édition. Pour activer le mode d'édition, dans les menus, choisissez :

Affichage > Mode Edition

Sélection

Les options d'édition disponibles dépendent de la sélection effectuée. Différentes options de barre d'outils et de palette de propriétés sont activées selon les éléments sélectionnés. Seules les options activées s'appliquent à la sélection courante. Par exemple, si un axe est sélectionné, les onglets Echelle, Graduations principales et Graduations secondaires sont disponibles dans la palette des propriétés.

Voici quelques conseils pour sélectionner des éléments dans la visualisation :

- Cliquez sur un élément pour le sélectionner.
- Sélectionnez un élément graphique (par exemple, des points dans un diagramme de dispersion ou des barres dans un diagramme à barres), en cliquant une seule fois. Après la sélection initiale, cliquez de nouveau pour réduire la sélection à des groupes d'éléments graphiques ou à un seul élément graphique.
- Appuyez sur la touche Echap pour désélectionner tous les éléments.

Palettes

Lorqu'un élément est sélectionné dans la visualisation, les différentes palettes sont mises à jour pour refléter cette sélection. Les palettes comprennent les commandes nécessaires pour apporter des modifications à la sélection. Les palettes peuvent être des barres d'outils ou un panel avec plusieurs commandes et onglets. Les palettes peuvent être masquées. Par conséquent, vérifiez que la palette appropriée est affichée pour effectuer vos modifications. Vérifiez dans le menu Affichage les palettes affichées actuellement.

Vous pouvez repositionner les palettes en cliquant sur et en faisant glisser l'espace vide dans une palette de barres d'outils ou sur le côté gauche d'autres palettes. L'affichage à l'écran vous indique où vous pouvez ancrer la palette. Pour les palettes ne contenant pas de barres d'outils, vous pouvez également cliquer sur le bouton Fermer pour masquer la palette et le bouton Détacher pour afficher la palette dans une fenêtre séparée. Cliquez sur le bouton Aide pour afficher l'aide d'une palette spécifique.

Paramètres automatiques

Certains paramètres fournissent une option -Automatique-. Cette option indique que des valeurs automatiques sont appliquées. Les paramètres automatiques utilisés dépendent de la visualisation et des valeurs de données spécifiques. Vous pouvez entrer une valeur pour remplacer le paramètre automatique. Pour restaurer le paramètre automatique, supprimez la valeur actuelle et appuyez sur Entrée. Le paramètre affiche de nouveau -Automatique-.

Suppression/Masquage d'éléments

Vous pouvez supprimer/masquer différents éléments dans la visualisation. Par exemple, vous pouvez masquer la légende ou l'étiquette d'axe. Pour supprimer un élément, sélectionnez-le et appuyez sur Supprimer. Si la suppression de l'élément n'est pas autorisée, rien ne se produit. Si vous supprimez un élément par erreur, appuyez sur Ctrl+Z pour annuler la suppression.

Etat

Certaines barres d'outils reflètent l'état de la sélection actuelle, contrairement à d'autres. La palette des propriétés reflète toujours cet état. Si une barre d'outils ne reflète *pas* l'état, ceci est mentionné dans la rubrique décrivant cette barre d'outils.

Edition et formatage de texte

Vous pouvez éditer le texte en place et changer le formatage d'un bloc de texte entier. Vous ne pouvez pas éditer un texte directement lié à des valeurs de données. Par exemple, vous ne pouvez pas éditer une étiquette de graduation car le contenu de l'étiquette provient des données sous-jacentes. Mais vous pouvez formater n'importe quel texte de la visualisation.

Comment éditer le texte en place

- Double-cliquez sur le bloc de texte. Cette action sélectionne tout le texte. Toutes les barres d'outils sont alors désactivées, car vous ne pouvez modifier aucune autre partie de la visualisation pendant la modification du texte.
- Tapez le nouveau texte pour remplacer le texte existant. Vous pouvez également cliquer de nouveau sur le texte pour afficher un curseur. Placez le curseur à la position souhaitée et entrez le texte supplémentaire.

Comment formater du texte

- Sélectionnez le cadre qui contient le texte. Ne double-cliquez pas sur le texte.
- Formatez le texte à l'aide de la barre d'outils des polices. Si cette barre d'outils n'est pas activée, vérifiez que seul le *cadre* contenant le texte est sélectionné. Si le texte lui-même est sélectionné, la barre d'outils est désactivée.

Figure 5-102
Barre d'outils des polices

Vous pouvez changer la police :

- Couleur
- Famille (par exemple, Arial ou Verdana)
- Taille (l'unité utilisée est le point, sauf si vous indiquez une unité différente telle que le pica, pc)

- Pondération
- Alignement par rapport au cadre du texte

Le formatage s'applique à tout le texte figurant dans le cadre. Vous ne pouvez pas changer le formatage de certaines lettres ou de certains mots dans un bloc de texte spécifique.

Modification des couleurs, des motifs, des pointillés et de la transparence

Plusieurs éléments différents d'une visualisation ont un remplissage et des bordures. L'exemple le plus évident est celui d'une barre dans un diagramme à barres. La couleur des barres est la couleur de remplissage. Les barres peuvent également être entourées d'une bordure unie noire.

Il existe d'autres éléments moins apparents dans la visualisation qui ont des couleurs de remplissage. Si la couleur de remplissage est transparente, le remplissage n'est pas nécessairement visible. Par exemple, considérons le texte d'une étiquette d'axe. Ce texte semble « flotter » mais figure en fait dans un cadre comportant une couleur de remplissage transparente. Le cadre est visible lorsque vous sélectionnez l'étiquette d'axe.

Tout cadre dans la visualisation peut avoir un style de remplissage et de bordure, y compris le cadre autour de la visualisation entière. De plus, tout remplissage possède un niveau d'opacité/transparence qui lui est associé et qui peut être ajusté.

Modifier les couleurs, les motifs, les pointillés et la transparence

▶ Sélectionnez l'élément à formater. Par exemple, sélectionnez les barres d'un diagramme à barres ou un cadre contenant du texte. Si la visualisation est scindée par une variable ou un champ catégoriel, vous pouvez également sélectionner le groupe correspondant à une catégorie individuelle. Vous pouvez ainsi changer l'apparence par défaut attribuée à ce groupe. Par exemple, vous pouvez changer la couleur de l'un des groupes de superposition d'un diagramme à barres superposées.

▶ Pour changer la couleur de remplissage, la couleur de bordure ou le motif de remplissage, utilisez la barre d'outils des couleurs.

Figure 5-103
Barre d'outils des couleurs

Remarque : Cette barre d'outils ne reflète pas l'état de la sélection courante.

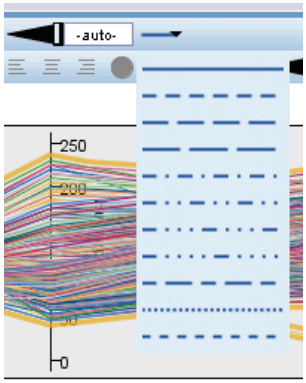
Pour changer une couleur ou un remplissage, vous pouvez cliquer sur le bouton pour sélectionner l'option affichée ou cliquer sur la flèche du menu déroulant pour choisir une autre option. Pour les couleurs, il existe une couleur paraissant blanche et traversée d'une ligne diagonale rouge. Il s'agit de la couleur transparente. Cette couleur peut être utilisée par exemple pour masquer les bordures des barres dans un histogramme.

- Le premier bouton contrôle la couleur de remplissage.
- Le deuxième bouton contrôle la couleur de bordure.

- Le troisième bouton contrôle le motif de remplissage. Le motif de remplissage utilise la couleur de bordure. Par conséquent, le motif de remplissage n'est visible que s'il existe une couleur de bordure visible.
- La quatrième commande est un curseur et une zone de texte qui contrôlent l'opacité de la couleur de remplissage et du motif. Un pourcentage peu élevé signifie moins d'opacité et plus de transparence. 100% est totalement opaque (aucune transparence).

▶ Pour changer les tirets d'une bordure ou d'une ligne, utilisez la barre d'outils de ligne.

Figure 5-104
Barre d'outils de ligne



Remarque : Cette barre d'outils ne reflète pas l'état de la sélection courante.

Comme dans le cas de l'autre barre d'outils, vous pouvez cliquer sur le bouton pour sélectionner l'option affichée ou sur la flèche de la liste déroulante pour sélectionner une autre option.

Changement de la forme et du rapport d'aspect des points et rotation des points

Vous pouvez faire pivoter des points, attribuer une forme prédéfinie différente ou changer le rapport d'aspect (le rapport entre la largeur et la hauteur).

Comment modifier des points

▶ Sélectionnez les points. Vous ne pouvez pas changer la forme et le rapport d'aspect de points individuels ni faire pivoter ces points.

▶ Utilisez la barre d'outils des symboles pour modifier les points.

Figure 5-105
Barre d'outils des symboles

- Le premier bouton permet de changer la forme des points. Cliquez sur la flèche de la liste déroulante et sélectionnez une forme prédéfinie.

- Le deuxième bouton permet de faire pivoter les points sur une position de compas spécifique. Cliquez sur la flèche de la liste déroulante, puis faites glisser l'aiguille vers la position désirée.
- Le troisième bouton permet de changer le rapport d'aspect. Cliquez sur la flèche de la liste déroulante, puis cliquez sur le rectangle qui apparaît et faites-le glisser. La forme du rectangle représente le rapport d'aspect.

Changement de la taille des éléments graphiques

Vous pouvez modifier la taille des éléments graphiques dans la visualisation. Ces éléments sont notamment les barres, les lignes et les points. Si la taille de l'élément graphique est déterminée par une variable ou un champ, la taille spécifiée est la taille *minimale*.

Comment changer la taille des éléments graphiques

▶ Sélectionnez les éléments graphiques à redimensionner.

▶ Utilisez le curseur ou entrez une taille spécifique pour l'option disponible dans la barre d'outils des symboles. L'unité utilisée est le pixel, sauf si vous indiquez une unité différente (vous trouverez ci-dessous une liste complète des abréviations d'unité). Vous pouvez également spécifier un pourcentage (par exemple, 30 %), ce qui signifie qu'un élément graphique utilise le pourcentage d'espace disponible spécifié. L'espace disponible dépend du type de l'élément graphique et de la visualisation.

Table 5-3
Abréviations d'unités valides

Abréviation	Unité
cm	centimètre
entrée	pouce
mm	millimètre
pc	pica
pt	point
px	pixel

Figure 5-106
Contrôle de la taille dans la barre d'outils des symboles

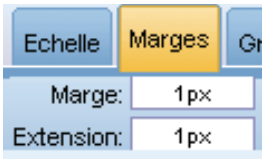
Spécification des marges et de l'extension

S'il y a trop ou pas assez d'espace autour ou à l'intérieur d'un cadre dans la visualisation, vous pouvez modifier ses paramètres de marge et d'encadrement. La **marge** est la quantité d'espace séparant le cadre des autres éléments situés autour de ce cadre. L'**extension** est la quantité d'espace situé entre la bordure et le *contenu* du cadre.

Comment spécifier les marges et l'extension

- Sélectionnez le cadre pour lequel vous souhaitez spécifier des marges et une extension. Il peut s'agir d'un cadre de texte, d'un cadre entourant une légende ou même d'un cadre de données affichant des éléments graphiques (par exemple, des barres et des points).
- Utilisez l'onglet Marges de la palette des propriétés pour spécifier les paramètres. Toutes les tailles sont exprimées en pixels, sauf si vous indiquez une unité différente (par exemple, le centimètre cm ou le pouce in).

Figure 5-107
Onglet Marges



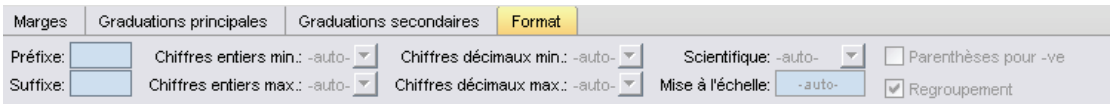
Formatage des nombres

Vous pouvez spécifier le format des nombres figurant dans les étiquettes de graduation sur un axe continu ou dans les étiquettes de valeurs de données affichant un nombre. Par exemple, vous souhaitez spécifier que les nombres affichés sur les étiquettes de graduation soient en milliers.

Spécifier les formats des nombres

- Sélectionnez les étiquettes de graduation de l'axe continu ou les étiquettes de valeur de données si elles comportent des nombres.
- Cliquez sur l'onglet Format dans la palette des propriétés.

Figure 5-108
Onglet Format



- Sélectionnez les options de formatage des nombres désirées :

Préfixe. Un caractère à afficher devant le nombre. Par exemple, saisissez le symbole ($) si les nombres correspondent à des salaires en dollars U.S.

Suffixe. Un caractère à afficher après le nombre. Par exemple, saisissez le symbole du pourcentage (%) si les nombres sont des pourcentages.

Chiffres entiers min.. Nombre de chiffres minimum à afficher dans la partie entière d'une représentation décimale. Si la valeur réelle ne contient pas le nombre de chiffres minimum, la partie entière de cette valeur sera complétée par des zéros.

Chiffres entiers max.. Nombre de chiffres maximum à afficher dans la partie entière d'une représentation décimale. Si la valeur réelle dépasse le nombre de chiffres minimum, la partie entière de cette valeur sera remplacée par des astérisques.

Décimales min.. Nombre de chiffres minimum à afficher dans la partie décimale d'une représentation décimale ou scientifique. Si la valeur réelle ne contient pas le nombre de chiffres minimum, la partie décimale de cette valeur sera complétée par des zéros.

Décimales max.. Nombre de chiffres maximum à afficher dans la partie décimale d'une représentation décimale ou scientifique. Si la valeur réelle dépasse le nombre de chiffres minimum, la décimale est arrrondie au nombre de chiffres approprié.

Scientifique. Affichage ou non des chiffres en notation scientifique. Cette notation est utile pour des nombres très grands ou très petits. -auto- permet à l'application de choisir si la notation scientifique est appropriée.

Echelle. Un facteur d'échelle, qui est un nombre par lequel la valeur d'origine est divisée. Utilisez un facteur d'échelle si vous souhaitez que l'étiquette ne s'étende pas trop pour s'ajuster aux nombres élevés. Si vous modifiez le format des nombres des étiquettes de graduation, veillez à modifier le titre de l'axe pour indiquer comment les nombres doivent être interprétés. Par exemple, si votre axe d'échelle affiche des salaires et que les étiquettes sont 30 000, 50 000 et 70 000, vous pouvez saisir un facteur d'échelle de 1 000 pour afficher 30, 50 et 70. Vous devez ensuite modifier l'axe d'échelle pour inclure le texte en milliers.

Parenthèses pour -ve. Si des parenthèses doivent entourer les valeurs négatives.

Regroupement. Si un caractère doit être placé entre les groupes de chiffres. Les paramètres régionaux en cours de votre ordinateur déterminent le caractère utilisé pour le regroupement de chiffres.

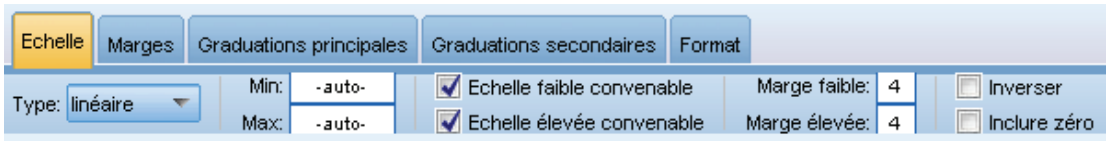
Changement des paramètres d'axe et d'échelle

Plusieurs options permettent de changer les axes et les échelles.

Comment changer les paramètres d'axe et d'échelle

- Sélectionnez une partie de l'axe (par exemple, l'étiquette d'axe ou les étiquettes de graduation).
- Utilisez les onglets Echelle, Graduations principales et Graduations secondaires de la palette des propriétés pour changer les paramètres d'axe et d'échelle.

Figure 5-109
Palette Propriétés



Onglet Echelle

Remarque : L'onglet Echelle n'apparaît pas pour les graphiques dans lesquels les données sont pré-agrégées (par exemple, les histogrammes).

Type. Indique si l'échelle est linéaire ou transformée. Les transformations d'échelle aident à comprendre les données ou à émettre les hypothèses nécessaires à la déduction statistique. Dans les diagrammes de dispersion, vous pouvez utiliser une échelle transformée si la relation entre les variables indépendantes et dépendantes ou les champs est non linéaire. Les transformations d'échelle permettent également de rendre un histogramme asymétrique plus symétrique de sorte qu'il ressemble à une distribution normale. Vous ne transformez que l'échelle à laquelle les données sont affichées et non pas les données elles-mêmes.

- **linéaire.** Indique une échelle non transformée linéaire.
- **log.** Indique une échelle transformée logarithmique décimale. Pour prendre en compte les valeurs nulles (zéro) et négatives, cette transformation utilise une version modifiée de la fonction log. Cette fonction log « sûre » est définie selon la formule sign(x) * log(1 + abs(x)). De manière à ce que safeLog(-99) soit égal à :

  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
- **exposant.** Indique une échelle transformée de puissance à l'aide de l'exposant 0,5. Pour prendre en compte les valeurs négatives, cette transformation utilise une version modifiée de la fonction de puissance. Cette fonction de puissance « sûre » est définie selon la formule sign(x) * pow(abs(x), 0.5). De manière à ce que safePower(-100) soit égal à :

  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10

Minimum/Maximum/Echelle faible convenable/Echelle élevée convenable. Indique l'intervalle de l'échelle. Le fait de sélectionner Echelle faible convenable et Echelle élevée convenable permet à l'application de sélectionner une échelle appropriée en fonction des données. Le minimum et le maximum sont convenables car ils constituent généralement des valeurs entières supérieures ou inférieures aux valeurs de données maximum et minimum. Par exemple, si l'intervalle de données va de 4 à 92, la valeur Echelle faible convenable ou Echelle élevée convenable pour l'échelle peut être 0 ou 100 respectivement plutôt que les valeurs réelles de données maximum et minimum. Veillez à ne pas définir unintervalle trop restreint et qui risquerait de masquer des éléments importants. En outre, vous ne pouvez pas définir de valeurs minimum et maximum explicites si l'option Inclure zéro est sélectionnée.

Marge basse/Marge haute. Crée des marges à l'extrémité inférieure et/ou supérieure de l'axe. La marge est perpendiculaire à l'axe sélectionné. L'unité utilisée est le pixel, sauf si vous indiquez une unité différente (par exemple, le centimètre, cm ou le pouce, in). Par exemple, si vous définissez la marge Marge élevée sur 5 pour l'axe vertical, une marge horizontale de 5 pixels s'étend le long de la partie supérieure du cadre de données.

Inverser. Indique si l'échelle est inversée.

Inclure le zéro. Indique que l'échelle doit inclure 0. Cette option est généralement utilisée pour les diagrammes en barres afin de s'assurer que les barres commencent à 0 plutôt qu'à une valeur proche de la hauteur de la barre la plus petite. Si cette option est sélectionnée, les options Minimum et Maximum sont désactivées car vous ne pouvez pas définir de valeur minimum et maximum personnalisée pour l'intervalle de l'échelle.

Onglets Graduations principales/Graduations secondaires

Les **graduations** ou **marques de graduation** sont les lignes qui apparaissent sur un axe. Celles-ci indiquent des valeurs à des intervalles ou catégories spécifiques. Les **graduations principales** sont les marques de graduation avec étiquettes. Celles-ci sont également plus longues que les autres marques de graduation. Les **graduations secondaires** sont les marques de graduation qui apparaissent entre les marques de graduation principales. Certaines options sont spécifiques au type de graduation mais la plupart des options sont disponibles pour les graduations principales et secondaires.

Afficher les graduations. Indique si les graduations principales ou secondaires apparaissent sur un diagramme.

Afficher les quadrillages. Indique si les quadrillages apparaissent au niveau des graduations principales ou secondaires. Le **quadrillage** est un ensemble de lignes qui quadrillent un diagramme entier d'un axe à l'autre.

Position : Indique la position des marques de graduation par rapport à l'axe.

Longueur. Spécifie la longueur des marques de graduation. L'unité utilisée est le pixel, sauf si vous indiquez une unité différente (par exemple, le centimètre, cm ou le pouce, in).

Base. *S'applique uniquement aux graduations principales.* Indique la valeur à laquelle la première graduation principale apparaît.

Delta. *S'applique uniquement aux graduations principales.* Indique la différence entre les graduations principales. En d'autres termes, les graduations principales apparaîtront à chaque *n*e valeur, *n* étant la valeur delta.

Divisions. *S'applique uniquement aux graduations secondaires.* Indique le nombre de divisions de graduation secondaire entre les graduations principales. Le nombre de graduations secondaires est inférieur de une unité au nombre de divisions. Par exemple, supposons qu'il existe des graduations principales à 0 et 100. Si vous entrez 2 comme nombre de divisions de graduation secondaire, il y aura *une* graduation secondaire à 50, divisant l'intervalle 0–100 et créant *deux* divisions.

Modification des modalités

Il existe plusieurs méthodes de modification des catégories sur un axe vertical :

- Changer l'ordre de tri pour l'affichage des catégories.
- Exclure des catégories spécifiques.
- Ajoutez une modalité qui n'apparaît pas dans l'ensemble de données.
- Fusionnez/combinez de petites modalités en une seule modalité.

Comment changer l'ordre de tri des catégories

▶ Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans IBM® SPSS® Modeler du menu Affichage, choisissez Catégories.

▶ Dans la palette Catégories, sélectionnez une option de tri dans la liste déroulante :

Personnalisée. Trier les catégories en fonction de l'ordre dans lequel elles apparaissent dans la palette. Utilisez les flèches pour placer les catégories en haut ou en bas de la liste, ou les déplacer vers le haut ou vers le bas.

Données. Trier les catégories en fonction de l'ordre dans lequel elles apparaissent dans l'ensemble de données.

Nom. Trier les catégories dans l'ordre alphabétique, en utilisant les noms affichés dans la palette. Il peut s'agir de la valeur ou de l'étiquette, selon si le bouton de la barre d'outils permettant d'afficher les valeurs ou les étiquettes est sélectionné ou non.

Valeur. Triez les modalités en fonction de la valeur de données sous-jacente en utilisant les valeurs entre parenthèses de la palette. Seules les sources de données avec des métadonnées (fichier de données IBM® SPSS® Statistics par exemple) prennent en charge cette option.

Statistique. Trier les catégories en fonction de la statistique calculée pour chaque catégorie. Il peut s'agir par exemple de nombres, de pourcentages et de moyennes. Cette option est uniquement disponible si une statistique est utilisée dans le graphique.

Ajout d'une modalité

Par défaut, seules les modalités apparaissant dans l'ensemble de données sont disponibles. Si nécessaire, vous pouvez ajouter une modalité à la visualisation.

▶ Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans SPSS Modeler du menu Affichage, choisissez Catégories.

▶ Dans la palette Modalités, cliquez sur le bouton Ajouter une modalité :

Figure 5-110
Bouton Ajouter une modalité

▶ Dans la boîte dialogue Ajouter une nouvelle modalité, saisissez le nom de la modalité.

▶ Cliquez sur OK.

Comment exclure des catégories spécifiques

▶ Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans SPSS Modeler du menu Affichage, choisissez Catégories.

▶ Dans la palette Catégories, sélectionnez un nom de catégorie dans la liste Inclure, puis cliquez sur le bouton X. Pour inclure de nouveau la catégorie, sélectionnez son nom dans la liste Exclus, puis cliquez sur la flèche à droite de la liste.

Fusionner /combiner de petites modalités

Vous pouvez associer des catégories que vous n'avez pas besoin d'afficher séparément du fait de leur petite taille. Par exemple, si votre diagramme à secteurs comporte de nombreuses modalités, il est conseillé de fusionner les modalités dont le pourcentage est inférieur à 10. La fusion n'est disponible que pour les statistiques additives. Par exemple, il est impossible d'ajouter des moyennes ensemble car les moyennes ne sont pas additives. Par conséquent, la combinaison/fusion de modalités à l'aide d'une moyenne n'est pas disponible.

- Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

  Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans SPSS Modeler du menu Affichage, choisissez Catégories.

- Dans la palette Catégories, sélectionnez Réduire et indiquez un pourcentage. Toutes les catégories dont le pourcentage du total est inférieur au nombre indiqué sont associées en une seule catégorie. Le pourcentage est basé sur la statistique présentée dans le tableau. La réduction est disponible uniquement pour les statistiques basées sur un dénombrement ou une addition (somme).

Modification de l'orientation des panels

Si vous utilisez des panels dans votre visualisation, vous pouvez modifier leur orientation.

Comment changer l'orientation des panneaux

- Sélectionnez une partie de la visualisation.
- Cliquez sur l'onglet Panels dans la palette des propriétés.

Figure 5-111
Onglet Panneaux

- Sélectionnez une option dans Présentation :

  Table. Dispose les panneaux comme une table : une ligne ou une colonne est attribuée à chaque valeur individuelle.

  Transposé. Dispose les panneaux comme une table, mais permute également les lignes et les colonnes d'origine. Cette option est différente de l'option de transposition du graphique lui-même. Les axes *x* et *y* restent inchangés lorsque vous sélectionnez cette option.

  Liste. Dispose les panneaux comme une liste : chaque cellule représente une combinaison de valeurs. Les colonnes et les lignes ne sont plus assignées à des valeurs individuelles. Cette option permet de réorganiser les panneaux si nécessaire.

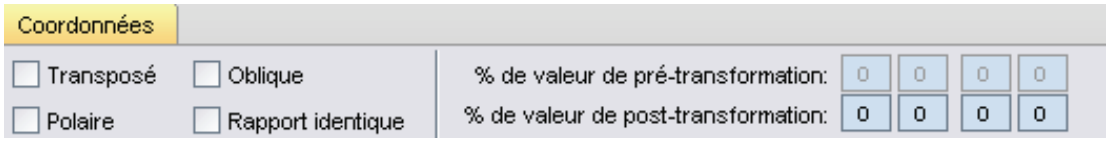
Transformation du système de coordonnées

De nombreuses visualisations sont affichées dans un système de coordonnnées plat et rectangulaire. Vous pouvez transformer le système de coordonnées si nécessaire. Par exemple, vous pouvez appliquer une transformation polaire au système de coordonnées, ajouter des effets d'ombrage oblique et transposer des axes. Vous pouvez également annuler toutes ces transformations si elles ont déjà été appliquées à la visualisation actuelle. Par exemple, un diagramme en secteurs est placé dans un système de coordonnées polaires. Si vous le souhaitez, vous pouvez annuler la transformation polaire et afficher le diagramme en secteurs sous la forme d'un seul diagramme en bâtons empilés dans un système de coordonnées rectangulaire.

Transformer le système de coordonnées

- Sélectionnez le système de coordonnées à transformer. Vous sélectionnez le système de coordonnées en sélectionnant le cadre autour du diagramme individuel.
- Cliquez sur l'onglet Coordonnées dans la palette des propriétés.

Figure 5-112
Onglet Coordonnées



- Sélectionnez les transformations que vous voulez appliquer au système de coordonnées. Vous pouvez également désélectionner une transformation pour l'annuler.

Transposé. L'opération consistant à changer l'orientation des axes est nommée **transposition**. Elle ressemble à l'inversion de l'axe vertical et de l'axe horizontal dans une visualisation en 2-D.

Polaire. Une transformation polaire dessine les éléments graphiques à un angle et une distance spécifiques du centre du diagramme. Un diagramme en secteurs est une visualisation en 1-D avec une transformation polaire qui dessine les bâtons individuels à des angles spécifiques. Un diagramme radar est une visualisation 2-D avec une transformation polaire qui dessine des éléments gaphiques à un angle et une distance spécifiques du centre du diagramme. Une visualisation 3-D inclurait également une dimension de profondeur supplémentaire.

Oblique. Une transformation oblique ajoute un effet 3-D aux éléments grahiques. Cette transformation ajoute de la profondeur aux éléments graphiques mais la profondeur est purement décorative. Elle n'est influencée par aucune valeur de donnée particulière.

Même rapport. Appliquer le même rapport indique que la même distance sur chaque échelle représente le même écart entre les valeurs de données. Par exemple, 2 cm sur les deux échelles représentent un écart de 1000.

% de marge avant transformation. Si les axes sont tronqués après la transformation, il est conseillé d'ajouter des marges au diagramme avant d'effectuer la transformation. Les marges réduisent les dimensions selon un certain pourcentage avant que toute transformation ne soit effectuée sur le système de coordonnées. Vous pouvez contrôler les dimensions de l'axe *x* inférieur,*x* supérieur, *y* inférieur *y* supérieur, dans cet ordre.

% de marge après transformation. Si vous souhaitez modifier le rapport hauteur/largeur du diagramme, vous pouvez y ajouter des marges après avoir effectué la transformation. Les marges réduisent les dimensions selon un certain pourcentage après que les transformations aient été effectuées sur le système de coordonnées. Ces marges peuvent également être appliquées même si aucune transformation n'est effectuée sur le diagramme. Vous pouvez contrôler les dimensions de l'axe *x* inférieur,*x* supérieur, *y* inférieur *y* supérieur, dans cet ordre.

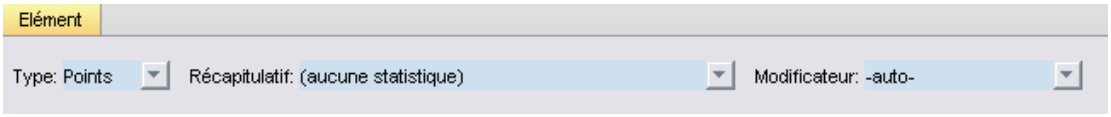
Modification des statistiques et des éléments graphiques

Vous pouvez convertir un en un autre type, modifier la statistique utilisée pour dessiner l'élément graphique ou spécifier le modificateur de collision qui détermine ce qui se passe lorsque les éléments graphiques se chevauchent.

Convertir un élément graphique

- Sélectionnez l'élément graphique que vous souhaitez convertir.
- Cliquez sur l'onglet Elément dans la palette des propriétés.

Figure 5-113
Onglet Elément



- Choisissez un nouveau type d'élément graphique dans la liste Type.

Type d'élément graphique	Description
Point	Un marqueur identifiant un point de données spécifique. Un point est utilisé dans les diagrammes de dispersion et dans d'autres visualisations associées.
Intervalle	Une forme rectangulaire dessinée pour une valeur de données spécifique et remplissant l'espace entre une valeur de données d'origine et une autre valeur. Un intervalle est utilisé dans les diagrammes en bâtons et les histogrammes.
Ligne	Une ligne qui relie des valeurs de données.
Chemin	Une ligne qui relie des valeurs de données selon l'ordre dans lequel elles apparaissent dans l'ensemble de données.
Aire	Une ligne qui relie des données, l'aire entre la ligne et la valeur d'origine étant pleine.
Polygone	Une forme à plusieurs côtés entourant une zone de données. Un élément de polygone peut être utilisé dans un diagramme de dispersion mis en intervalles ou une carte.
Schéma	Un élément consistant en une boîte à moustaches et des marqueurs indiquant des valeurs éloignées. Un élément de schéma est utilisé pour les boîtes à moustaches.

Modifier la statistique

- Sélectionnez l'élément graphique dont vous souhaitez modifier la statistique.

- Cliquez sur l'onglet Elément dans la palette des propriétés.
- Dans la liste déroulante Récapitulatif, sélectionnez une nouvelle statistique. Veuillez noter que sélectionner une statistique agrège les données. Si vous préférez que la visualisation affiche les données non agrégées, sélectionnez (aucune statistique) dans la liste Récapitulatif.

Statistiques récapitulatives calculées à partir d'un champ continu

- **Moyenne.** Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.
- **Median.** Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.
- **Mode.** Valeur qui revient le plus fréquemment. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.
- **Minimum.** Valeur la plus petite d'une variable numérique.
- **Maximum.** Plus grande valeur d'une variable numérique.
- **Intervalle**. La différence entre les valeurs minimale et maximale.
- **Intervalle de milieu**. Le milieu de l'intervalle est la valeur pour laquelle la distance du minimum est égale à la distance du maximum.
- **Sum.** Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- **Somme cumulée**. Somme cumulée des valeurs. Chaque élément graphique affiche la somme correspondant à un sous-groupe à laquelle la somme totale des groupes précédents est ajoutée.
- **Somme de pourcentage**. Le pourcentage de chaque sous-groupe basé sur une valeur de champ comparée à la somme de tous les groupes.
- **Somme de pourcentage cumulé**. Le pourcentage cumulatif de chaque sous-groupe basé sur une valeur de champ comparée à la somme de tous les groupes. Chaque élément graphique affiche le pourcentage correspondant à un sous-groupe auquel le pourcentage total des groupes précédents est ajouté.
- **Variance.** Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.
- **Ecart-type.** Mesure de dispersion par rapport à la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si la moyenne d'âge est de 45 avec un écart-type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.
- **Erreur standard.** Mesure du degré de variation de la valeur d'une statistique test, d'un échantillon à l'autre. Il s'agit de l'écart-type de la distribution de l'échantillon pour une statistique. Par exemple, l'erreur standard de la moyenne est l'écart-type des moyennes d'échantillon.

- **Kurtosis.** Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.
- **Skewness.** Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Les statistiques suivantes de zone peuvent résulter en plusieurs éléments graphiques par sous-groupe. Lorsque des éléments graphiques tels les intervalles, les aires ou les bords sont utilisés, une statistique de zone résulte en un élément graphique affichant l'intervalle. Tous les autres éléments graphiques résultent en deux éléments distincts, l'un affichant le début de l'intervalle et l'autre affichant la fin de celui-ci.

- **Région : Intervalle**. L'intervalle de valeurs entre les valeurs minimale et maximale.
- **Région : Intervalle de confiance de moyenne de 95 %**. Intervalle de valeurs ayant 95% de chances de contenir la valeur moyenne de la population.
- **Région : Intervalle de confiance de 95 % d'un individu**. Intervalle de valeurs ayant 95% de chances de contenir la valeur prédite d'une observation individuelle.
- **Région : Ecart-type de 1 inférieur/supérieur à la moyenne**. Intervalle de valeurs entre un **écart-type** de 1 au-dessus et au-dessous de la **moyenne**.
- **Région : Erreur standard de 1 inférieure/supérieure à la moyenne**. Intervalle de valeurs entre une **erreur standard** de 1 au-dessus et au-dessous de la **moyenne**.

Statistiques récapitulatives basées sur un comptage

- **Comptage**. Nombre d'observations / de lignes
- **Nombre cumulé**. Nombre d'observations / de lignes cumulé. Chaque élément graphique affiche l'effectif correspondant à un sous-groupe auquel l'effectif total des groupes précédents est ajouté.
- **Pourcentage de compte**. Le pourcentage des observations / lignes dans chaque sous-groupe comparé au nombre total d'observations / de lignes.
- **Pourcentage de compte cumulé**. Le pourcentage cumulé des observations / lignes dans chaque sous-groupe comparé au nombre total d'observations / de lignes. Chaque élément graphique affiche le pourcentage correspondant à un sous-groupe auquel le pourcentage total des groupes précédents est ajouté.

Spécifier le modificateur de collision

Le modificateur de collision détermine ce qui arrive lorsque des éléments graphiques se chevauchent.

- Sélectionnez l'élément graphique dont vous souhaitez spécifier le modificateur de collision.
- Cliquez sur l'onglet Elément dans la palette des propriétés.
- Dans la liste déroulante Modificateur, sélectionnez un modificateur de collision. -auto- permet à l'application de déterminer quel modificateur de collision est adapté au type de l'élément graphique et à la statistique.

Superposé. Tracez des éléments graphiques les uns sur les autres lorsqu'ils possèdent la même valeur.

Empilement. Empile des éléments graphiques qui devraient normalement être superposés lorsqu'ils ont les mêmes valeurs de date.

Dodge (regroupement). Déplace les éléments graphiques près d'autres éléments graphiques qui ont la même valeur, plutôt que de les superposer. Les éléments graphiques sont arrangés de manière symétrique. C'est-à-dire que les éléments graphiques sont déplacés sur les côtés opposés d'une position centrale. Le Dodge (regroupement) est très semblable à la classification hiérarchique.

Pile. Déplace les éléments graphiques près d'autres éléments graphiques qui ont la même valeur, plutôt que de les superposer. Les éléments graphiques sont arrangés de manière asymétrique. C'est-à-dire que les éléments graphiques sont empilés les uns au-dessus des autres, avec l'élément graphique du bas positionné sur une valeur spécifique de l'échelle.

Brouillage (normal). Repositionne de manière aléatoire les éléments graphiques à la même valeur de date en utilisant une distribution normale.

Brouillage (uniforme). Repositionne de manière aléatoire les éléments graphiques à la même valeur de date en utilisant une distribution uniforme.

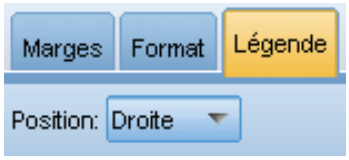
Changement de la position de la légende

Si un diagramme contient une légende, cette légende apparaît généralement à droite du diagramme. Vous pouvez changer cette position si nécessaire.

Comment changer la position de la légende

- Sélectionnez la légende.
- Cliquez sur l'onglet Légende dans la palette des propriétés.

Figure 5-114
Onglet Légende



▶ Sélectionnez une position.

Copie d'une visualisation et des données de visualisation

La palette Générale comprend des boutons pour copier la visualisation et ses données.

Figure 5-116
Bouton Copie de la visualisation

Copie de la visualisation. Cette action copie la visualisation dans le presse-papiers comme une image. Plusieurs formats d'image sont disponibles. Lorsque vous collez l'image dans une autre application, vous pouvez choisir une option "Collage spécial" pour sélectionner un des formats d'image disponibles pour le collage.

Figure 5-117
Bouton Copie des données de la visualisation

Copie des données de la visualisation. Cette action copie les données sous-jacentes utilisées pour dessiner la visualisation. Ces données sont copiées dans le presse-papiers comme texte brut ou comme texte au format HTML. Lorsque vous collez des données dans une autre application, vous pouvez choisir une option "Collage spécial" pour sélectionner un des formats d'image disponibles pour le collage.

Raccourcis clavier

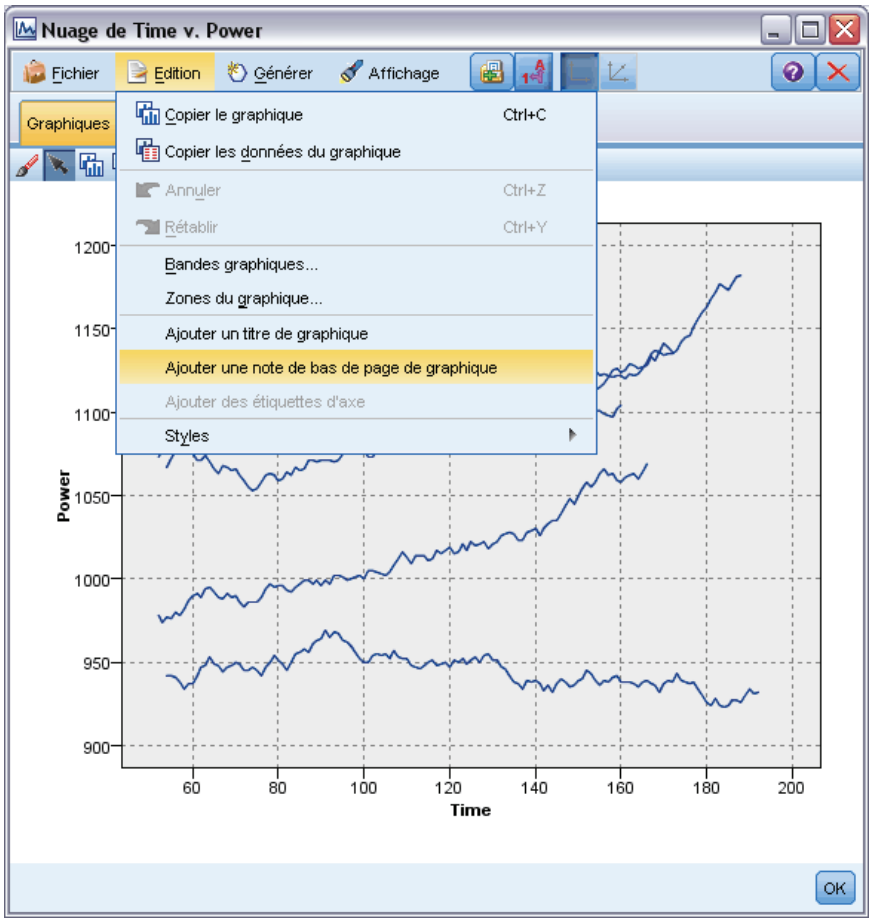
Table 5-4
Raccourcis clavier

Touche de raccourci	Fonction
Ctrl+Espace	Passer du mode Explorer au mode Edition
Suppr	Supprimer un élément de la visualisation
Ctrl+Z	Annuler
Ctrl+Y	Rétablir
F2	Afficher des conseils pour sélectionner des éléments du graphique

Ajout de titres et de notes de bas de page

Pour tous les types de graphique, vous pouvez ajouter un titre unique, une note de bas de page ou des étiquettes d'axe afin d'identifier ce que représente le graphique.

Figure 5-118
Ajout d'une note de bas de page de graphique



Ajout de titre aux graphiques

- Dans les menus, sélectionnez Edition > Ajouter un titre de graphique. Une zone de texte contenant <TITRE> apparaît au-dessus du graphique.
- Vérifiez que vous êtes en mode d'édition. Dans les menus, sélectionnez Affichage > Mode d'édition.
- Double-cliquez sur le texte <TITRE>.
- Entrez le titre souhaité et appuyez sur Entrée.

Ajout de notes de bas de page aux graphiques

- Dans les menus, sélectionnez Edition > Ajouter une note de bas de page de graphique. Une zone de texte contenant <NOTE DE BAS DE PAGE> apparaît sous le graphique.
- Vérifiez que vous êtes en mode d'édition. Dans les menus, sélectionnez Affichage > Mode d'édition.
- Double-cliquez sur le texte <NOTE DE BAS DE PAGE>.

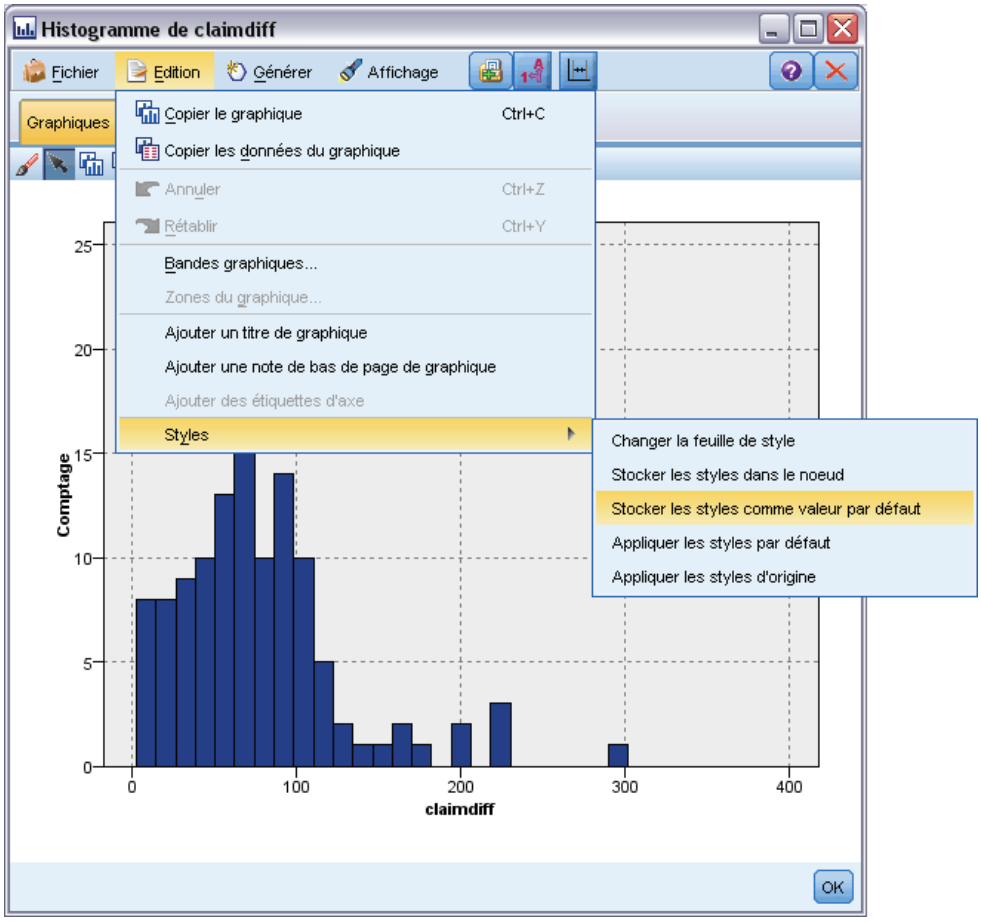
▶ Entrez le titre souhaité et appuyez sur Entrée.

Utilisation de feuilles de style de graphique

Les informations de base relatives à l'affichage du graphique telles que les couleurs, polices, symboles et épaisseurs de lignes sont contrôlées par une feuille de style. Une feuille de style par défaut est fournie avec IBM® SPSS® Modeler ; néanmoins, vous pouvez la modifier si nécessaire. Par exemple, vous pouvez disposer d'un modèle de couleurs d'entreprise pour les présentations à utiliser dans les graphiques. Pour plus d'informations, reportez-vous à la section Modification des visualisations sur p. 372.

Dans les noeuds Graphiques, vous pouvez utiliser le Mode d'édition pour modifier les styles d'apparence d'un graphique. Vous pouvez ensuite utiliser le menu Edition > Styles pour enregistrer les modifications sous forme de feuille de style qui s'applique à tous les graphiques que vous générez ensuite à partir du noeud Graphique actuel ou sous forme de nouvelle feuille de style par défaut pour tous les graphiques que vous produisez à l'aide de SPSS Modeler.

Figure 5-119
Sélection de styles de graphique



Quatre options de feuille de style sont disponibles à partir de l'option Styles du menu Edition :

- **Changer la feuille de style.** Cette option affiche une liste des feuilles de styles différentes stockées parmi lesquelles vous pouvez choisir afin de modifier l'apparence de vos graphiques. Pour plus d'informations, reportez-vous à la section Application des feuilles de style sur p. 393.
- **Stocker les styles dans le noeud.** Stocke les modifications apportées aux styles du graphique sélectionné de façon à ce qu'elles soient appliquées aux graphiques futurs créés à partir du même noeud Graphique, dans le flux actuel.
- **Stocker les styles comme valeur par défaut.** Stocke les modifications apportées aux styles du graphique sélectionné de façon à ce qu'elles soient appliquées aux graphiques futurs créés à partir d'un noeud Graphique, dans n'importe quel flux. Une fois cette option sélectionnée, vous pouvez utiliser l'option Appliquer les styles par défaut pour modifier d'autres graphiques existants afin qu'ils utilisent les mêmes styles.
- **Appliquer les styles par défaut.** Remplace les styles du graphique sélectionné par les styles actuellement enregistrés comme styles par défaut.
- **Appliquer les styles d'origine.** Rétablit les styles d'un graphique sur les styles fournis comme styles d'origine par défaut.

Application des feuilles de style

Vous pouvez appliquer une feuille de style de visualisation qui spécifie les propriétés de style de la visualisation. Par exemple, la feuille de style peut définir les polices, les tirets, les couleurs, parmi d'autres options. Dans une certaine mesure, les feuilles de style fournissent un raccourci pour les modifications que vous auriez effectuées manuellement. Remarque : les feuilles de style sont limitées aux modifications de *style*. Les autres modifications telles la position des légendes ou l'intervalle de l'échelle ne sont pas stockées dans la feuille de style.

Appliquer une feuille de style

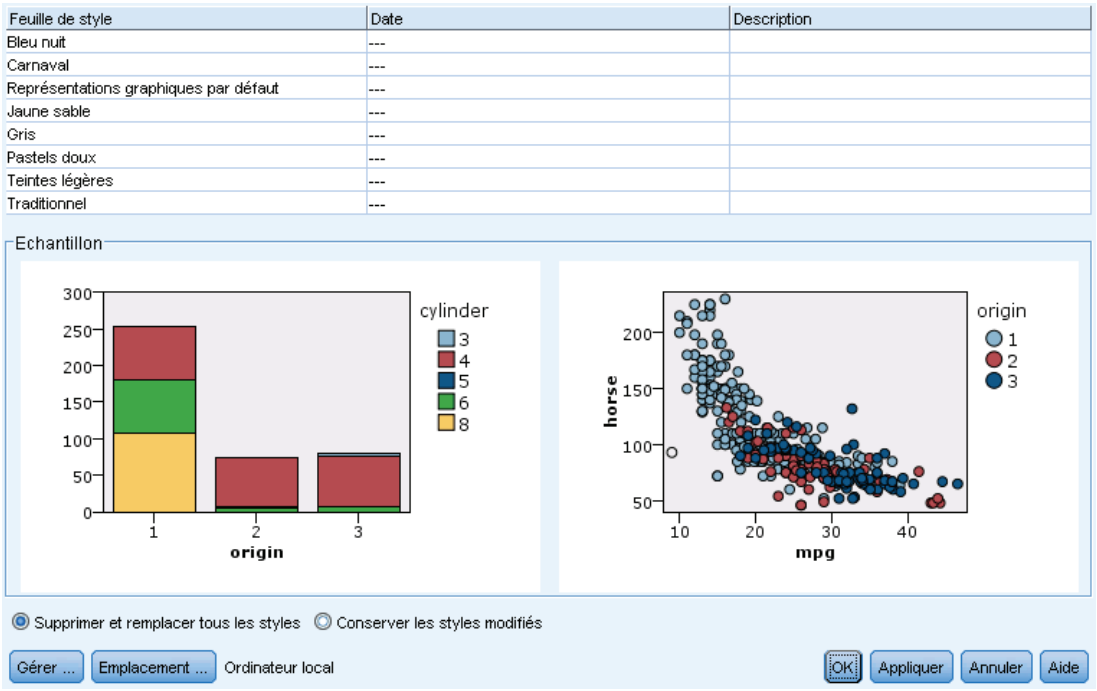
▶ Dans les menus, sélectionnez :
Edition > Styles > Changer la feuille de style

▶ Utilisez la boîte de dialogue Changer la feuille de style pour sélectionner une feuille de style.

▶ Cliquez sur Appliquer pour appliquer la feuille de style à la visualisation sans fermer la boîte de dialogue. Cliquez sur OK pour appliquer la feuille de style et fermer la boîte de dialogue.

Boîte de dialogue Changer/Sélectionner la feuille de style

Figure 5-120
Boîte de dialogue Changer la feuille de style



Le tableau en haut de la boîte de dialogue répertorie toutes les feuilles de style de visualisation actuellement disponibles. Quelques feuilles de style sont préinstallées, tandis que d'autres ont été créées dans IBM® SPSS® Visualization Designer (un produit distinct).

Le bas de la boîte de dialogue affiche des exemples de visualisations contenant des exemples de données. Sélectionnez une des feuilles de style pour l'appliquer aux exemples de visualisation. Ces exemples peuvent vous aider à déterminer la manière dont la feuille de style va affecter votre visualisation.

La boîte de dialogue offre également d'autres options.

Styles existants. Par défaut, une feuille de style peut remplacer tous les styles dans une visualisation. Vous pouvez modifier ce comportement.

- **Remplacer tous les styles**. Lorsque vous appliquez une feuille de style, cette option permet de remplacer tous les styles de la visualisation, y compris les styles modifiés durant la session en cours.
- **Conserver les styles modifiés**. Lorsque vous appliquez une feuille de style, cette option permet de remplacer uniquement les styles de la visualisation qui n'ont *pas* été modifiés durant la session en cours. Les styles modifiés durant la session en cours sont conservés.

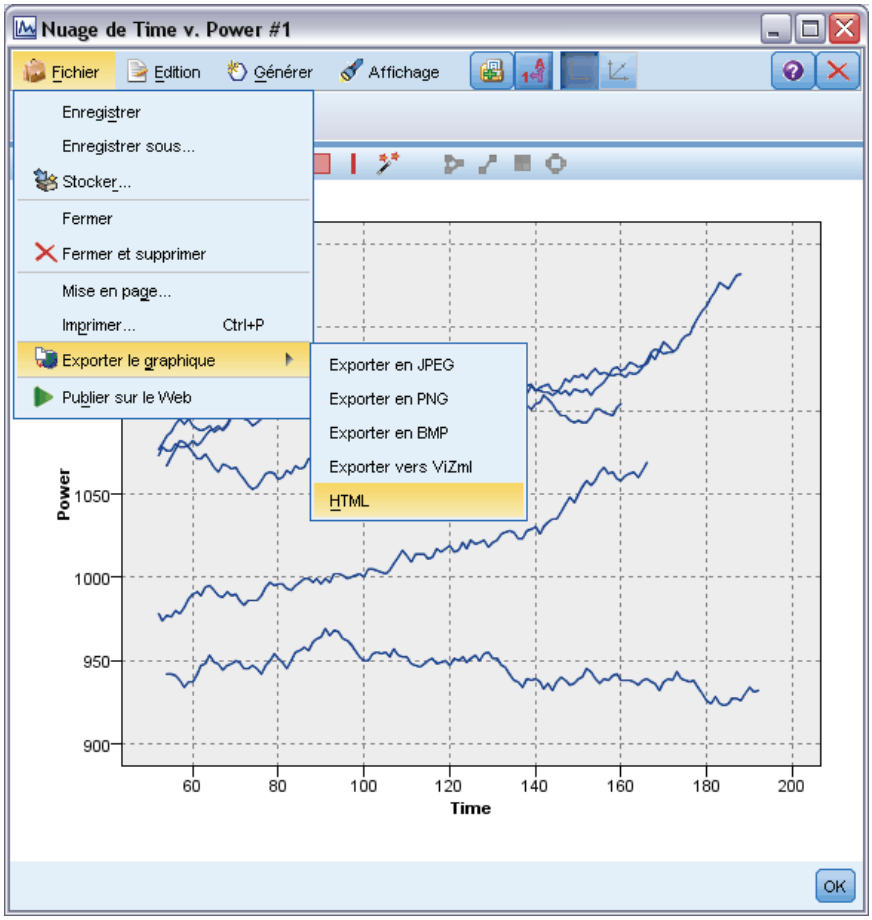
Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, reportez-vous à la section Gestion des modèles, des feuilles de style et des fichiers cartes sur p. 292.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est noté à droite du bouton. Pour plus d'informations, reportez-vous à la section Définition de l'emplacement des modèles, des feuilles de style et des cartes sur p. 290.

Impression, enregistrement, copie et exportation de graphiques

Chaque graphique présente un certain nombre d'options relatives à l'enregistrement, à l'impression ou bien encore à l'exportation vers un autre format. La plupart de ces options sont disponibles dans le menu Fichier. En outre, dans le menu Edition, vous pouvez choisir de copier le graphique ou les données afin de les utiliser dans une autre application.

Figure 5-121
Menu Fichier et barre d'outils des fenêtres de graphique



Impression

- Pour imprimer le graphique, utilisez l'option de menu ou le bouton Imprimer. Avant l'impression, vous pouvez utiliser les options Mise en page et Aperçu avant impression pour définir les options d'impression et prévisualiser la sortie.

Enregistrement des graphiques

▶ Pour enregistrer le graphique dans un fichier de sortie IBM® SPSS® Modeler (**.cou*), choisissez l'option Fichier > Enregistrer ou Fichier > Enregistrer sous dans les menus.

ou

Pour enregistrer le graphique dans le référentiel, choisissez Fichier > Stocker la sortie dans les menus.

Copier les graphiques

▶ Pour copier le graphique en vue d'une utilisation dans une autre application, telle que MS Word ou MS PowerPoint, sélectionnez Edition > Copier le graphique dans les menus.

Copier des données

▶ Pour copier les données en vue d'une utilisation dans une autre application, telle que MS Excel ou MS Word, sélectionnez Edition > Copier les données dans les menus. Par défaut, les données sont formatées en HTML. Utilisez Collage spécial dans l'autre application pour voir d'autres options de formatage lors du collage.

Exportation de graphiques

L'option Exporter le graphique vous permet d'exporter le graphique dans l'un des formats suivants : Bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*), ou document ViZml (*.xml*) à utiliser dans d'autres applications IBM® SPSS® Statistics

▶ Pour exporter des graphiques, sélectionnez Fichier > Exporter le graphique dans les menus puis sélectionnez le format.

Exportation de tables

L'option Exporter la table vous permet d'exporter la table dans l'un des formats suivants : délimité par des tabulations (*.tab*), délimité par des virgules (*.csv*), ou HTML (*.html*)

▶ Pour exporter des tables, sélectionnez Fichier > Exporter la table dans les menus puis sélectionnez le format.

Chapitre
6

Noeuds de sortie

Présentation des noeuds de sortie

Les noeuds de sortie permettent d'obtenir des informations sur vos données et vos modèles. Ils permettent également d'exporter les données dans divers formats, afin de pouvoir les utiliser avec d'autres logiciels.

Les noeuds de sortie disponibles sont les suivants :

Le noeud Table affiche les données au format tabulaire (ces données peuvent également être écrites dans un fichier). Ainsi, vous pouvez passer en revue les valeurs de données ou les exporter dans un format facilement lisible. Pour plus d'informations, reportez-vous à la section Noeud Table sur p. 404.

Le noeud Matrice permet de créer un tableau dans lequel les relations entre les champs sont indiquées. Il s'agit généralement de deux champs symboliques, mais il peut également s'agir de champs booléens ou numériques. Pour plus d'informations, reportez-vous à la section Noeud Matrice sur p. 410.

Le noeud Analyse évalue la capacité des modèles prédictifs à générer des prévisions précises. Les noeuds Analyse comparent les valeurs prédites et les valeurs réelles d'un ou de plusieurs nuggets de modèle. Ils peuvent également comparer entre eux les modèles prédictifs. Pour plus d'informations, reportez-vous à la section Noeud Analyse sur p. 415.

Le noeud Audit données fournit un premier aperçu complet des données, notamment des statistiques récapitulatives, des histogrammes et distributions pour chaque champ, ainsi que des informations sur les valeurs éloignées, les valeurs manquantes et les valeurs extrêmes. Les résultats sont affichés dans une matrice facile à lire pouvant être triée et utilisée pour générer les noeuds de préparation des données et des graphiques grandeur nature. Pour plus d'informations, reportez-vous à la section Noeud Audit données sur p. 420.

Le noeud Transformation vous permet de sélectionner et de prévisualiser les résultats des transformations avant de les appliquer aux champs sélectionnés. Pour plus d'informations, reportez-vous à la section Noeud Transformation sur p. 435.

Le noeud Statistiques fournit des informations récapitulatives de base sur les champs numériques. Ill calcule les statistiques récapitulatives des champs individuels et des corrélations entre les champs. Pour plus d'informations, reportez-vous à la section Noeud Statistiques sur p. 441.

Le noeud Moyennes compare les moyennes de groupes indépendants ou de paires de champs associés, afin de détecter toute différence sensible. Par exemple, vous pouvez comparer les revenus moyens avant et après l'application d'une augmentation, ou comparer les revenus des personnes ayant obtenu une augmentation avec ceux des personnes qui n'en ont pas eu. Pour plus d'informations, reportez-vous à la section Noeud Moyennes sur p. 446.

© Copyright IBM Corporation 1994, 2012.

Ce noeud permet de créer des rapports formatés contenant du texte fixe et des données, ainsi que des expressions calculées à partir de ces dernières. Le format du rapport est déterminé par des modèles texte définissant la structure du texte fixe et de la sortie de données. Vous pouvez définir un formatage de texte personnalisé en utilisant des balises HTML dans le modèle et en définissant des options dans l'onglet Sortie. Vous pouvez inclure des valeurs de données et d'autres sorties conditionnelles à l'aide des expressions CLEM du modèle. Pour plus d'informations, reportez-vous à la section Noeud Rapport sur p. 452.

Le nœud V. globales (Valeurs globales) analyse les données et calcule des valeurs récapitulatives pouvant être utilisées dans des expressions CLEM. Par exemple, vous pouvez utiliser ce noeud pour calculer les statistiques d'un champ *âge*, puis utiliser la moyenne globale du champ *age* dans des expressions CLEM en insérant la fonction @GLOBAL_MEAN(age). Pour plus d'informations, reportez-vous à la section Noeud V. globales (Valeurs globales) sur p. 455.

Gestion des sorties

Le gestionnaire des sorties affiche les diagrammes, les graphiques et les tableaux générés lors d'une session IBM® SPSS® Modeler. Vous pouvez toujours rouvrir une sortie en double-cliquant dessus dans le gestionnaire ; il est inutile de réexécuter le flux ou le noeud correspondant.

Pour afficher le gestionnaire des sorties

▶ Ouvrez le menu Affichage et choisissez Gestionnaires. Cliquez sur l'onglet Sorties.

Figure 6-1
Gestionnaire des sorties

Dans le gestionnaire des sorties, vous pouvez effectuer les opérations suivantes :

- Afficher des objets de sortie existants, tels que des histogrammes, des graphiques Evaluation et des tableaux.
- Renommer des objets de sortie.
- Enregistrer des objets de sortie sur disque ou dans le IBM® SPSS® Collaboration and Deployment Services Repository (s'il est disponible).
- Ajouter des fichiers de sortie au projet actuel.
- Supprimer des objets de sortie non enregistrés de la session actuelle.
- Ouvrir les objets de sortie enregistrés ou les récupérer dans le IBM SPSS Collaboration and Deployment Services Repository (s'il est disponible).

Pour accéder à ces options, cliquez avec le bouton droit de la souris sur l'onglet Sorties.

Affichage des sorties

La sortie à l'écran est affichée dans une fenêtre du navigateur de sortie. La fenêtre du navigateur de sortie comporte ses propres menus, lesquels vous permettent d'imprimer ou d'enregistrer la sortie, ou de l'exporter dans un autre format. Notez que les options proposées peuvent varier en fonction du type de sortie.

Impression, enregistrement et exportation de données. Pour plus d'informations, procédez comme suit :

- Pour imprimer la sortie, utilisez le bouton ou l'option de menu Imprimer. Avant l'impression, vous pouvez utiliser les options Mise en page et Aperçu avant impression pour définir les options d'impression et prévisualiser la sortie.
- Pour enregistrer la sortie dans un fichier de sortie IBM® SPSS® Modeler (*.cou*), choisissez l'option Enregistrer ou Enregistrer sous dans le menu Fichier.
- Pour enregistrer la sortie dans un autre format (texte ou HTML, par exemple), choisissez Exporter dans le menu Fichier. Pour plus d'informations, reportez-vous à la section Exportation des sorties sur p. 402.
- Pour enregistrer la sortie dans un référentiel partagé afin que les autres utilisateurs puissent le consulter via le IBM® SPSS® Collaboration and Deployment Services Deployment Portal, choisissez Publier sur le Web dans le menu Fichier. Notez que cette option requiert une licence distincte pour IBM® SPSS® Collaboration and Deployment Services.

Sélection de cellules et de colonnes. Le menu Edition contient plusieurs options permettant de sélectionner, de désélectionner et de copier des cellules et des colonnes, pour le type de sortie actuel. Pour plus d'informations, reportez-vous à la section Sélection de cellules et de colonnes sur p. 403.

Création de noeuds. Le menu Générer permet de créer des noeuds sur la base du contenu du navigateur de sortie. Les options varient en fonction du type de sortie et des éléments de la sortie actuellement sélectionnés. Pour plus d'informations sur les options de création de noeud pour un type de sortie donné, reportez-vous à la documentation propre à la sortie.

Publication sur le Web

La fonctionnalité Publier sur le Web vous permet de publier certains types de flux de sortie dans un IBM® SPSS® Collaboration and Deployment Services Repository partagé central qui forme la base de IBM® SPSS® Collaboration and Deployment Services. Si vous utilisez cette option, d'autres utilisateurs qui ont besoin de visualiser cette sortie peuvent le faire en utilisant un accès Internet et un compte IBM SPSS Collaboration and Deployment Services : il est inutile d'installer IBM® SPSS® Modeler.

Remarque : Une licence distincte est requise pour accéder à un référentiel IBM SPSS Collaboration and Deployment Services. Pour plus d'informations, reportez-vous à *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*

Le tableau suivant répertorie les noeuds SPSS Modeler qui prennent en charge la fonctionnalité Publier sur le Web. La sortie à partir de ces noeuds est stockée dans le IBM SPSS Collaboration and Deployment Services Repository dans un format d'objet de sortie (*.cou*), et peut-être affichée directement dans le IBM® SPSS® Collaboration and Deployment Services Deployment Portal.

D'autres types de sorties peuvent être affichés uniquement si l'application appropriée (par exemple,. SPSS Modeler, pour des objets de flux) est installée sur l'ordinateur de l'utilisateur.

Table 6-1
Noeuds qui prennent en charge la fonctionnalité Publier sur le Web

Type de noeud	**Noeud**
Graphiques	all
Sortie	Table
	Matrice
	Audit données
	Transformation
	Moyennes
	Analyse
	Statistiques
	Rapport (HTML)
IBM® SPSS® Statistics	Sortie du nœud Statistiques

Publication d'un résultat sur le Web

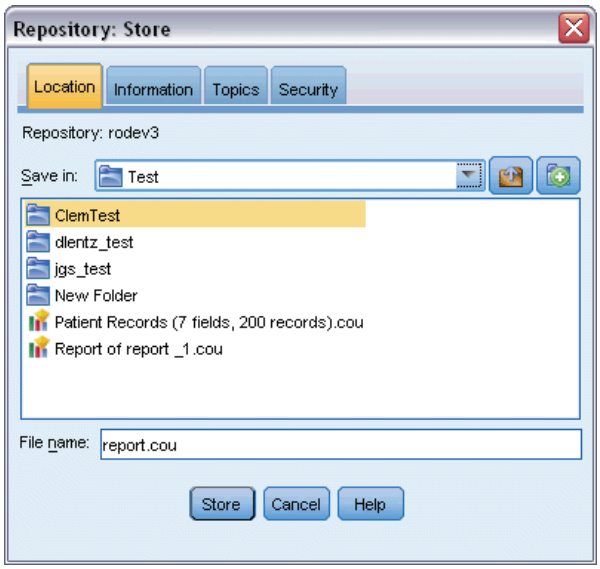
Pour publier un résultat sur le Web :

- Dans un flux IBM® SPSS® Modeler, exécutez l'un des noeuds répertoriés dans le tableau. Un objet de sortie est alors créé (par exemple, un objet tableau , matrice ou rapport) dans une nouvelle fenêtre.
- Dans la fenêtre des objets de sortie, sélectionnez :
  Fichier > Publication sur le Web

  Remarque : pour exporter de simples fichiers HTML à utiliser avec un navigateur Web standard, choisissez Exporter dans le menu Fichier et sélectionnez HTML.
- Connectez-vous au IBM® SPSS® Collaboration and Deployment Services Repository

  Lorsque vous vous êtes connecté avec succès, le référentiel : boîte de dialogue Stocker apparaît et vous propose plusieurs options de stockage.

Figure 6-2
Référentiel : boîte de dialogue Stocker



▶ Lorsque vous avez choisi l'option de stockage de votre choix, cliquez sur Stocker.

Affichage du résultat publié sur le Web

Vous devez disposer d'un compte IBM SPSS Collaboration and Deployment Services configuré pour utiliser cette fonctionnalité. Si l'application appropriée est installée pour le type d'objet que vous souhaitez afficher (par exemple, IBM® SPSS® Modeler ou IBM® SPSS® Statistics), la sortie est affichée dans l'application même plutôt que dans le navigateur.

*Remarque :*Une licence distincte est requise pour accéder à IBM® SPSS® Collaboration and Deployment Services. Pour plus d'informations, reportez-vous à *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*.

Pour afficher un résultat publié sur le Web :

▶ Saisissez l'adresse *http://<repos_host>:<repos_port>/peb* dans votre navigateur

où *repos_host* et *repos_port* sont le nom d'hôte et le numéro de port de l'hôte IBM SPSS Collaboration and Deployment Services.

▶ Saisissez les détails de connexion de votre compte IBM SPSS Collaboration and Deployment Services.

▶ Cliquez sur Référentiel de contenu.

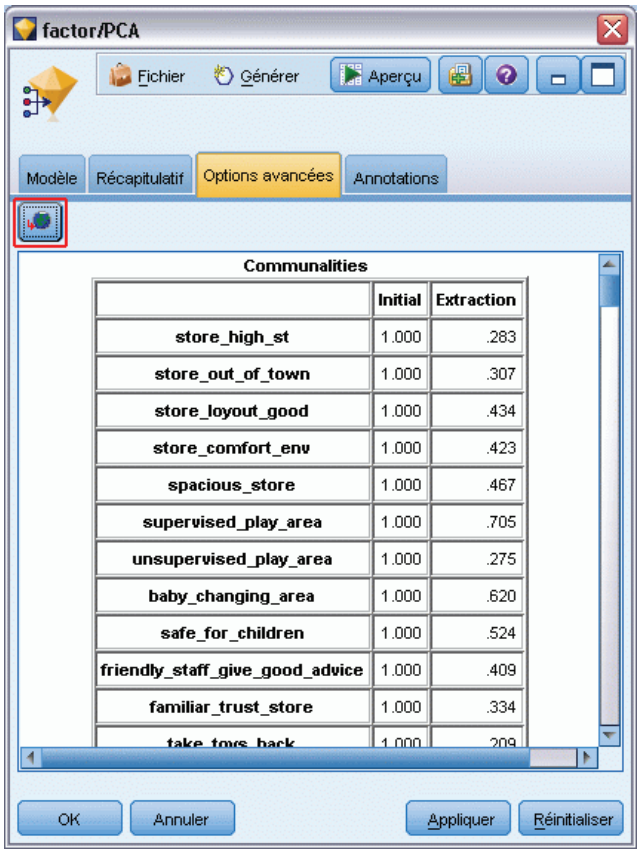
▶ Recherchez ou accédez à l'objet que vous souhaitez afficher.

▶ Cliquez sur le nom de l'objet. Pour certains types d'objets, tels que des diagrammes, il est possible qu'il y ait un délai car l'objet est rendu dans le navigateur.

Affichage de la sortie dans un navigateur HTML

A partir de l'onglet Options avancées des nuggets de modèle Linéaire, Logistique et ACP/Facteur, vous pouvez afficher les informations affichées dans un navigateur distinct tel qu'Internet Explorer. Le format de sortie des informations est HTML ; vous pouvez alors les enregistrer et les réutiliser ailleurs, par exemple sur un réseau Intranet d'entreprise ou un site Internet.

Figure 6-3
Bouton Lancer sur l'onglet Options avancées du nugget de modèle



Pour afficher les informations dans un navigateur, cliquez sur le bouton de lancement, situé sous l'icône de modèle, en haut à gauche de la boîte de dialogue de l'onglet Options avancées du modèle de nugget.

Exportation des sorties

Dans la fenêtre du navigateur de sortie, vous pouvez choisir d'exporter la sortie dans un autre format (texte ou HTML, par exemple). Les formats d'exportation varient selon le type de sortie, mais sont en général semblables aux options de type de fichier disponibles si vous sélectionnez l'option d'enregistrement dans un fichier dans le noeud de génération de la sortie.

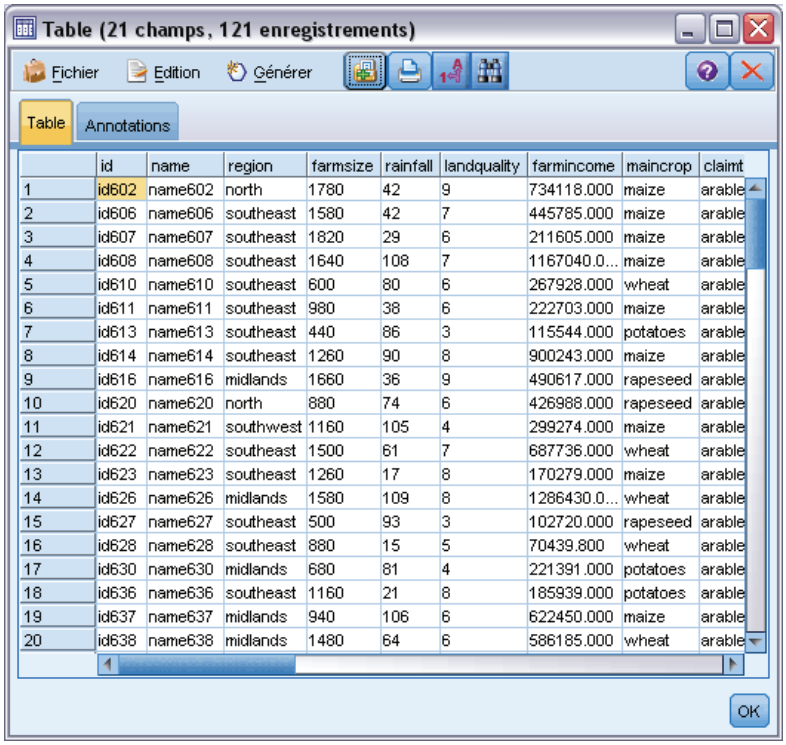
Pour exporter le résultat

▶ Dans le navigateur de sortie, ouvrez le menu Fichier et choisissez Exporter. Sélectionnez ensuite le type de fichier à créer :

- **Délimité par des tabulations (*.tab).** Cette option crée un fichier texte formaté contenant les valeurs de données. Ce style est souvent utilisé pour créer une représentation en texte brut des informations susceptibles d'être importées dans d'autres applications. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
- **Délimité par des virgules (*.dat).** Cette option crée un fichier texte séparé par des virgules contenant les valeurs de données. Ce style est souvent utilisé pour générer rapidement un fichier de données susceptible d'être importé dans un tableur ou un autre logiciel d'analyse de données. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
- **Format délimité par des tabulations transposé (*.tab).** Cette option est identique à l'option Délimité par des tabulations, à une exception près toutefois : les données sont transposées de façon à ce que les lignes et les colonnes représentent respectivement les champs et les enregistrements.
- **Format délimité par des virgules transposé (*.dat).** Cette option est identique à l'option Délimité par des virgules, à une exception près : les données sont transposées de façon à ce que les lignes et les colonnes représentent respectivement les champs et les enregistrements.
- **HTML (*.html).** Cette option permet d'écrire une sortie au format HTML dans des fichiers.

Sélection de cellules et de colonnes

Figure 6-4
Fenêtre du navigateur du nœud Table



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

Un certain nombre de noeuds, notamment les noeuds Table, Matrice et Moyennes, génèrent une sortie tabulaire. Les tableaux de sortie peuvent tous être affichés et utilisés de la même manière. Ainsi, il est possible, entre autres, de sélectionner des cellules, de copier tout ou partie du tableau dans le Presse-papiers, de générer de nouveaux noeuds à partir de la sélection actuelle, et d'enregistrer et d'imprimer le tableau.

Sélection de cellules. Pour sélectionner une cellule, cliquez dessus. Pour sélectionner un groupe de cellules, cliquez sur un angle du groupe voulu, faites glisser le pointeur de la souris jusqu'à l'angle opposé, puis relâchez le bouton de la souris. Pour sélectionner une colonne tout entière, cliquez sur son en-tête. Pour sélectionner plusieurs colonnes, cliquez sur leur en-tête en maintenant la touche Maj ou Ctrl enfoncée.

Toute nouvelle sélection annule la précédente. Si vous maintenez la touche Ctrl enfoncée lorsque vous effectuez une sélection, celle-ci sera ajoutée aux sélections existantes (la sélection précédente n'est pas supprimée). Cette fonction permet de sélectionner plusieurs zones non contiguës du tableau. Le menu Edition contient également les options Sélectionner tout et Effacer la sélection.

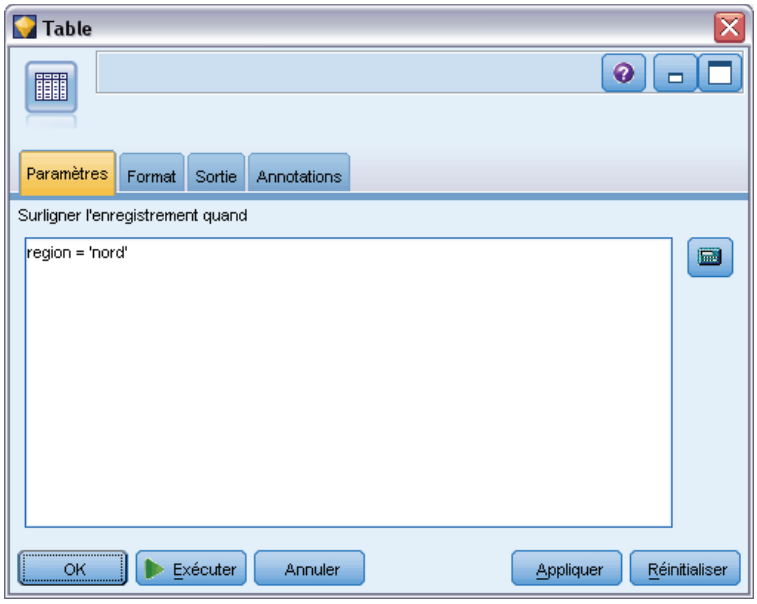
Réorganisation des colonnes. Les navigateurs de sortie des noeuds Table et Moyennes vous permettent de déplacer des colonnes du tableau en cliquant sur l'en-tête correspondant, puis en le faisant glisser vers l'emplacement souhaité. Vous ne pouvez déplacer qu'une seule colonne à la fois.

Noeud Table

Le noeud Table crée un tableau qui répertorie les valeurs dans vos données. Tous les champs et toutes les valeurs du flux sont comprises, ce qui facilite l'inspection des valeurs de vos données ou leur exportation sous une forme facilement lisible. De façon facultative, vous pouvez mettre en évidence des enregistrements qui satisfont à une certaine condition.

Noeud Table - Onglet Paramètres

Figure 6-5
Noeud Table : onglet Paramètres



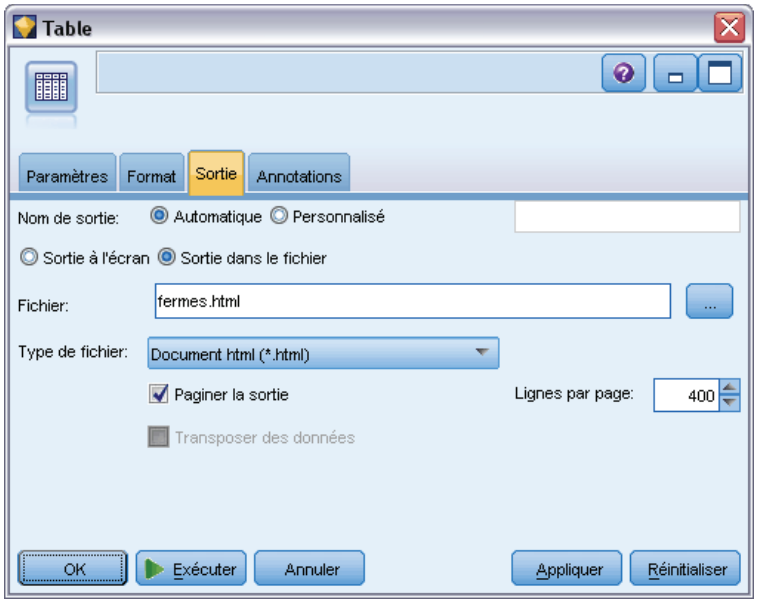
Surligner l'enregistrement quand. Pour mettre en évidence certains enregistrements du tableau, entrez une expression CLEM vraie pour chacun de ces enregistrements. Cette option est activée uniquement lorsque l'option Sortie à l'écran est sélectionnée.

Noeud Table - Onglet Format

L'onglet Format inclut les options permettant d'indiquer le formatage de chaque champ. Cet onglet est partagé avec le nœud Typer. Pour plus d'informations, reportez-vous à la section Onglet Paramètres du champ sur p. 153.

Noeud de sortie - Onglet Sortie

Figure 6-6
Noeud de sortie - Onglet Sortie



Pour les noeuds générant une sortie de type tableau, l'onglet Sortie vous permet de définir le format et l'emplacement des résultats.

Nom de sortie. Spécifie le nom de la sortie générée lorsque le noeud est exécuté. L'option Automatique sélectionne un nom en fonction du nœud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir Personnalisé pour indiquer un autre nom.

Sortie à l'écran (option par défaut)**.** Crée un objet de sortie à afficher en ligne. L'objet de sortie apparaît dans l'onglet Sorties de la fenêtre du gestionnaire lors de l'exécution du noeud de sortie.

Sortie dans le fichier. Enregistre la sortie dans un fichier lors de l'exécution du noeud. Si vous choisissez cette option, entrez un nom de fichier (ou parcourez l'arborescence et indiquez un nom de fichier à l'aide du sélecteur de fichiers), puis sélectionnez un type de fichier. Il se peut que tous les types de fichier ne soient pas disponibles pour certains types de sortie.

Les données sont sorties dans le format d'encodage par défaut du système qui est spécifié dans le Panneau de configuration de Windows, ou si le système est en mode réparti, sur l'ordinateur serveur.

- **Données (délimitées par des tabulations) (*.tab).** Cette option crée un fichier texte formaté contenant les valeurs de données. Ce style est souvent utilisé pour créer une représentation en texte brut des informations susceptibles d'être importées dans d'autres applications. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.

- **Données (séparées par des virgules) (*.dat).** Cette option crée un fichier texte séparé par des virgules contenant les valeurs de données. Ce style est souvent utilisé pour générer rapidement un fichier de données susceptible d'être importé dans un tableur ou un autre logiciel d'analyse de données. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
- **HTML (*.html).** Cette option permet d'écrire une sortie au format HTML dans des fichiers. Dans les sorties tabulaires (noeud Table, Matrice ou Moyennes), les ensembles de fichiers HTML comportent un panneau de contenu répertoriant les noms des champs, ainsi que les données sous forme de tableau HTML. Le tableau peut être partagé entre plusieurs fichiers HTML s'il contient plus de lignes que la valeur définie pour le paramètre Lignes par page. Dans ce cas, le panneau de contenu contient des liens correspondant à toutes les pages du tableau et permet de naviguer dans ce dernier. Dans le cas d'une sortie non tabulaire, un seul fichier HTML contenant les résultats du noeud est créé.

  Remarque : si la sortie HTML ne contient des données de formatage que pour la première page, sélectionnez Paginer la sortie et ajustez le paramètre Lignes par page afin de regrouper toutes les sorties sur une même page. Si le modèle de sortie des noeuds (noeud Rapport, par exemple) contient des balises HTML personnalisées, assurez-vous d'avoir choisi le type de format Personnalisé.
- **Fichier texte (*.txt).** Cette option crée un fichier texte contenant la sortie. Ce style est souvent utilisé pour générer une sortie susceptible d'être importée dans d'autres applications (par exemple, traitement de texte ou création de présentations). Vous ne pouvez pas utiliser cette option avec tous les noeuds.
- **Objet de sortie (*.cou).** Les objets de sortie enregistrés dans ce format peuvent être ouverts et affichés dans IBM® SPSS® Modeler, ajoutés à des projets, ainsi que publiés et suivis via le IBM® SPSS® Collaboration and Deployment Services Repository.

Vue de sortie. Pour le noeud Moyennes, vous pouvez indiquer si vous souhaitez afficher par défaut une sortie simple ou avancée. Vous pouvez également basculer entre ces vues lorsque vous parcourez la sortie générée. Pour plus d'informations, reportez-vous à la section Navigateur de sortie du noeud Moyennes sur p. 449.

Format : Pour le nœud Rapport, vous pouvez choisir de formater automatiquement la sortie ou de la formater à l'aide des paramètres HTML indiqués dans le modèle. Sélectionnez Personnalisé pour permettre le formatage HTML dans le modèle.

Titre : Dans le cas du nœud Rapport, vous pouvez indiquer un titre facultatif qui apparaîtra en haut de la sortie sous forme de rapport.

Surligner le texte inséré. Dans le cas du noeud Rapport, cette option permet de surligner le texte généré par les expressions CLEM dans le modèle de rapport. Pour plus d'informations, reportez-vous à la section Noeud Rapport - Onglet Modèle sur p. 453. Cette option n'est pas recommandée si vous avez opté pour un formatage Personnalisé.

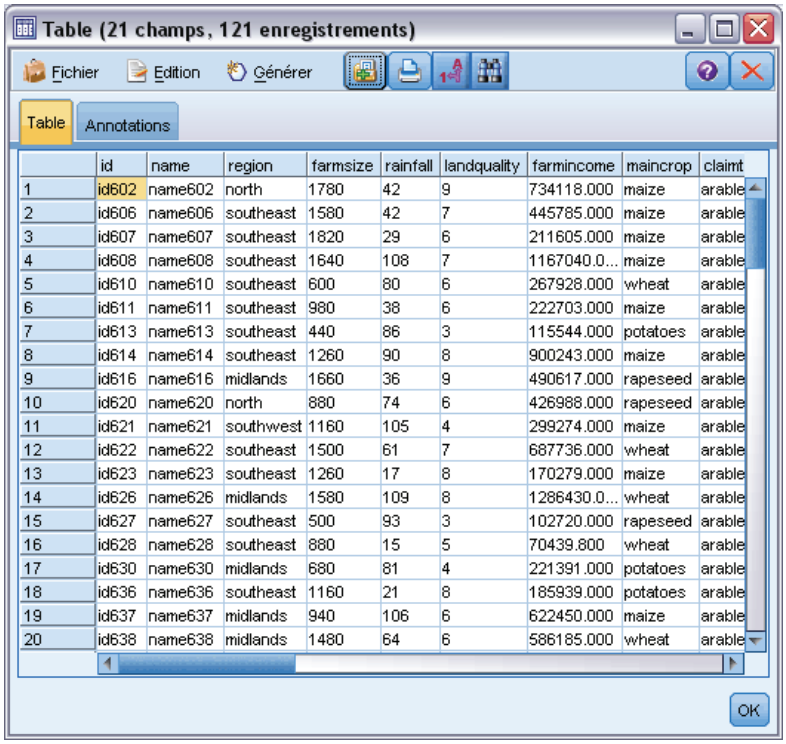
Lignes par page. Pour le nœud Rapport, indiquez le nombre de lignes à inclure sur chaque page lors du formatage automatique du rapport de sortie.

Transposer des données. Cette option transpose les données avant l'exportation, de sorte que les lignes et les colonnes représentent respectivement les champs et les enregistrements.

Remarque : dans le cas de tableaux volumineux, les options ci-dessus ne fonctionnent pas toujours bien, surtout si vous utilisez un serveur distant. Il est alors préférable d'utiliser un noeud de sortie Fichier. Pour plus d'informations, reportez-vous à la section noeud Export Fichier plat dans le chapitre 7 sur p. 481.

Navigateur du nœud Table

Figure 6-7
Fenêtre du navigateur du nœud Table



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

Le navigateur du noeud Table affiche les données tabulaires et vous permet d'exécuter des opérations standard : sélection et copie de cellules, réorganisation des colonnes, enregistrement et impression du tableau. Pour plus d'informations, reportez-vous à la section Sélection de cellules et de colonnes sur p. 403.Il s'agit des mêmes opérations que vous pouvez effectuer lors de l'aperçu des données dans un noeud.

Exportation des données de table. Vous pouvez exporter des données depuis le navigateur du noeud Table en choisissant :

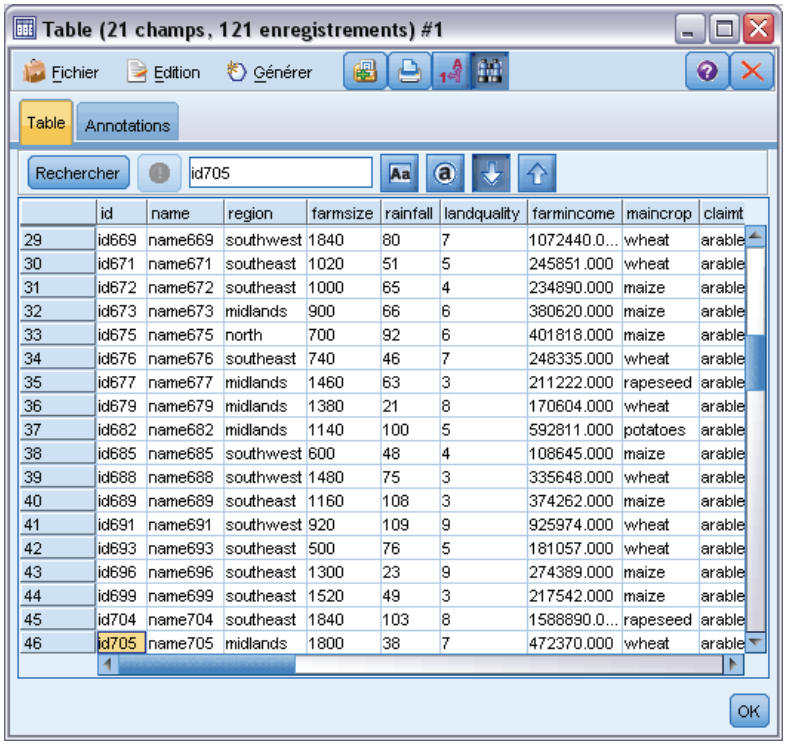
Fichier > Exporter

Pour plus d'informations, reportez-vous à la section Exportation des sorties sur p. 402.

Les données sont exportées dans le format d'encodage par défaut du système qui est spécifié dans le Panneau de configuration de Windows, ou si le système est en mode réparti, sur l'ordinateur serveur.

Recherche dans le tableau. Le bouton de recherche (icône représentant des jumelles) de la barre d'outils principale active la barre d'outils de recherche, qui permet de trouver des valeurs précises dans le tableau. Vous pouvez effectuer une recherche vers le début ou vers la fin du tableau, indiquer si la recherche doit respecter la casse ou non (bouton Aa), et interrompre une recherche en cours à l'aide du bouton Interrompre la recherche.

Figure 6-8
Tableau avec commandes de recherche activées



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
29	id669	name669	southwest	1840	80	7	1072440.0...	wheat	arable
30	id671	name671	southeast	1020	51	5	245851.000	wheat	arable
31	id672	name672	southeast	1000	65	4	234890.000	maize	arable
32	id673	name673	midlands	900	66	6	380620.000	maize	arable
33	id675	name675	north	700	92	6	401818.000	maize	arable
34	id676	name676	southeast	740	46	7	248335.000	wheat	arable
35	id677	name677	midlands	1460	63	3	211222.000	rapeseed	arable
36	id679	name679	midlands	1380	21	8	170604.000	wheat	arable
37	id682	name682	midlands	1140	100	5	592811.000	potatoes	arable
38	id685	name685	southwest	600	48	4	108645.000	maize	arable
39	id688	name688	southwest	1480	75	3	335648.000	wheat	arable
40	id689	name689	southeast	1160	108	3	374262.000	maize	arable
41	id691	name691	southwest	920	109	9	925974.000	wheat	arable
42	id693	name693	southeast	500	76	5	181057.000	wheat	arable
43	id696	name696	southeast	1300	23	9	274389.000	maize	arable
44	id699	name699	southeast	1520	49	3	217542.000	maize	arable
45	id704	name704	southeast	1840	103	8	1588890.0...	rapeseed	arable
46	id705	name705	midlands	1800	38	7	472370.000	wheat	arable

Création de noeuds. Le menu Générer contient des options permettant de générer des noeuds.

- **Noeud Sélectionner (Enregistrements).** Crée un nœud Sélectionner permettant de sélectionner les enregistrements auxquels au moins une cellule sélectionnée dans le tableau est associée.
- **Sélectionner (Et).** Crée un nœud Sélectionner permettant de sélectionner les enregistrements contenant *toutes* les valeurs sélectionnées dans le tableau.
- **Sélectionner (Ou).** Crée un nœud Sélectionner permettant de sélectionner les enregistrements contenant *n'importe quelle* valeur sélectionnée dans le tableau.
- **Calculer (Enregistrements).** Crée un nœud Calculer permettant de créer un champ booléen. Ce dernier contient la valeur *T* pour les enregistrements pour lesquels au moins une cellule du tableau est sélectionnée, et *F* (false - faux) pour les autres.
- **Calculer (Et).** Crée un nœud Calculer permettant de créer un champ booléen. Ce dernier indique la valeur *T* (true - vrai) pour les enregistrements contenant *toutes* les valeurs sélectionnées dans le tableau, et *F* (false - faux) pour les autres.
- **Calculer (Ou).** Crée un nœud Calculer permettant de créer un champ booléen. Ce dernier indique la valeur *T* (true - vrai) pour les enregistrements contenant *n'importe quelle* valeur sélectionnée dans le tableau, et *F* (false - faux) pour les autres.

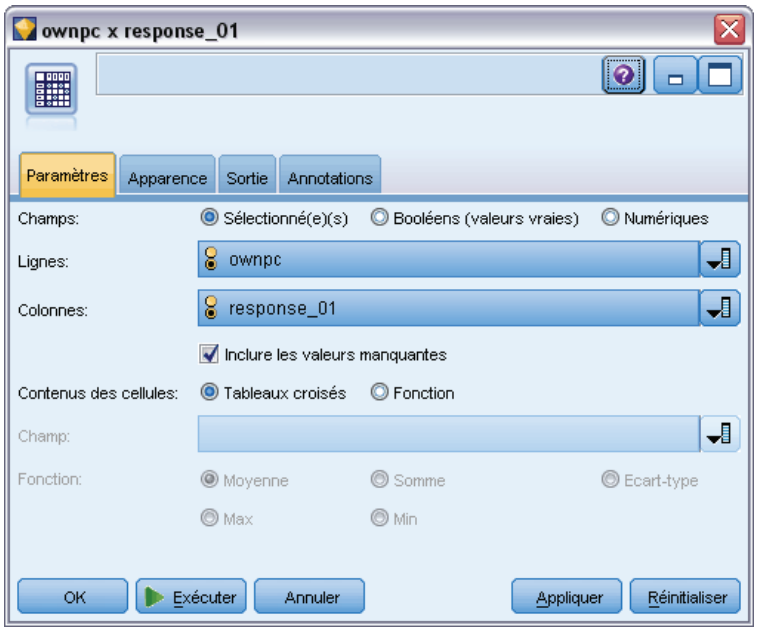
Noeud Matrice

Le nœud Matrice permet de créer un tableau dans lequel les relations entre les champs sont indiquées. Il s'agit généralement de la relation entre deux champs catégoriels (booléens, nominaux ou ordinaux), mais il peut également s'agir de la relation entre des champs continus (intervalle numérique).

Noeud Matrice - Onglet Paramètres

L'onglet Paramètres permet de définir des options pour la structure de la matrice.

Figure 6-9
Noeud Matrice : onglet Paramètres



Champs. Permet de choisir l'un des types de sélection de champ suivants :

- **Sélectionné(e)(s).** Cette option permet de sélectionner un champ catégoriel pour les lignes de la matrice et un pour les colonnes. Les lignes et les colonnes de la matrice sont définies par la liste des valeurs du champ catégoriel sélectionné. Les cellules de la matrice contiennent les statistiques récapitulatives sélectionnées plus bas.
- **Tous les booléens (valeurs vraies).** Cette option crée une matrice contenant une ligne et une colonne pour chaque champ booléen présent dans les données. Les cellules de la matrice indiquent le nombre d'enregistrements pour lesquels une combinaison de deux champs booléens est vraie. En d'autres termes, pour une ligne correspondant à *pain acheté* et une colonne correspondant à *fromage acheté*, la cellule à l'intersection de cette ligne et de cette colonne contient le nombre d'enregistrements pour lesquels *pain acheté* et *fromage acheté* sont vrais.
- **Tous les numériques.** Cette option crée une matrice contenant une ligne et une colonne pour chaque champ numérique. Les cellules de la matrice indiquent la somme des produits croisés pour la paire de champs correspondante. En d'autres termes, pour chaque cellule de la

matrice, les valeurs du champ ligne et du champ colonne sont multipliées pour chaque enregistrement, puis additionnées.

Inclure les valeurs manquantes. Inclut les valeurs manquantes utilisateur (blancs) et les valeurs manquantes système ($null$) dans la sortie des lignes et des colonnes. Par exemple, si la valeur *Non applicable* est définie comme valeur manquante utilisateur pour le champ de colonne sélectionné, une autre colonne *Non applicable* est ajoutée, comme toute autre catégorie, au tableau (en supposant que cette valeur figure réellement dans les données). Si cette option est désélectionnée, la colonne *Non applicable* est exclue, quelle que soit sa fréquence.

Remarque : L'option d'ajout de valeurs manquantes ne s'applique que lorsque les champs sélectionnés sont affichés sous forme de tableau croisé. Les valeurs vides sont mappées avec les valeurs nulles ($null$) et sont exclues de l'agrégation pour le champ de fonction lorsque vous vous trouvez en mode Sélectionné(e)(s) et que le contenu est paramétré sur Fonction, et pour tous les champs numériques lorsque le mode est paramétré sur Numériques.

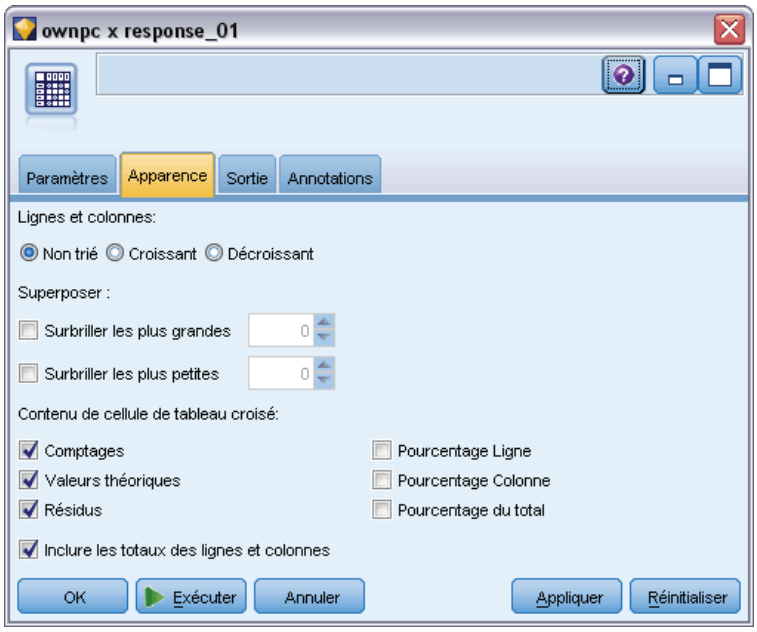
Contenus des cellules. Si vous avez choisi Sélectionné(e)(s) dans la zone Champs, vous pouvez indiquer le type de statistique à utiliser dans les cellules de la matrice. Sélectionnez des statistiques basées sur un comptage, ou un champ de superposition récapitulant les valeurs d'un champ numérique en fonction des valeurs des champs ligne et colonne.

- **Tableaux croisés.** Les valeurs des cellules indiquent le nombre et/ou le pourcentage d'enregistrements auxquels la combinaison de valeurs correspondante est associée. Vous pouvez indiquer les récapitulatifs de tableaux croisés de votre choix à l'aide des options de l'onglet Apparence. La valeur Chi-deux globale et la signification sont également affichées. Pour plus d'informations, reportez-vous à la section Navigateur de sortie du noeud Matrice sur p. 413.
- **Fonction.** Si vous sélectionnez une fonction récapitulative, les valeurs des cellules sont une fonction des valeurs de champ de superposition sélectionnées pour les observations où les valeurs de ligne et de colonne appropriées existent. Par exemple, si le champ ligne est *Région*, le champ colonne *Produit* et le champ de superposition *Revenu*, la cellule située à l'intersection de la ligne *Nord-est* et de la colonne *Objet* contiendra la somme (ou la moyenne, ou la valeur minimale ou maximale) des revenus provenant de la vente d'objets dans la région nord-est. La fonction récapitulative par défaut est Moyenne. Vous pouvez sélectionner une autre fonction pour la récapitulation du champ Fonction. Les options possibles sont les suivantes : Moyenne, Somme, Ecart-type, Maximum et Minimum.

Noeud Matrice - Onglet Apparence

L'onglet Apparence permet de définir des options de tri et de surlignage pour la matrice, ainsi que les statistiques présentées pour les matrices de tableau croisé.

Figure 6-10
Noeud Matrice : onglet Apparence



Lignes et colonnes. Permet de contrôler le tri des en-têtes de ligne et de colonne de la matrice. La valeur par défaut est Non trié. Sélectionnez Croissant ou Décroissant en fonction de l'ordre de tri des en-têtes de ligne et de colonne voulu.

Superposer. Permet de surligner les valeurs extrêmes de la matrice. Les valeurs sont surlignées sur la base du nombre de cellules (pour les matrices de tableau croisé) ou des valeurs calculées (pour les matrices de fonction).

- **Surligner le haut.** Cette option permet de surligner en rouge les valeurs les plus élevées de la matrice. Vous devez indiquer le nombre de valeurs à surligner.
- **Surligner le bas.** Cette option permet de surligner en vert les valeurs les moins élevées de la matrice. Vous devez indiquer le nombre de valeurs à surligner.

Remarque : s'il existe des valeurs ex-aequo, il est possible que le nombre de valeurs surlignées soit supérieur au nombre indiqué. Par exemple, dans le cas d'une matrice dont six cellules contiennent des zéros, si vous sélectionnez Surligner le bas 5, les six zéros seront surlignés.

Contenu de cellule de tableau croisé. Pour les tableaux croisés, vous pouvez indiquer les statistiques récapitulatives contenues dans la matrice dédiée aux matrices de tableau croisé. Ces options ne sont pas disponibles lorsque la fonction Numériques ou Fonction est sélectionnée dans l'onglet Paramètres.

- **Effectifs.** Les cellules indiquent le nombre d'enregistrements dont la valeur de ligne a la valeur de colonne correspondante. Il s'agit uniquement du contenu de cellule par défaut.

- **Effectifs théoriques :** Valeur théorique du nombre d'enregistrements dans la cellule, en supposant qu'il n'existe aucune relation entre les lignes et les colonnes. Les valeurs théoriques sont basées sur la formule suivante :

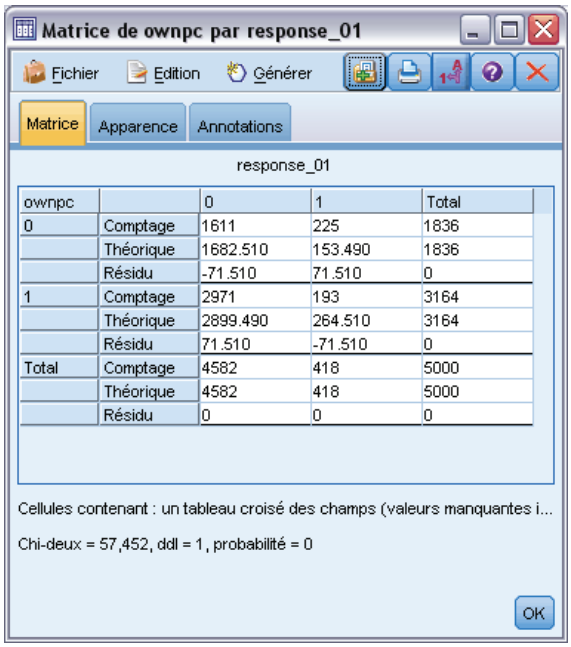
  p(*row value*) * p(*column value*) * *total number of records*

- **Résidus :** Différence entre les valeurs observées et les valeurs théoriques.
- **Pourcentage de lignes.** Pourcentage de tous les enregistrements dont la valeur de ligne a la valeur de colonne correspondante. Le pourcentage maximal pour une ligne est égal à 100.
- **Pourcentage de colonnes.** Pourcentage de tous les enregistrements dont la valeur de colonne a la valeur de ligne correspondante. Le pourcentage maximal pour une colonne est égal à 100.
- **Pourcentage du total.** Pourcentage de tous les enregistrements présentant la combinaison valeur de colonne/valeur de ligne. Le pourcentage maximal pour la matrice est égal à 100.
- **Inclure les totaux des lignes et colonnes.** Ajoute une ligne et une colonne à la matrice pour les totaux.
- **Appliquer les paramètres.** (Navigateur de sortie seulement) Vous permet de modifier l'apparence de la sortie du nœud Matrice sans avoir besoin de fermer et rouvrir le navigateur de sortie. Effectuez les modifications dans cet onglet du navigateur de sortie, cliquez sur ce bouton et sélectionnez l'onglet Matrice pour visualiser l'impact des modifications.

Navigateur de sortie du noeud Matrice

Le navigateur du nœud Matrice affiche les données sous la forme d'un tableau croisé dans lequel vous pouvez effectuer un certain nombre d'opérations : sélection de cellules, copie totale ou partielle de la matrice dans le Presse-papiers, création de noeuds en fonction d'une sélection, enregistrement et impression de la matrice. Le navigateur du noeud Matrice permet également d'afficher les sorties de certains modèles, comme les modèles Naive Bayes d'Oracle.

Figure 6-11
Navigateur du noeud Matrice



ownpc		0	1	Total
0	Comptage	1611	225	1836
	Théorique	1682.510	153.490	1836
	Résidu	-71.510	71.510	0
1	Comptage	2971	193	3164
	Théorique	2899.490	264.510	3164
	Résidu	71.510	-71.510	0
Total	Comptage	4582	418	5000
	Théorique	4582	418	5000
	Résidu	0	0	0

Les menus Fichier et Edition offrent les fonctions habituelles d'impression, d'enregistrement et d'exportation de sortie, ainsi que de sélection et de copie des données. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Khi-deux : Pour le tableau croisé de deux champs catégoriels, le Pearson du Chi-deux global apparaît également sous le tableau. Ce test indique la probabilité que les deux champs ne soient pas liés, sur la base de la différence entre les valeurs observées et les valeurs théoriques en l'absence de relation. Par exemple, s'il n'existe aucune relation entre la satisfaction client et l'emplacement des magasins, vous vous attendez à des taux de satisfaction semblables pour tous les magasins. En revanche, si certains magasins enregistrent de forts taux de satisfaction par rapport aux autres, tout laisse à penser qu'il ne s'agit pas d'une simple coïncidence. Plus la différence est importante, plus la probabilité que cela soit dû uniquement à une erreur d'échantillonnage aléatoire est faible.

- Le test du Chi-deux indique la probabilité que les deux champs ne soient pas liés, auquel cas les différences éventuelles entre les fréquences observées et les prévisions de fréquence ne relèvent que du hasard. Si cette probabilité est très faible (en général, inférieure à 5 %), la relation entre les deux champs est considérée comme significative.
- Si une seule colonne ou une seule ligne est utilisée (test du Chi-deux unilatéral), les degrés de liberté correspondent au nombre de cellules moins un. Pour un test du Chi-deux bilatéral, les degrés de liberté correspondent au nombre de lignes moins le nombre de colonnes moins un.
- Soyez vigilant en interprétant les statistiques Chi-deux lorsque l'une des prévisions de fréquence de cellule est inférieure à cinq.
- Le test du Chi-deux n'est disponible que pour le tableau croisé de deux champs. (Lorsque l'option Tous les booléens ou Numériques est sélectionnée dans l'onglet Paramètres, ce test n'apparaît pas.)

Menu Générer. Le menu Générer contient des options permettant de générer des noeuds. Ces options sont disponibles uniquement pour les matrices à tableau croisé et vous devez avoir sélectionné au moins une cellule dans la matrice.

- **Noeud Sélectionner.** Crée un nœud Sélectionner permettant de sélectionner les enregistrements correspondant à au moins une cellule sélectionnée dans la matrice.
- **Noeud Calculer (Booléen).** Crée un nœud Calculer permettant de créer un champ booléen. Ce dernier contient la valeur *T* (true - vrai) pour les enregistrements correspondant à au moins une cellule sélectionnée dans la matrice, et *F* (false - faux) pour les autres.
- **Noeud Calculer (Ensemble).** Crée un nœud Calculer permettant de créer un champ nominal. Le champ nominal contient une catégorie pour chaque ensemble contigu de cellules sélectionnées dans la matrice.

Noeud Analyse

Les noeuds Analyse vous permettent d'évaluer la capacité d'un modèle à générer des prévisions précises. Les noeuds Analyse comparent les valeurs prédites et les valeurs réelles (votre champ cible) d'un ou de plusieurs nuggets de modèle. Ils peuvent également être utilisés pour comparer des modèles prédictifs entre eux.

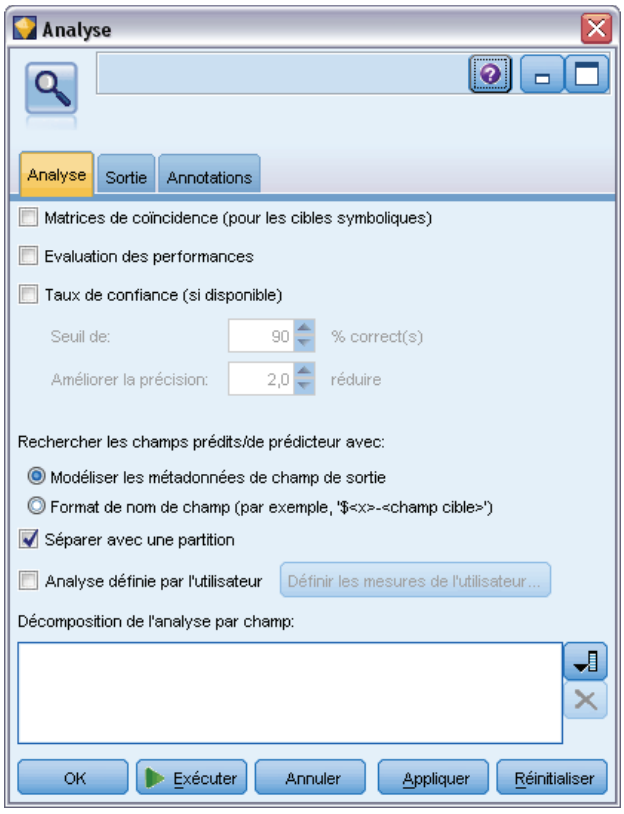
Lorsque que vous exécutez un nœud Analyse, un récapitulatif des résultats de l'analyse est automatiquement ajouté à la section Analyse de l'onglet Récapitulatif pour chaque nugget de modèle du flux exécuté. Les résultats détaillés de l'analyse apparaissent dans l'onglet Sorties de la fenêtre de gestionnaire ; ils peuvent également être écrits directement dans un fichier.

Remarque : Etant donné que les noeuds Analyse comparent les valeurs prédites aux valeurs réelles, ils ne sont utiles qu'avec les modèles supervisés (ceux qui requièrent un champ cible). Pour les modèles non supervisés, comme les algorithmes de classification non supervisée, aucun résultat réel n'est disponible pour servir de base à la comparaison.

Noeud Analyse - Onglet Analyse

L'onglet Analyse permet d'indiquer les détails de l'analyse.

Figure 6-12
Noeud Analyse : onglet Analyse



Matrices de coïncidence (pour les cibles symboliques ou catégorielles). Affiche le motif des correspondances entre chaque champ généré (prédit) et le champ cible associé pour les cibles catégorielles (booléen, nominal ou ordinal). Un tableau apparaît, dans lequel les lignes sont définies par des valeurs réelles et les colonnes par des valeurs prédites, chaque cellule indiquant le nombre d'enregistrements auxquels ce motif correspond. Cette fonction permet notamment d'identifier les erreurs systématiques dans les prévisions. Si plusieurs champs générés sont reliés au même champ de sortie alors qu'ils sont issus de modèles différents, le nombre de fois où ces champs sont en accord ou en désaccord est calculé et affiché. Lorsqu'ils sont en accord, d'autres statistiques correctes/incorrectes apparaissent.

Evaluation des performances. Affiche les statistiques d'évaluation des performances des modèles produisant des sorties catégorielles. Ces statistiques, affichées pour chaque catégorie des champs de sortie, indiquent la taille moyenne (en bits) des informations générées par le modèle utilisé pour la prévision des enregistrements appartenant à la catégorie en question. Elles tiennent compte des difficultés liées à la classification ; par conséquent, l'index d'évaluation de performances des prévisions précises portant sur des catégories rares sera supérieur à celui des prévisions précises portant sur des catégories courantes. Si le modèle ne permet pas d'obtenir des résultats pertinents pour une catégorie, l'index d'évaluation de performances de cette dernière sera de zéro.

Taux de confiance (si disponible). Pour les modèles qui génèrent un champ de confiance, cette option affiche des statistiques sur les valeurs de confiance et leurs relations avec les prévisions. Deux paramètres peuvent être définis pour cette option :

- **Seuil de.** Indique le niveau de confiance au-delà duquel la précision sera égale au pourcentage spécifié.
- **Améliorer la précision.** Indique le niveau de confiance au-delà duquel la précision sera améliorée par le facteur spécifié. Par exemple, si la précision globale est de 90 % et que cette option est paramétrée sur 2,0 la valeur affichée correspondra au niveau de confiance requis pour une précision de 95 %.

Rechercher les champs prédits/prédicteurs avec. Détermine la façon dont les champs prédits sont en correspondance avec le champ cible d'origine.

- **Modéliser les métadonnées de champ de sortie.** Fait correspondre les champs prédits à la cible en fonction des informations du champ de modèle, ce qui autorise une correspondance même si un champ prédit a été renommé. Les informations du champ de modèle peuvent aussi être accédées pour tout champ prédit à partir de la boîte de dialogue Valeurs grâce à un noeud Typer. Pour plus d'informations, reportez-vous à la section Utilisation de la boîte de dialogue Valeurs dans le chapitre 4 sur p. 144.
- **Format de nom de champ.** Fait correspondre des champs en fonction de la convention de dénomination. Par exemple, des valeurs prédites générées par un nugget de modèle C5.0 pour une cible nommée *réponse* doivent se trouver dans un champ nommé *$C-réponse*.

Séparer avec une partition. Si un champ de partition est utilisé pour diviser des enregistrements en échantillons d'apprentissage, de test et de validation, sélectionnez cette option pour afficher les résultats séparément pour chaque partition. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 sur p. 207.

Remarque : lorsque vous séparez des enregistrements par partition, ceux dont le champ de partition contient des valeurs nulles sont exclus de l'analyse. Ce problème ne se pose jamais si un noeud Partitionner est utilisé, car ce type de noeud ne génère aucune valeur nulle.

Analyse définie par l'utilisateur. Permet d'indiquer le calcul d'analyse à utiliser pour l'évaluation des modèles. Vous pouvez utiliser des expressions CLEM pour spécifier les éléments calculés pour chaque enregistrement et pour combiner les scores de niveau enregistrement en un score global. Utilisez les fonctions @TARGET et @PREDICTED pour faire référence respectivement à la valeur cible (sortie réelle) et à la valeur prédite.

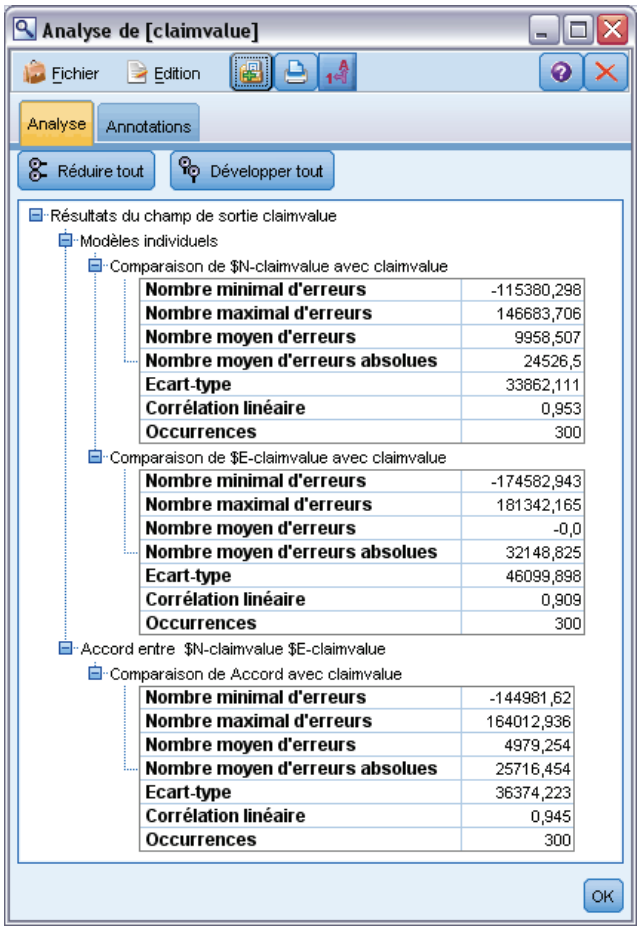
- **Si.** Indiquez une expression conditionnelle pour utiliser des calculs différents en fonction de certaines conditions.
- **Donc.** Indiquez le calcul à utiliser si la condition Si a la valeur true (vrai).
- **Sinon.** Indiquez le calcul à utiliser si la condition Si a la valeur false (faux).
- **Utiliser.** Sélectionnez les statistiques à utiliser pour calculer un score global à partir des scores individuels.

Décomposition de l'analyse par champ. Affiche les champs catégoriels disponibles pour la décomposition de l'analyse. Outre l'analyse globale, une analyse distincte sera effectuée pour chaque catégorie de chaque champ de décomposition.

Navigateur de sortie du nœud Analyse

Le navigateur de sortie du nœud Analyse permet de visualiser les résultats de l'exécution du nœud Analyse. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu Fichier. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Figure 6-13
Navigateur de sortie du nœud Analyse



Lorsque vous accédez pour la première fois à la sortie du nœud Analyse, les résultats sont développés. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement située à gauche des résultats à masquer ou cliquez sur le bouton Réduire tout pour réduire tous les résultats. Pour afficher de nouveau les résultats, utilisez la commande de développement située à gauche des résultats à afficher ou cliquez sur le bouton Développer tout pour développer tous les résultats.

Résultats du champ de sortie. La sortie du nœud Analyse contient une section pour chaque champ de sortie pour lequel il existe un champ de prévision créé par un modèle généré.

Comparaison. La section du champ de sortie contient une sous-section pour chaque champ de prévision associé au champ de sortie. Pour les champs de sortie catégoriels, la partie supérieure de cette section contient un tableau indiquant le nombre et le pourcentage de prévisions correctes et incorrectes, et le nombre total d'enregistrements dans le flux. Pour les champs de sortie numériques, cette section contient les informations suivantes :

- **Nombre minimal d'erreurs.** Affiche le nombre minimal d'erreurs (différence entre les valeurs observées et les valeurs prédites).
- **Nombre maximal d'erreurs.** Affiche le nombre maximal d'erreurs.
- **Nombre moyen d'erreurs.** Affiche le nombre moyen d'erreurs sur l'ensemble des enregistrements. Indique s'il existe un **biais** systématique (tendance à surestimer au lieu de sous-estimer ou inversement) dans le modèle.
- **Nombre moyen d'erreurs absolues.** Affiche la moyenne des valeurs absolues des erreurs sur l'ensemble des enregistrements. Indique la grandeur moyenne des erreurs, indépendamment de la direction.
- **Ecart-type.** Indique l'écart-type des erreurs.
- **Corrélation linéaire.** Indique la corrélation linéaire entre les valeurs prédites et réelles. Ces statistiques varient entre –1,0 et 1,0. Les valeurs proches de +1,0 indiquent une association positive forte, de sorte que les valeurs prédites élevées sont associées à des valeurs réelles élevées et les valeurs prédites faibles à des valeurs réelles faibles. Les valeurs proches de –1 indiquent une association négative forte, de sorte que les valeurs prédites élevées sont associées à des valeurs réelles faibles, et inversement. Les valeurs proches de 0,0 indiquent une association faible, de sorte que les valeurs prédites sont plus ou moins indépendantes des valeurs réelles. *Remarque* : Une entrée vide présente ici indique que la corrélation linéaire ne peut pas être calculée dans ce cas, car les valeurs réelles ou prédites sont constantes.
- **Occurrences.** Indique le nombre d'enregistrements utilisés dans l'analyse.

Matrice de coïncidences. Pour les champs de sortie catégoriels, si vous avez indiqué une matrice de coïncidences dans les options d'analyse, une sous-section contenant cette matrice apparaît. Les lignes représentent les valeurs réelles observées tandis que les colonnes représentent les valeurs prédites. La cellule du tableau indique le nombre d'enregistrements pour chaque combinaison valeurs prédites/valeurs réelles.

Evaluation des performances. Pour les champs de sortie catégoriels, si vous avez spécifié des statistiques d'évaluation des performances dans les options d'analyse, les résultats d'évaluation des performances apparaissent ici. Chaque catégorie de sortie est répertoriée avec les statistiques d'évaluation des performances correspondantes.

Rapport de valeurs de confiance. Pour les champs de sortie catégoriels, si vous avez spécifié des valeurs de confiance dans les options d'analyse, ces valeurs apparaissent ici. Les statistiques suivantes sont indiquées pour les valeurs de confiance du modèle :

- **Intervalle :** Indique l'intervalle (valeurs les plus faibles et les plus élevées) des valeurs de confiance pour les enregistrements des données du flux.
- **Moyenne correcte.** Indique la valeur de confiance moyenne des enregistrements correctement classés.
- **Moyenne incorrecte.** Indique la valeur de confiance moyenne des enregistrements non correctement classés.

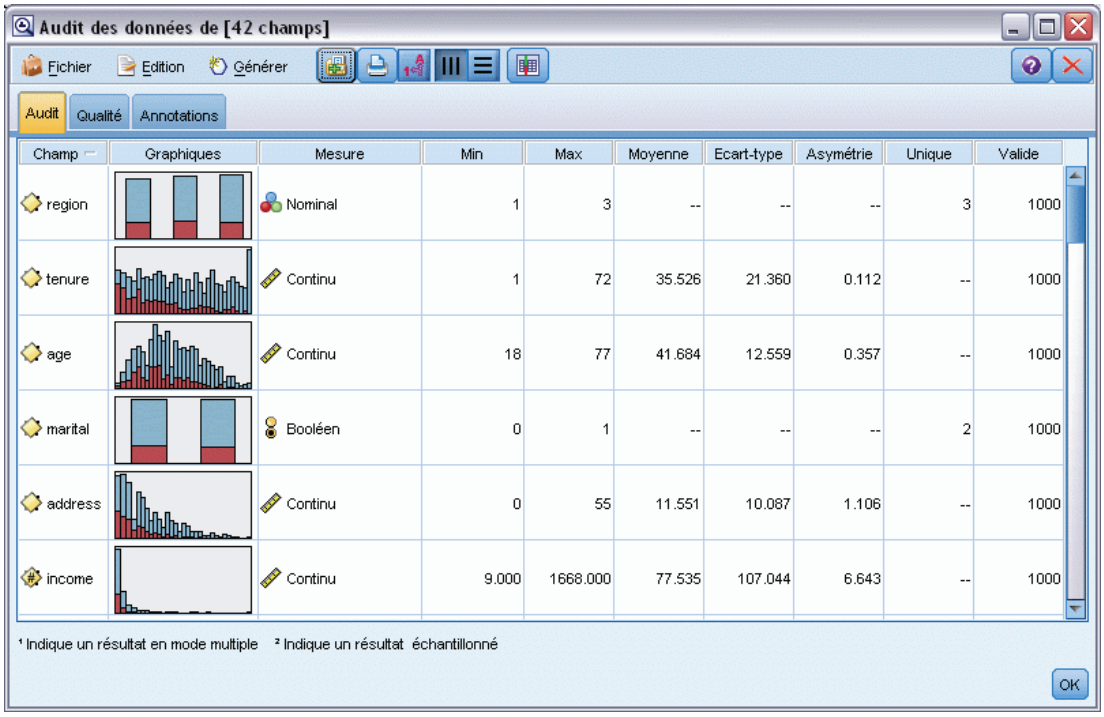
- **Toujours correct au-dessus de.** Indique le seuil de confiance au-dessus duquel les prévisions sont toujours correctes et le pourcentage d'observations qui répondent à ce critère.
- **Toujours incorrect au-dessous de.** Indique le seuil de confiance en dessous duquel les prévisions sont toujours fausses et le pourcentage d'observations qui répondent à ce critère.
- **Degré de précision au-dessus de = X %.** Indique le niveau de confiance correspondant à une précision de *X %*. *X* est approximativement la valeur spécifiée pour Seuil de dans les options d'analyse. Pour certains modèles et ensembles de données, il n'est pas possible de choisir une valeur de confiance qui indique le seuil exact spécifié dans les options (généralement en raison de classes d'observations similaires ayant la même valeur de confiance à proximité du seuil). La valeur de seuil indiquée est la valeur la plus proche du critère de précision spécifié pouvant être obtenue avec un même seuil de valeur de confiance.
- **Réduction correcte au-dessus = X.** Indique la valeur de confiance au niveau de laquelle la précision est *X* fois meilleure qu'elle ne l'est pour l'ensemble de données global. *X* est la valeur spécifiée dans le champ Améliorer la précision des options d'analyse.

Accord entre. Si plusieurs modèles générés prédisant le même champ de sortie sont inclus dans le flux, vous pouvez également consulter des statistiques sur l'**accord** entre les prévisions générées par les modèles. Il peut s'agir du nombre et du pourcentage d'enregistrements pour lesquels les prévisions sont concordantes (pour les champs de sortie catégoriels), ou de statistiques récapitulant les erreurs (pour les champs de sortie continus). Pour les champs catégoriels, une analyse comparant les prévisions aux valeurs réelles est incluse pour le sous-ensemble d'enregistrements sur lesquels les modèles sont concordants (c'est-à-dire, génèrent la même valeur prédite).

Noeud Audit données

Le noeud Audit données offre un premier aperçu complet des données que vous entrez dans IBM® SPSS® Modeler. Celles-ci sont présentées sous la forme d'une matrice très lisible que vous pouvez trier et à partir de laquelle vous pouvez générer des graphiques grandeur nature et divers noeuds de préparation des données.

Figure 6-14
Navigateur Audit données

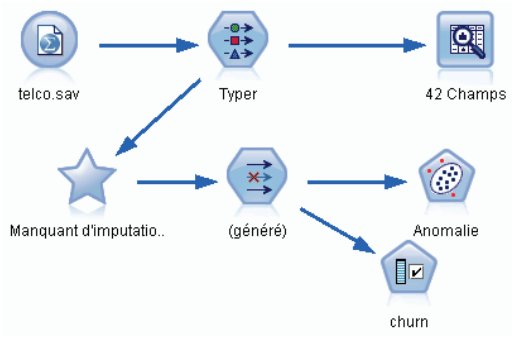


- L'onglet Audit affiche un rapport qui fournit des statistiques récapitulatives, des histogrammes et des graphiques Proportion qui peuvent contribuer à une première compréhension des données. Le rapport affiche aussi l'icône de stockage devant le nom de champ.
- L'onglet Qualité du rapport d'audit affiche des informations sur les valeurs éloignées, les extrêmes et les valeurs manquantes, et propose des outils de gestion de ces valeurs.

Utilisation du nœud Audit données

Le nœud Audit données peut être connecté directement à un nœud source ou en aval d'un nœud Typer instancié. Vous pouvez également générer des noeuds de préparation des données sur la base des résultats. Par exemple, vous pouvez générer un noeud Filtrer qui exclut les champs contenant trop de valeurs manquantes pour être utiles à la modélisation et générer un super noeud qui attribue les valeurs manquantes à l'un des champs ou à tous les champs restants. Voilà où la puissance réelle de l'audit intervient, vous permettant non seulement d'évaluer l'état actuel de vos données, mais également d'agir sur la base de cette évaluation.

Figure 6-15
Flux avec super noeud Valeurs manquantes

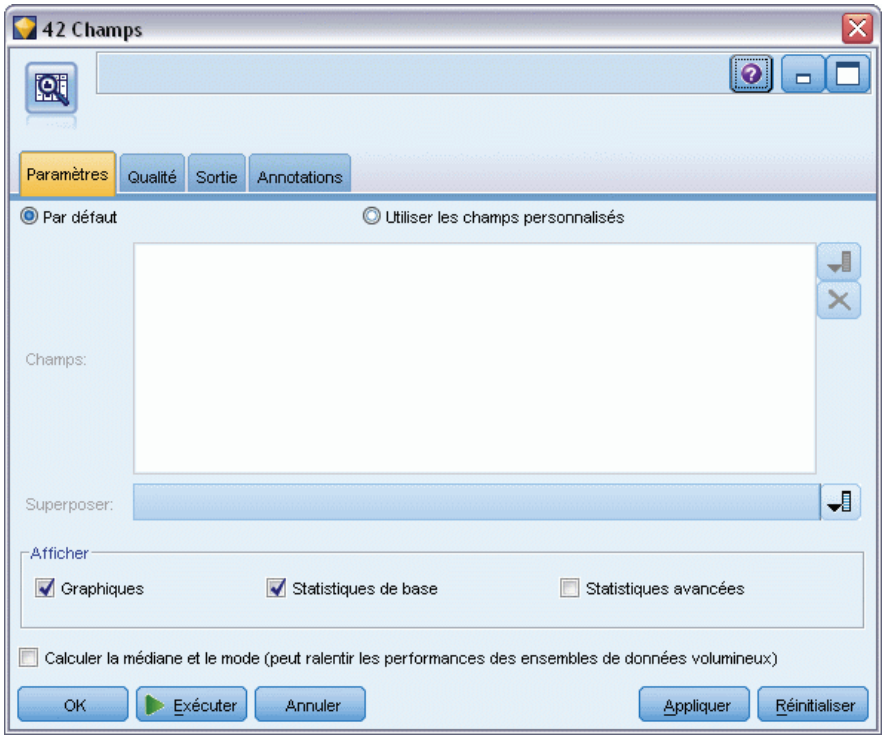


Filtrage ou échantillonnage des données. Etant donné que les audits initiaux sont particulièrement efficaces pour traiter les « données volumineuses », vous pouvez utiliser un noeud Echantillonner pour réduire le temps de traitement lors de l'exploration initiale en sélectionnant uniquement un sous-ensemble d'enregistrements. Vous pouvez également utiliser le noeud Audit données avec d'autres noeuds, tels que Sélection de fonction et Détection des anomalies, lors des phases exploratoires de l'analyse.

Noeud Audit données - Onglet Paramètres

L'onglet Paramètres vous permet de définir les paramètres de base de l'audit.

Figure 6-16
Noeud Audit données : onglet Paramètres



Défaut : Vous pouvez tout simplement connecter le noeud au flux et cliquer sur Exécuter pour générer un rapport d'audit pour tous les champs, sur la base des paramètres par défaut, comme suit :

- Si aucun paramètre n'a été défini pour le nœud Typer, tous les champs sont inclus dans le rapport.
- Si des paramètres ont été définis pour le noeud Typer (qu'ils soient instanciés ou non), tous les champs *Entrée*,*Cible* et *Les deux* sont inclus dans l'affichage. S'il existe un seul champ *Cible*, utilisez-le en tant que champ de superposition. Si plusieurs champs *Cible* ont été définis, aucune superposition par défaut n'est spécifiée.

Utiliser les champs personnalisés. Choisissez cette option pour sélectionner les champs manuellement. Utilisez le sélecteur de champs à droite pour sélectionner les champs un par un ou par type.

Champ de superposition. Le champ de superposition permet de tracer les graphiques en miniature affichés dans le rapport d'audit. Pour un champ continu (intervalle numérique), les statistiques à deux dimensions (covariance et corrélation) sont également calculées. Si un seul champ *Cible* est présent sur la base des paramètres de noeud Typer, ce champ est utilisé comme champ de superposition par défaut, conformément à la description précédente. Vous pouvez également sélectionner Utiliser les champs personnalisés pour définir une superposition.

Afficher : Permet d'indiquer si des graphiques sont disponibles dans la sortie et de choisir les statistiques affichées par défaut.

- **Graphiques.** Affiche un graphique pour chaque champ sélectionné : un graphique Proportion (en barres), un histogramme ou un diagramme de dispersion, selon le type de graphique adapté aux données. Les graphiques sont affichés sous forme de miniatures dans le rapport initial, mais vous pouvez également générer des graphiques en grandeur nature et des noeuds Graphiques. Pour plus d'informations, reportez-vous à la section Navigateur de sortie du nœud Audit données sur p. 425.
- **Statistiques avancées/de base.** Indique le niveau de statistiques affiché par défaut dans la sortie. Bien que ce paramètre détermine l'affichage initial, toutes les statistiques sont disponibles dans la sortie, quel que soit ce paramètre. Pour plus d'informations, reportez-vous à la section Afficher les statistiques sur p. 428.

Médiane et mode. Calcule la médiane et le mode de tous les champs figurant dans le rapport. Notez que, lorsque les ensembles de données sont volumineux, ces statistiques peuvent augmenter le temps de traitement, car leur calcul dure plus longtemps. Pour le calcul de la médiane uniquement et dans certaines conditions, vous pouvez baser la valeur figurant dans le rapport sur un échantillon de 2000 enregistrements (plutôt que sur l'ensemble de données complet). Cet échantillonnage est effectué sur la base de chaque champ lorsque les limites de mémoire risquent d'être dépassées. Lorsque l'échantillonnage est actif, les résultats sont étiquetés comme tels dans la sortie (*Médiane de l'échantillon* plutôt que *Médiane*). Toutes les statistiques autres que la médiane sont systématiquement calculées sur la base de l'ensemble de données complet.

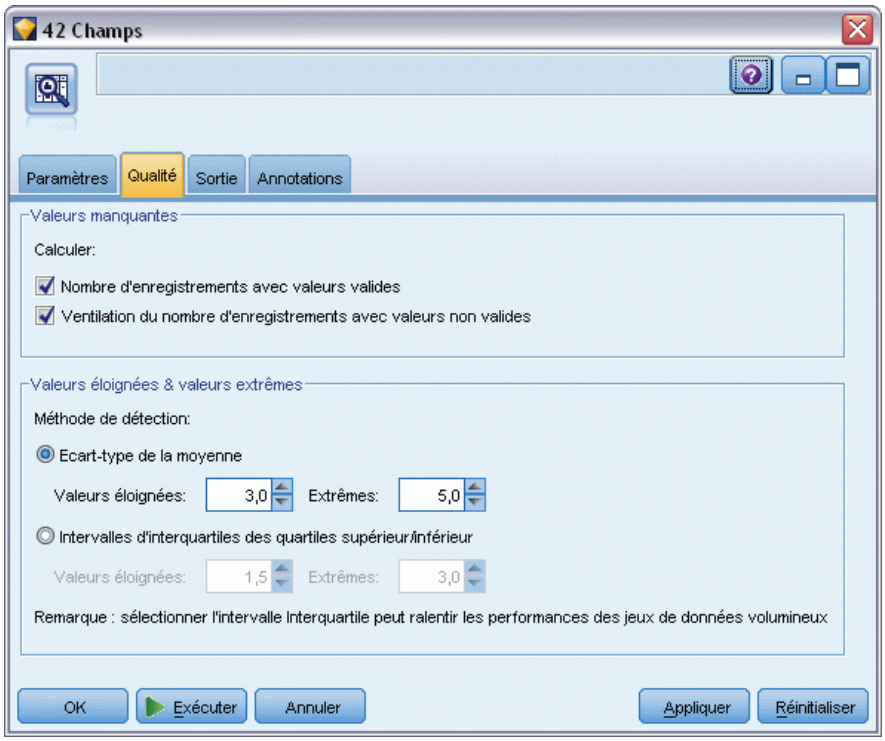
Champs vides ou sans type. Lorsqu'ils sont utilisés avec des données instanciées, les champs sans type ne sont pas inclus dans le rapport d'audit. Pour inclure des champs sans type (y compris les champs vides), sélectionnez Effacer toutes les valeurs dans les noeuds Typer en amont. Cela garantit que les données ne sont pas instanciées, ce qui entraîne l'inclusion de tous les champs dans le rapport. Par exemple, cela peut se révéler utile si vous souhaitez obtenir la liste complète de

tous les champs ou générer un noeud Filtrer qui exclut les champs vides. Pour plus d'informations, reportez-vous à la section Filtrage de champs contenant des données manquantes sur p. 433.

Audit données - Onglet Qualité

L'onglet Qualité du noeud Audit données fournit des options de traitement des valeurs manquantes, des valeurs éloignées et des extrêmes.

Figure 6-17
Noeud Audit données - Onglet Qualité



Valeurs manquantes

- **Nombre d'enregistrements avec valeurs valides.** Sélectionnez cette option pour afficher le nombre d'enregistrements contenant des valeurs valides pour chaque champ évalué. Notez que les valeurs nulles (non définies), les valeurs non renseignées, les espaces blancs et les chaînes vides sont toujours traités comme des valeurs non valides.
- **Ventilation du nombre d'enregistrements avec valeurs non valides.** Sélectionnez cette option pour afficher le nombre d'enregistrements contenant chaque type de valeur non valide pour chaque champ.

Valeurs éloignées et extrêmes

Méthode de détection des valeurs éloignées et extrêmes. Deux méthodes sont prises en charge :

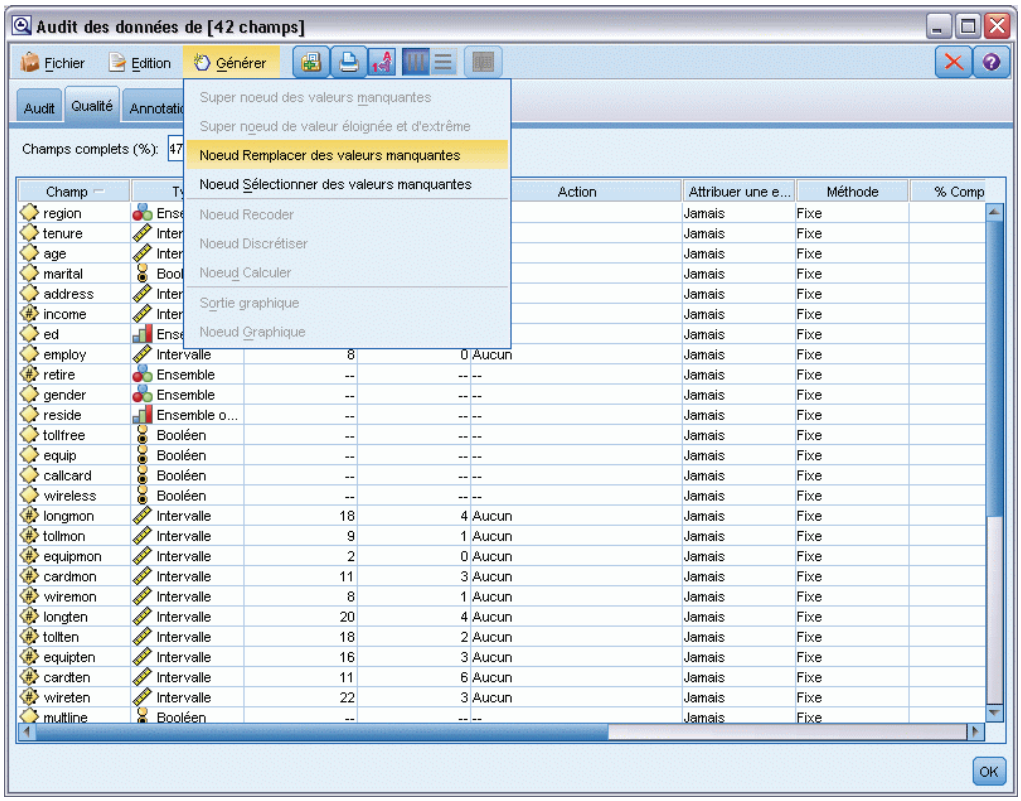
Ecart-type de la moyenne. Détecte les valeurs éloignées et extrêmes sur la base du nombre d'écarts-types par rapport à la moyenne. Par exemple, si un champ a une moyenne égale à 100 et un écart-type standard égal à 10, vous pouvez saisir 3,0 pour indiquer que toute valeur inférieure à 70 ou supérieure à 130 doit être traitée comme une valeur éloignée.

Intervalle interquartile. Détecte les valeurs éloignées et les extrêmes en fonction de l'intervalle interquartile (IQR), lequel représente l'intervalle dans lequel sont compris les deux quartiles centraux (entre les 25e et 75e centiles). Par exemple, si le paramètre par défaut est égal à 1,5, le seuil inférieur des valeurs éloignées est Q1 – 1,5 * IQR et le seuil supérieur, Q3 + 1,5*IQR. Cette option risque de ralentir les performances pour les ensembles de données volumineux.

Navigateur de sortie du nœud Audit données

Le navigateur Audit données est un outil puissant permettant d'obtenir une présentation de vos données. L'onglet Audit affiche les graphiques en miniature, des icônes de stockage et les statistiques pour tous les champs, alors que l'onglet Qualité contient des informations sur les valeurs éloignées, les extrêmes et les valeurs manquantes. Sur la base des statistiques récapitulatives et des graphiques initiaux, vous pouvez décider de recoder un champ numérique, de calculer un nouveau champ ou de reclasser les valeurs d'un champ nominal. Si vous le souhaitez, vous pouvez également procéder à une exploration plus approfondie à l'aide d'outils de visualisation avancés. Vous pouvez le faire directement à partir du navigateur de rapport d'audit via le menu Générer pour créer plusieurs noeuds permettant de transformer ou de visualiser vos données.

Figure 6-18
Génération d'un noeud Filtrer les valeurs manquantes

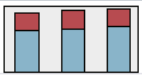
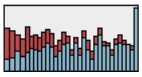


- Triez les colonnes en cliquant sur l'en-tête de colonne ou réorganisez-les en les faisant glisser. La plupart des opérations de sortie standard sont également prises en charge. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.
- Afficher les valeurs et les intervalles des champs en double-cliquant sur un champ de la colonne Mesure ou Unique.
- Utilisez la barre d'outils ou le menu Edition pour afficher ou masquer les étiquettes de valeur, ou pour sélectionner les statistiques à afficher. Pour plus d'informations, reportez-vous à la section Afficher les statistiques sur p. 428.
- Vérifiez les icônes de stockage à gauche des noms de champ. Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Pour plus d'informations, reportez-vous à la section Définition du stockage et du formatage des champs dans le chapitre 2 sur p. 32.

Affichage et génération de graphiques

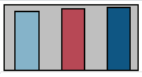
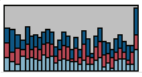
Si aucune superposition n'est sélectionnée, l'onglet Audit affiche des diagrammes en barres (pour les champs nominaux ou booléens) ou des histogrammes (pour les champs de type Continu).

Figure 6-19
Extrait de résultats d'audit sans champ de superposition

Champ	Graphiques	Type	Min	Max	Moyenne	Ecart-type	Asymétrie	Unique	Valide
region		Ensemble	1	3	--	--	--	3	1000
tenure		Intervalle	1	72	35.526	21.360	0.112	--	1000

Dans le cas de la superposition d'un champ nominal ou booléen, les valeurs de la superposition déterminent les couleurs des graphiques.

Figure 6-20
Extrait de résultats d'audit avec superposition d'un champ nominal

Champ	Graphiques	Type	Min	Max	Moyenne	Ecart-type	Asymétrie	Unique	Valide
region		Ensemble	1	3	--	--	--	3	1000
tenure		Intervalle	1	72	35.526	21.360	0.112	--	1000

Dans le cas de la superposition d'un champ continu, des diagrammes de dispersion en deux dimensions sont générés à la place des diagrammes en barres et des histogrammes unidimensionnels. Dans ce cas, l'axe *x* est mappé sur le champ de superposition, ce qui vous permet d'obtenir un tableau où tous les axes *x* sont à la même échelle.

Figure 6-21
Extrait de résultats d'audit avec superposition d'un champ continu

Champ	Graphiques	Type	Min	Max	Moyenne	Corrélation	Corrélation T	ddl. corrélation T
region		Ensemble	1	3	--	--	--	--
tenure		Intervalle	1	72	35.526	0.490	17.768	998.000

- Pour les champs de type Booléen ou Nominal, positionnez le curseur de la souris sur une barre pour afficher la valeur ou l'étiquette sous-jacente dans une info-bulle.
- Pour les champs de type Booléen ou Nominal, utilisez la barre d'outils pour rendre verticale l'orientation horizontale des graphiques en miniature.
- Pour générer un graphique en grandeur nature à partir d'une miniature, double-cliquez sur cette dernière, puis sélectionnez Sortie graphique dans le menu Générer. *Remarque* : Lorsqu'un graphique en miniature repose sur des données échantillonnées, le graphique généré contient toutes les observations si le flux de données d'origine est resté ouvert.

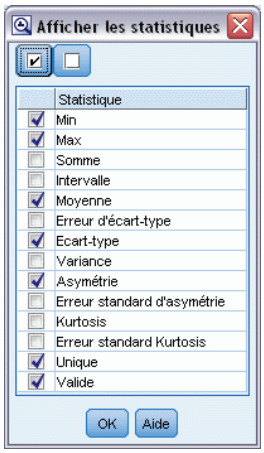
  Vous ne pouvez générer un diagramme que si le noeud Audit données qui a généré la sortie est connecté au flux.

- Pour générer un noeud Graphiques correspondant, sélectionnez un ou plusieurs champs dans l'onglet Audit, puis choisissez Noeud Graphique dans le menu Générer. Le noeud obtenu est ajouté à l'espace de travail de flux ; il permet de recréer le graphique lorsque le flux est exécuté.
- si un champ d'ensemble de superposition contient plus de 100 valeurs, un avertissement est généré et la superposition n'est pas incluse.

Afficher les statistiques

La boîte de dialogue Afficher les statistiques vous permet de sélectionner les statistiques affichées dans l'onglet Audit. Les paramètres initiaux sont indiqués dans le noeud Audit données. Pour plus d'informations, reportez-vous à la section Noeud Audit données - Onglet Paramètres sur p. 422.

Figure 6-22
Afficher les statistiques



Minimum. Valeur la plus petite d'une variable numérique.

Maximum. Plus grande valeur d'une variable numérique.

Somme. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Intervalle. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.

Erreur standard de la moyenne. Mesure du degré de variation de la moyenne d'un échantillon à l'autre au sein d'une même distribution. Cette mesure permet de comparer approximativement la moyenne observée avec une valeur hypothétique (autrement dit, vous pouvez conclure que ces deux valeurs sont différentes si le rapport de la différence avec l'erreur standard est inférieur à -2 ou supérieur à +2).

écart-type. Mesure de la dispersion des valeurs autour de la moyenne, égale à la racine carrée de la variance. L'écart-type est mesuré dans les mêmes unités que la variable d'origine.

Variance. Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Asymétrie. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Erreur standard du Skewness. Rapport de l'asymétrie avec son erreur standard, qui peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'asymétrie positive importante indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

Aplatissement. Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.

Erreur standard du Kurtosis. Le rapport de l'aplatissement à son erreur standard peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'aplatissement positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur d'aplatissement négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Unique. Evalue tous les effets simultanément, en ajustant chaque effet à tous les autres effets d'un type donné.

Valide. Observations valides, c'est-à-dire ne comportant ni la valeur manquante par défaut ni des valeurs définies comme manquantes.

Médiane. Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

Mode. Valeur qui revient le plus fréquemment. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.

La médiane et le mode sont supprimés par défaut pour améliorer les performances, mais peuvent être sélectionnés dans l'onglet Paramètres du noeud Audit données. Pour plus d'informations, reportez-vous à la section Noeud Audit données - Onglet Paramètres sur p. 422.

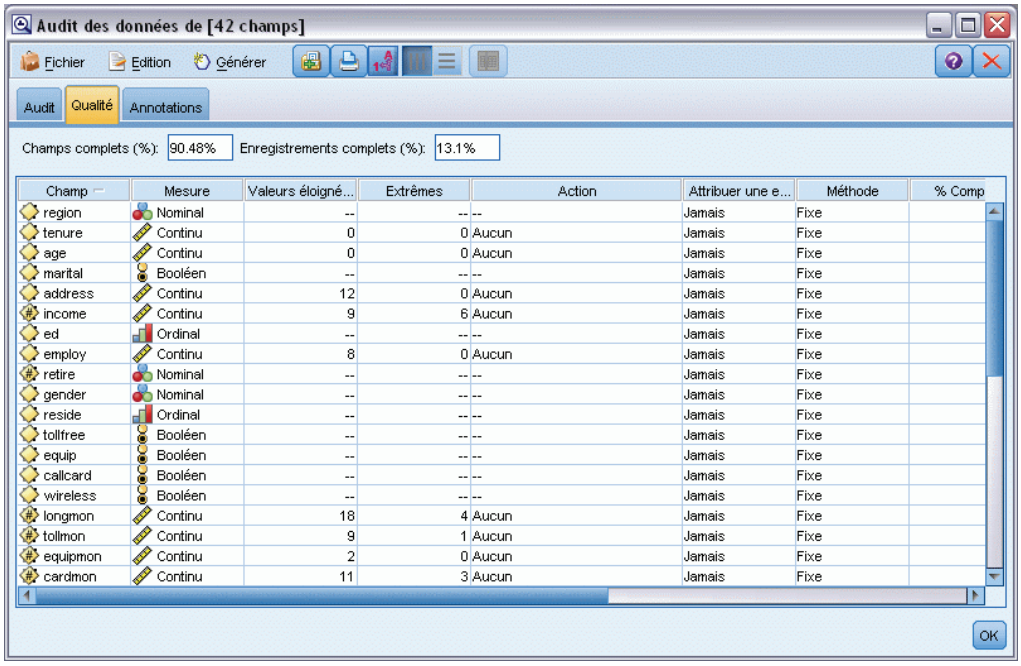
Statistiques de superpositions

Si un champ de superposition continu (intervalle numérique) est utilisé, les statistiques suivantes sont également disponibles :

Covariance. Mesure non normalisée de la relation entre deux variables, égale au produit des écarts divisé par N-1.

Navigateur Audit données - Onglet Qualité

Figure 6-23
Rapport sur la qualité du navigateur Audit données



Champ	Mesure	Valeurs éloigné...	Extrêmes	Action	Attribuer une e...	Méthode	% Comp
region	Nominal	--	--	--	Jamais	Fixe	
tenure	Continu	0	0	Aucun	Jamais	Fixe	
age	Continu	0	0	Aucun	Jamais	Fixe	
marital	Booléen	--	--	--	Jamais	Fixe	
address	Continu	12	0	Aucun	Jamais	Fixe	
income	Continu	9	6	Aucun	Jamais	Fixe	
ed	Ordinal	--	--	--	Jamais	Fixe	
employ	Continu	8	0	Aucun	Jamais	Fixe	
retire	Nominal	--	--	--	Jamais	Fixe	
gender	Nominal	--	--	--	Jamais	Fixe	
reside	Ordinal	--	--	--	Jamais	Fixe	
tollfree	Booléen	--	--	--	Jamais	Fixe	
equip	Booléen	--	--	--	Jamais	Fixe	
callcard	Booléen	--	--	--	Jamais	Fixe	
wireless	Booléen	--	--	--	Jamais	Fixe	
longmon	Continu	18	4	Aucun	Jamais	Fixe	
tollmon	Continu	9	1	Aucun	Jamais	Fixe	
equipmon	Continu	2	0	Aucun	Jamais	Fixe	
cardmon	Continu	11	3	Aucun	Jamais	Fixe	

L'onglet Qualité du navigateur Audit données affiche les résultats de l'analyse de la qualité des données. En outre, il vous permet de spécifier des traitements pour les valeurs éloignées, les extrêmes et les valeurs manquantes.

Attribution des valeurs manquantes

Le rapport d'audit répertorie le pourcentage d'enregistrements complets pour chaque champ, ainsi que le nombre de valeurs valides, de valeurs nulles et de valeurs non renseignées. Vous pouvez choisir d'attribuer les valeurs manquantes appropriées à des champs spécifiques, puis de générer un super noeud pour appliquer ces transformations.

- Dans la colonne Attribuer une entrée manquante, spécifiez le type de valeur à attribuer, le cas échéant. Vous pouvez choisir d'attribuer des valeurs non renseignées ou nulles, ou les deux, ou d'indiquer une condition ou une expression personnalisée sélectionnant les valeurs à attribuer.

Plusieurs types de valeur manquante sont reconnus par IBM® SPSS® Modeler :

- **Valeurs nulles ou manquantes système.** Ces valeurs sont des valeurs « non-chaîne » qui ne sont pas renseignées dans la base de données ou dans le fichier source, et qui n'ont pas été spécifiquement définies comme « manquantes » dans un noeud source ou un noeud Typer. Les valeurs manquantes système sont affichées sous la forme $null$. Les chaînes vides ne sont pas considérées comme des valeurs nulles dans SPSS Modeler, même si elles peuvent être traitées comme telles par certaines bases de données.

- **Chaînes vides et espaces blancs.** Les chaînes vides et les espaces blancs (chaînes sans caractère visible) sont traités différemment des valeurs nulles. Dans la plupart des cas, les chaînes vides sont considérées comme des espaces blancs. Par exemple, si vous choisissez de traiter les espaces blancs comme blancs dans un noeud source ou un noeud Typer, ce paramètre s'applique également aux chaînes vides.
- **Valeurs manquantes définies par l'utilisateur ou vides.** Ces valeurs sont des valeurs, telles que inconnu, 99, ou –1, qui sont explicitement définies comme manquantes dans un noeud source ou Typer. Vous pouvez également, si vous le souhaitez, préciser si les valeurs nulles et les espaces blancs doivent être traités comme des blancs ; un traitement spécial leur est alors appliqué et ils sont exclus de la plupart des calculs. Par exemple, vous pouvez utiliser la fonction @BLANK pour traiter comme des blancs ces valeurs, ainsi que d'autres types de valeur manquante. Pour plus d'informations, reportez-vous à la section Utilisation de la boîte de dialogue Valeurs dans le chapitre 4 sur p. 144.

▶ Dans la colonne Méthode, spécifiez la méthode à utiliser.

Les méthodes suivantes sont disponibles pour attribuer des valeurs manquantes :

Fixe. Remplacement par une valeur fixe (soit la moyenne du champ, soit la moitié de l'intervalle, soit une constante que vous indiquez).

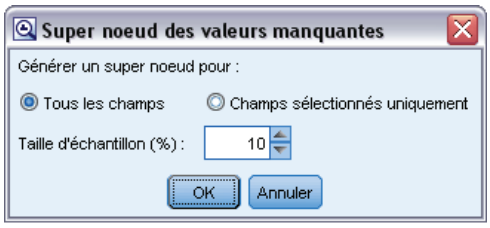
Aléatoire. Remplacement par une valeur aléatoire fondée sur une loi normale ou uniforme.

Expression. Permet d'indiquer une expression personnalisée. Par exemple, vous pourriez remplacer les valeurs par une variable globale créée par le noeud V. globales.

Algorithme. Remplacement par une valeur prévue par un modèle fondé sur l'algorithme C&RT. Chaque champ auquel une valeur est attribuée à l'aide de cette méthode est associé à un modèle C&RT distinct et à un noeud Remplacer qui remplace les valeurs non renseignées et les valeurs nulles par la valeur prédite par le modèle. Ensuite, un noeud Filtrer est utilisé pour supprimer les champs de prévision générés par le modèle.

▶ Pour générer un super noeud Valeurs manquantes, choisissez dans le menu les options suivantes :
Générer > Super noeud des valeurs manquantes

Figure 6-24
Boîte de dialogue Super noeud des valeurs manquantes

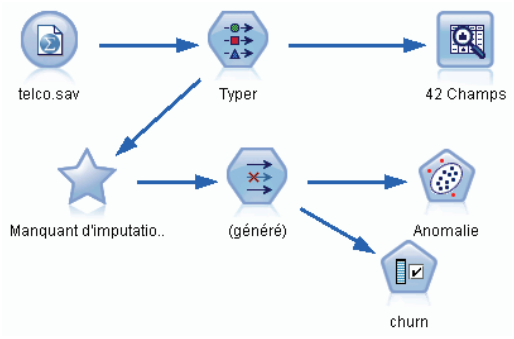


▶ Sélectionnez Tous les champs ou Champs sélectionnés uniquement, puis indiquez une taille d'échantillon si vous le souhaitez. (L'échantillon spécifié est un pourcentage. Par défaut, 10 % des enregistrements sont échantillonnés.)

▶ Cliquez sur OK pour ajouter le super noeud généré à l'espace de travail de flux.

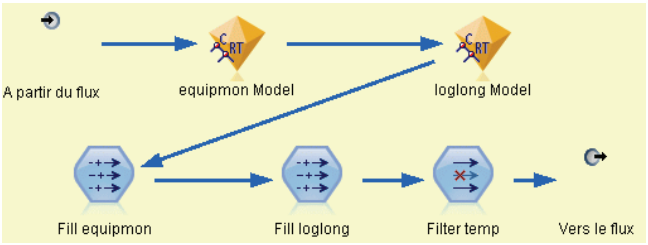
▶ Reliez le super noeud au flux pour appliquer les transformations.

Figure 6-25
Ajout du super noeud au flux



Dans le super noeud, une combinaison de noeuds Remplacer, Filtrer et de nugget de modèle est utilisée. Pour comprendre le fonctionnement du super noeud, vous pouvez l'éditer et cliquer sur Zoom avant, puis ajouter, éditer ou supprimer des noeuds spécifiques dans le super noeud pour en affiner le comportement.

Figure 6-26
Zoom avant sur le super noeud



Gestion des valeurs éloignées et extrêmes

Le rapport d'audit répertorie le nombre de valeurs éloignées et d'extrêmes pour chaque champ en fonction des options de détection spécifiées dans le noeud Audit données. Pour plus d'informations, reportez-vous à la section Audit données - Onglet Qualité sur p. 424.Vous pouvez choisir de forcer, d'isoler ou de rendre nulles ces valeurs pour des champs spécifiques, selon vos besoins, puis de générer un super noeud pour appliquer les transformations.

- Dans la colonne Action, spécifiez la gestion des valeurs éloignées et des extrêmes pour des champs spécifiques, si nécessaire.

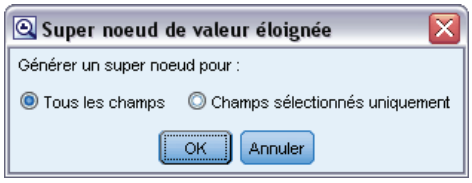
Les actions suivantes sont disponibles pour la gestion des valeurs éloignées et des extrêmes :

- **Forcer.** Remplace les valeurs éloignées et extrêmes par la valeur la plus proche qui ne sera pas considérée comme extrême. Par exemple, si une valeur éloignée est définie comme étant supérieure ou inférieure à trois écarts-types, toutes les valeurs éloignées sont remplacées par la valeur supérieure ou inférieure comprise dans cet intervalle.
- **Supprimer.** Ignore les enregistrements contenant des valeurs éloignées ou extrêmes pour le champ spécifié.

- **Rendre nul.** Remplace les valeurs éloignées et les extrêmes par la valeur nulle ou manquante système.
- **Forcer les valeurs éloignées/ignorer les extrêmes.** Ignore les valeurs extrêmes uniquement.
- **Forcer les valeurs éloignées/rendre nulles les extrêmes.** Rend nulles les valeurs extrêmes uniquement.

▶ Pour générer le super noeud, choisissez les options suivantes à partir des menus :
Générer > Super noeud de valeur éloignée et d'extrême

Figure 6-27
Boîte de dialogue Super noeud de valeur éloignée



▶ Sélectionnez Tous les champs ou Champs sélectionnés uniquement, puis cliquez sur OK pour ajouter le super noeud généré à l'espace de travail de flux.

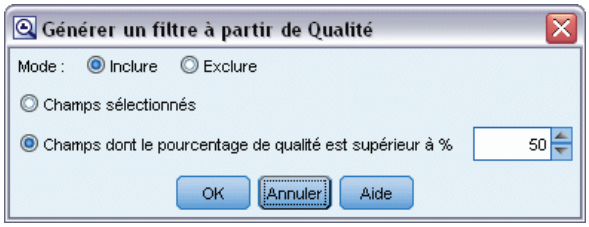
▶ Reliez le super noeud au flux pour appliquer les transformations.

Si nécessaire, vous pouvez éditer le super noeud et effectuer un zoom avant à des fins de navigation ou de modification. Dans le super noeud, les valeurs sont supprimées, forcées ou rendues nulles par le biais des noeuds Sélectionner et/ou Remplacer appropriés.

Filtrage de champs contenant des données manquantes

A partir du navigateur Data Audit, vous pouvez créer un noeud Filtrer sur la base des résultats de l'analyse de la qualité.

Figure 6-28
Boîte de dialogue Générer un filtre à partir de Qualité



Mode. Sélectionnez l'option souhaitée pour les champs indiqués : Enlever ou Isoler.

- **Champs sélectionnés.** Le noeud Filtrer inclut/exclut les champs sélectionnés dans l'onglet Qualité. Par exemple, vous pouvez trier le tableau en fonction de la colonne % terminé(s), maintenir la touche Maj enfoncée tout en cliquant sur les champs les moins complets pour les sélectionner, puis générer un noeud Filtrer excluant ces champs.
- **Champs dont le pourcentage de qualité est supérieur à.** Le nœud Filtrer inclut/exclut les champs dont le pourcentage d'enregistrements complets est supérieur au seuil indiqué. La valeur de seuil par défaut est 50 %.

Filtrage de champs vides ou sans type

Une fois les valeurs de données instanciées, les champs vides ou sans type sont exclus des résultats d'audit et de la plupart des autres résultats dans IBM® SPSS® Modeler. Ces champs sont ignorés à des fins de modélisation, mais peuvent amplifier ou encombrer les données. Dans ce cas, vous pouvez utiliser le navigateur Audit données pour générer un noeud Filtrer supprimant ces champs du flux.

- Pour vérifier que tous les champs figurent dans l'audit, y compris les champs vides ou sans type, cliquez sur Effacer toutes les valeurs dans le noeud Typer ou source en amont, ou définissez Valeurs sur <*Transférer*> pour tous les champs.

- Dans le navigateur Audit données, triez le tableau en fonction de la colonne % terminé(s), sélectionnez les champs ne contenant aucune valeur valide (ou un autre seuil) et utilisez le menu Générer pour générer un noeud Filtrer pouvant être ajouté au flux.

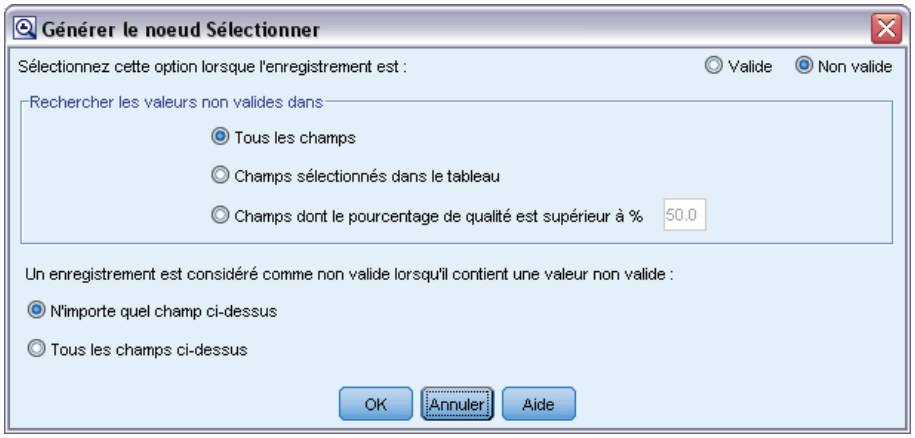
Sélection d'enregistrements contenant des valeurs manquantes

A partir du navigateur Audit données, vous pouvez créer un noeud Sélectionner sur la base des résultats de l'analyse de la qualité.

- Dans le navigateur Audit données, sélectionnez l'onglet Qualité.

- A partir du menu, sélectionnez :
  Générer > Noeud Sélectionner les valeurs manquantes

Figure 6-29
Boîte de dialogue Générer le nœud Sélectionner



Sélectionnez cette option lorsque l'enregistrement est. Indiquez si les enregistrements doivent être conservés lorsque leur statut est Valide ou Non valide.

Rechercher les valeurs non valides dans. Indiquez où rechercher des valeurs non valides.

- **Tous les champs.** Le nœud Sélectionner recherche les valeurs non valides dans tous les champs.

- **Champs sélectionnés dans le tableau.** Le nœud Sélectionner ne vérifie que les champs sélectionnés dans le tableau de sortie Qualité.
- **Champs dont le pourcentage de qualité est supérieur à.** Le nœud Sélectionner ne vérifie que les champs dont le pourcentage d'enregistrements complets est supérieur au seuil indiqué. La valeur de seuil par défaut est 50 %.

Un enregistrement est considéré comme non valide lorsqu'il contient une valeur non valide. Indiquez la condition d'identification d'un enregistrement comme non valide.

- **N'importe quel champ ci-dessus.** Le nœud Sélectionner considère qu'un enregistrement n'est pas valide si l'*un* des champs spécifiés ci-dessus contient une valeur non valide pour cet enregistrement.
- **Tous les champs ci-dessus.** Le nœud Sélectionner considère qu'un enregistrement n'est pas valide si *tous* les champs spécifiés ci-dessus contiennent des valeurs non valides pour cet enregistrement.

Génération d'autres noeuds en vue d'une préparation de données

La plupart des noeuds servant à la préparation des données peuvent être générés directement à partir du navigateur Audit données, y compris les noeuds Recoder, Discrétiser et Calculer. Par exemple :

- Vous pouvez calculer un nouveau champ sur la base des valeurs *valeurréclamation* et *revenuferme* en les sélectionnant dans le rapport d'audit et en choisissant Calculer dans le menu Générer. Le nouveau nœud est ajouté à l'espace de travail de flux.
- De même, sur la base des résultats de l'audit, vous pouvez déterminer si le recodage de *revenuferme* en intervalles de type centile fournit une analyse plus précise. Pour générer un nœud Discrétiser, sélectionnez la ligne de champ dans l'affichage et choisissez Discrétiser dans le menu Générer.

Une fois le nœud généré et ajouté à l'espace de travail de flux, vous devez le relier au flux et ouvrir le nœud afin de spécifier les options des champs sélectionnés.

Noeud Transformation

La normalisation des champs d'entrée est une étape importante préalable à l'application de techniques de scoring traditionnelles, telles que la régression, la régression logistique et l'analyse discriminante. Ces techniques reposent sur des hypothèses relatives aux proportions normales des données qui peuvent ne pas s'appliquer à de nombreux fichiers de données brutes. Une méthode de traitement des données concrètes consiste à appliquer des transformations qui rapprochent un élément de données brutes d'une proportion plus normale. En outre, les champs normalisés sont facilement comparables entre eux. Par exemple, les revenus et l'âge se situent sur des échelles totalement différentes dans un fichier de données brutes. Une fois ces éléments normalisés, leur impact relatif est facile à interpréter.

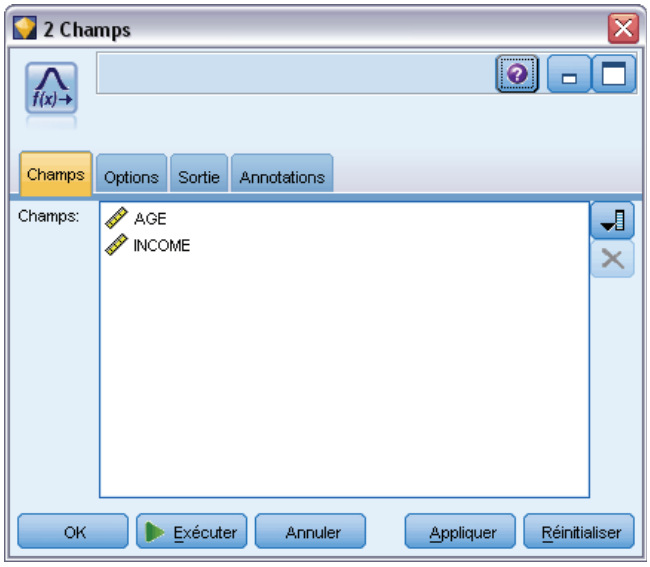
Le noeud Transformation fournit un afficheur de résultats qui vous permet de procéder à une évaluation visuelle rapide de la meilleure transformation à utiliser. Vous pouvez voir en un coup d'oeil si les variables sont normalement réparties et, si nécessaire, choisir la transformation à appliquer. Vous pouvez choisir plusieurs champs et appliquer une transformation par champ.

Après avoir sélectionné les transformations préférées pour les champs, vous pouvez générer des noeuds Calculer ou Remplacer qui exécutent les transformations et connecter ces noeuds au flux. Le noeud Calculer crée des champs tandis que le noeud Remplacer transforme les champs existants. Pour plus d'informations, reportez-vous à la section Génération de graphiques sur p. 440.

Onglet Champs du noeud Transformation

Dans l'onglet Champs, indiquez les champs de données à utiliser pour afficher les transformations possibles et les appliquer. Seuls les champs numériques peuvent être transformés. Cliquez sur le bouton de sélection de champ, puis sélectionnez un ou plusieurs champs numériques dans la liste affichée.

Figure 6-30
Noeud Transformation : Onglet Champs



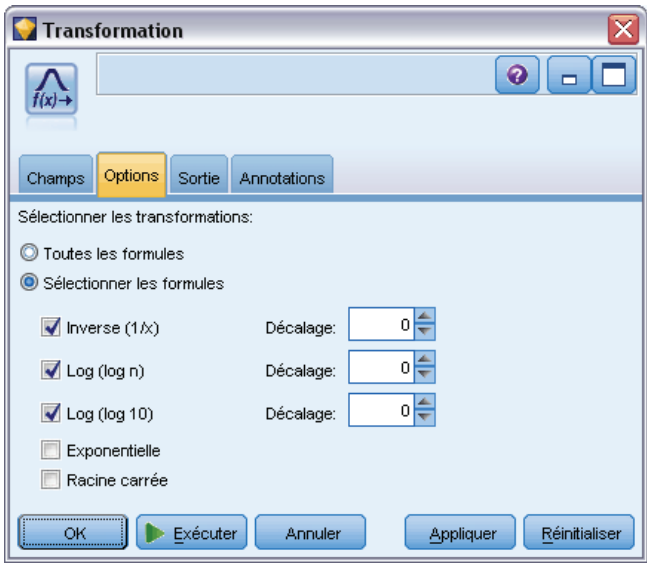
Onglet Options du noeud Transformation

L'onglet Options permet d'indiquer les types de transformation à inclure. Vous pouvez décider d'inclure toutes les transformations disponibles ou sélectionner des transformations distinctes.

Dans ce dernier cas, vous pouvez également entrer un nombre pour décaler les données en vue des transformations inverse et logarithmique. Cela se révèle utile dans les cas où une large proportion de zéros dans les données pourrait biaiser les résultats de moyenne et d'écart-type.

Par exemple, supposez que vous avez un champ nommé *BALANCE* qui contient des valeurs nulles et que vous souhaitez lui appliquer une transformation inverse. Pour éviter tout biais indésirable, vous sélectionnez Inverse (1/x) et entrez 1 dans le champ Utiliser un décalage de données. (Notez que ce décalage n'est pas lié au décalage appliqué par la fonction séquentielle @OFFSET dans IBM® SPSS® Modeler.)

Figure 6-31
Noeud Transformation : Onglet Options



Toutes les formules. Indique que toutes les transformations disponibles doivent être calculées et figurer dans la sortie.

Sélectionner les formules. Permet de sélectionner les transformations à calculer et à afficher dans la sortie.

- **Inverse (1/x).** Indique que la transformation inverse doit être affichée dans la sortie.
- **Log (log n).** Indique que la transformation $\log_n$ doit être affichée dans la sortie.
- **Log (log 10).** Indique que la transformation $\log_{10}$ doit être affichée dans la sortie.
- **Exponentielle.** Indique que la transformation exponentielle (e^x) doit figurer dans la sortie.
- **Racine carrée.** Indique que la transformation racine carrée doit être affichée dans la sortie.

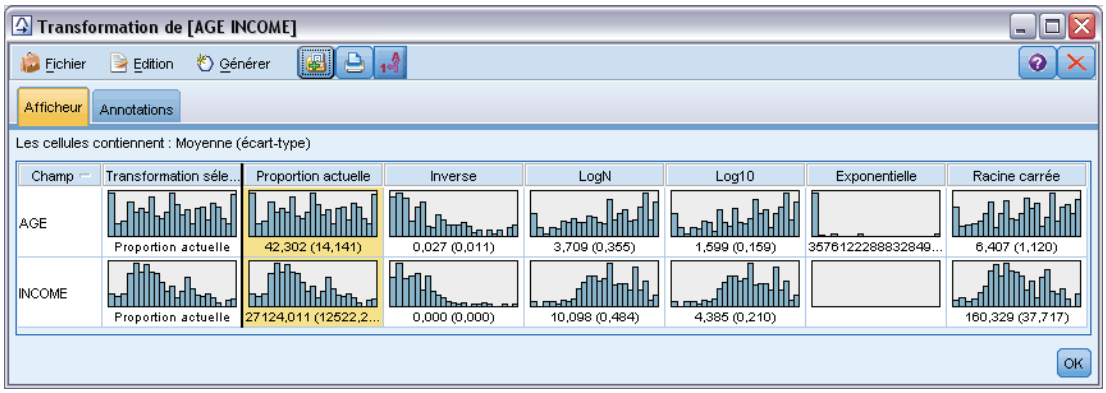
Onglet Sortie du noeud Transformation

L'onglet Sortie vous permet de préciser le format et l'emplacement de la sortie. Vous pouvez choisir d'afficher les résultats à l'écran ou de les envoyer vers un des types de fichier standard. Pour plus d'informations, reportez-vous à la section Noeud de sortie - Onglet Sortie sur p. 406.

Afficheur de résultats du noeud Transformation

L'afficheur de résultats vous permet de consulter les résultats de l'exécution du noeud Transformation. L'afficheur est un outil puissant qui affiche plusieurs transformations par champ dans des vues miniatures de la transformation, vous permettant ainsi de comparer rapidement les champs. Utilisez les options du menu Fichier pour enregistrer, exporter ou imprimer la sortie. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Figure 6-32
Affichage des transformations disponibles par champ



Sous chaque transformation (autre que Transformation sélectionnée), une légende est affichée sous le format :

Mean (Standard deviation)

Génération des noeuds pour les transformations

L'afficheur de résultats fournit un point de départ à la préparation des données. Par exemple, vous souhaitez normaliser le champ *AGE* pour pouvoir utiliser une technique de scoring (telle que la régression logistique ou l'analyse discriminante) qui suppose une proportion normale. D'après les graphiques initiaux et les statistiques récapitulatives, vous pouvez décider de transformer le champ *AGE* en fonction d'une distribution particulière (par exemple, log). Après avoir sélectionné la proportion préférée, vous pouvez générer un noeud de dérivation avec une transformation standardisée à utiliser pour l'évaluation.

Vous pouvez générer les noeuds d'opérations de champ suivants à partir de l'afficheur de résultats :

- Calculer
- Remplacer

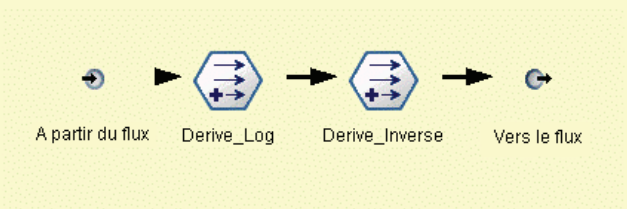
Un noeud Calculer crée des champs avec les transformations souhaitées, tandis que le noeud Remplacer transforme les champs existants. Les noeuds sont placés dans l'espace de travail, sous la forme d'un super noeud.

Si vous sélectionnez la même transformation pour différents champs, un noeud Calculer ou Remplacer contient les formules de ce type de transformation pour tous les champs auxquels cette transformation s'applique. Par exemple, supposez que vous avez sélectionné les champs et les transformations suivants pour générer un noeud Calculer :

Champ	**Transformation**
AGE	Proportion actuelle
INCOME	Log
OPEN_BAL	Inverse
BALANCE	Inverse

Les noeuds suivants sont inclus dans le super noeud :

Figure 6-33
Super noeud dans l'espace de travail



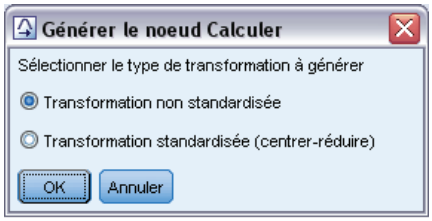
Dans cet exemple, le noeud Derive_Log contient la formule logarithmique du champ *INCOME* et le noeud Derive_Inverse, les formules inverses des champs *OPEN_BAL* et *BALANCE*.

Pour générer un noeud

- Pour chaque champ apparaissant dans l'afficheur de résultats, sélectionnez la transformation souhaitée.
- Dans le menu Générer, choisissez Noeud Calculer ou Noeud Remplacer.

Cela affiche la boîte de dialogue Générer le noeud Calculer ou Générer le noeud Remplacer, selon le cas.

Figure 6-34
Choix d'une transformation standardisée ou non standardisée



Choisissez Transformation non standardisée ou Transformation standardisée (centrer-réduire), comme vous le souhaitez. La seconde option applique un score *z* à la transformation ; les scores *z* représentent les valeurs en tant que fonction de la distance par rapport à la moyenne de la variable dans les écarts-types. Par exemple, si vous appliquez la transformation logarithmique au champ *AGE* et que vous choisissez une transformation standardisée, l'équation finale du noeud généré est :

```
(log(AGE)-Mean)/SD
```

Lorsqu'un noeud est généré et qu'il apparaît dans l'espace de travail de flux :

- Connectez-le au flux.
- Dans le cas d'un super noeud, vous pouvez double-cliquer sur le noeud pour consulter son contenu.
- Vous pouvez double-cliquer sur un noeud Calculer ou Remplacer pour modifier les options des champs sélectionnés.

Génération de graphiques

Vous pouvez générer une sortie d'histogramme grandeur nature à partir d'un histogramme en miniature dans l'afficheur de résultats.

Pour générer un graphique

- Double-cliquez sur un graphique en miniature dans l'afficheur de résultats.

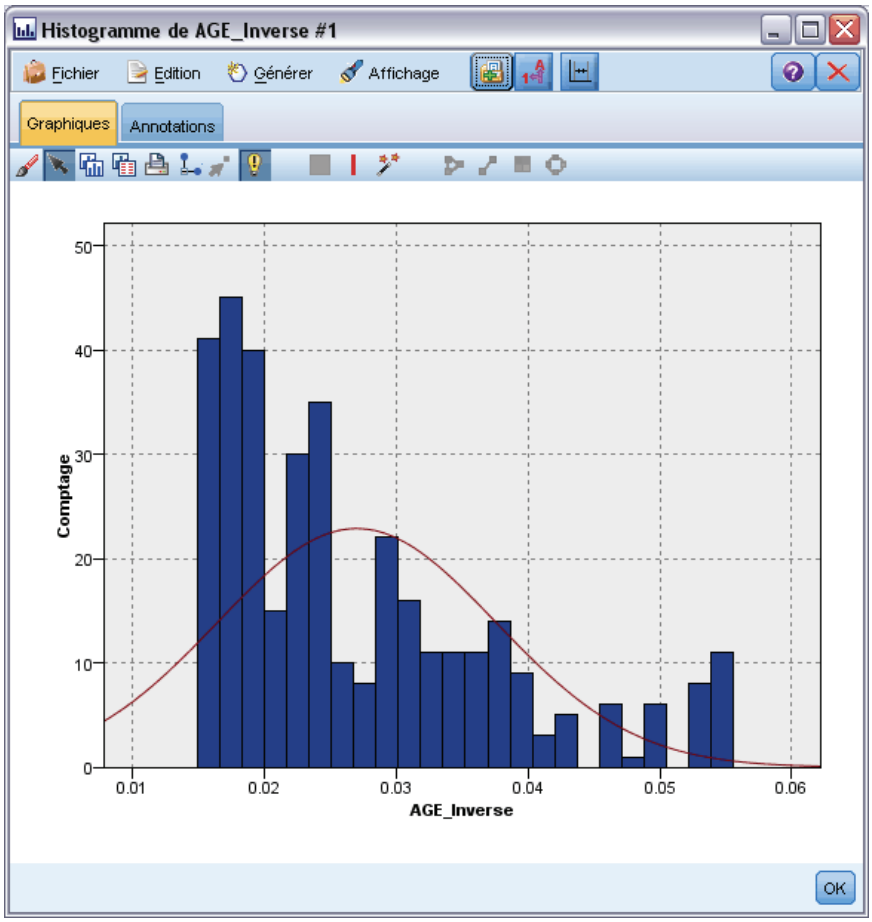
  ou

- Sélectionnez un graphique en miniature dans l'afficheur de résultats.
- Dans le menu Générer, sélectionnez Sortie graphique.

Vous affichez ainsi un histogramme avec une courbe de distribution normale en superposition. Cela vous permet de déterminer à quel point chaque transformation disponible correspond à la distribution normale.

Remarque : Vous ne pouvez générer un diagramme que si le noeud Transformer qui a créé la sortie est connecté au flux.

Figure 6-35
Histogramme de transformation avec courbe de distribution normale en superposition



Autres opérations

Dans l'afficheur des résultats, vous pouvez effectuer les opérations suivantes :

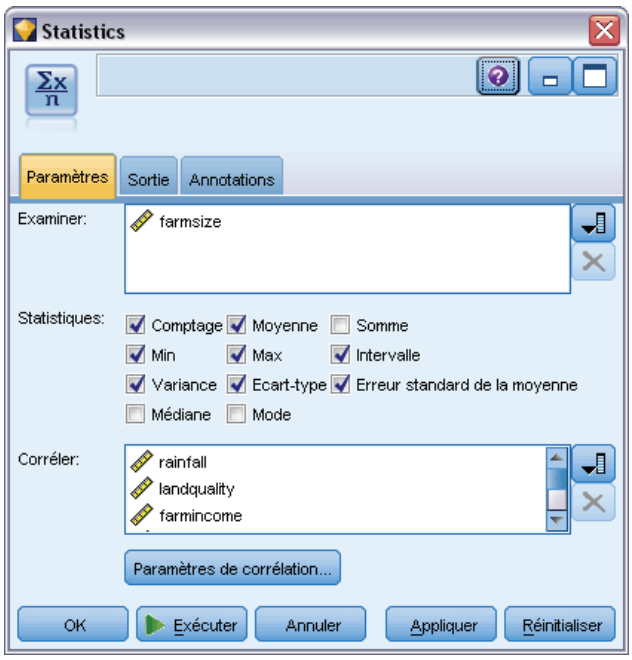
- Triez la grille de sortie sur la base de la colonne Champ.
- Exportez la sortie vers un fichier HTML. Pour plus d'informations, reportez-vous à la section Exportation des sorties sur p. 402.

Noeud Statistiques

Le nœud Statistiques fournit des informations récapitulatives de base sur les champs numériques. Ces statistiques peuvent porter sur des champs individuels et sur les corrélations entre les champs.

Noeud Statistiques - Onglet Paramètres

Figure 6-36
Noeud Statistiques : onglet Paramètres



Examiner. Sélectionnez les champs sur lesquels obtenir des statistiques récapitulatives individuelles. Vous pouvez sélectionner plusieurs champs.

Statistics. Sélectionnez les statistiques à créer. Les options disponibles sont les suivantes : Comptage, Moyenne, Somme, Minimum, Maximum, Intervalle, Variance, Ecart-type, Erreur standard de la moyenne, Médiane et Mode.

Corréler. Sélectionnez les champs à mettre en corrélation. Vous pouvez sélectionner plusieurs champs. Lorsque vous sélectionnez des champs de corrélation, la corrélation entre chaque champ Examiner et les champs de corrélation est indiquée dans la sortie.

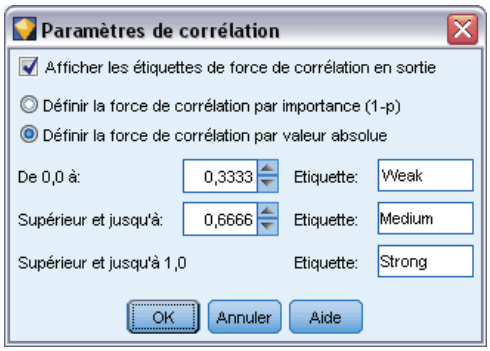
Paramètres de corrélation. Vous pouvez définir les options d'affichage de la force des corrélations dans la sortie.

Paramètres de corrélation

IBM® SPSS® Modeler permet de définir les corrélations à l'aide d'étiquettes descriptives afin de mettre en évidence des relations importantes. La **corrélation** mesure la force de la relation entre deux champs continus (intervalle numérique). Ses valeurs sont comprises entre –1,0 et 1,0. Les valeurs proches de +1,0 indiquent une association positive forte, de sorte que les valeurs élevées d'un champ sont associées aux valeurs élevées d'un autre champ, et les valeurs faibles du champ aux valeurs faibles de l'autre champ. Les valeurs proches de –1 indiquent une association négative forte, de sorte que les valeurs élevées d'un champ sont associées aux valeurs faibles de l'autre, et inversement. Les valeurs proches de 0,0 indiquent une association faible, de sorte que les valeurs des deux champs sont plus ou moins indépendantes.

Vous pouvez contrôler l'affichage des étiquettes de corrélation, modifier les seuils définissant les catégories et modifier les étiquettes utilisées pour chaque intervalle. Etant donné que la manière dont vous définissez les valeurs de corrélation dépend essentiellement du type de problème, vous pouvez personnaliser les intervalles et les étiquettes en fonction de votre situation.

Figure 6-37
Boîte de dialogue Paramètres de corrélation



Afficher les étiquettes de force de corrélation en sortie. Par défaut, cette option est sélectionnée. Désélectionnez-la pour ne pas insérer les étiquettes descriptives dans la sortie.

Force de corrélation. Deux options permettent de définir et d'étiqueter la force des corrélations :

- **Définir la force de corrélation par importance (1-p).** Applique une étiquette aux corrélations en fonction de leur importance, cette dernière étant égale à 1 moins la signification, ou à 1 moins la probabilité que la différence de moyenne ne soit due qu'au hasard. Plus cette valeur est proche de 1, plus la probabilité que les deux champs ne soient *pas* indépendants (en d'autres termes, qu'une relation existe entre eux) est forte. En général, il est recommandé d'étiqueter les corrélations en fonction de leur importance plutôt qu'en fonction des valeurs absolues, car cela rend compte de la variabilité des données. Par exemple, un coefficient de 0,6 peut s'avérer très significatif dans un ensemble de données et pas du tout dans un autre. Par défaut, les valeurs d'importance comprises entre 0 et 0,9 sont repérées par une étiquette *Faible*, celles entre 0,9 et 0,95 par une étiquette *Moyen*, et celles entre 0,95 et 1 par une étiquette *Elevé*.
- **Définir la force de corrélation par valeur absolue.** Applique une étiquette aux corrélations en fonction de la valeur absolue du coefficient de corrélation de Pearson, qui, comme indiqué précédemment, est comprise entre –1 et 1. Plus la valeur absolue de cette mesure est proche de 1, plus la corrélation est forte. Par défaut, les corrélations comprises entre 0 et 0,3333 (en valeur absolue) sont repérées par une étiquette *Faible*, celles entre 0,3333 et 0,6666 par une étiquette *Moyen*, et celles entre 0,6666 et 1 par une étiquette *Elevé*. Toutefois, la signification d'une valeur donnée peut difficilement être généralisée d'un ensemble de donnée à un autre. C'est pourquoi, dans la plupart des cas, il est recommandé de définir des corrélations sur la base des probabilités et non des valeurs absolues.

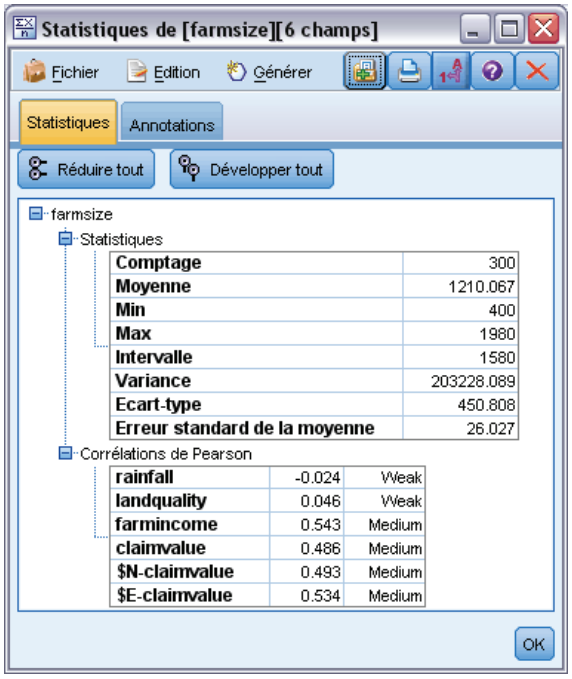
Navigateur de sortie du nœud Statistiques

Le navigateur de sortie du nœud Statistics affiche les résultats de l'analyse statistique et permet d'effectuer un certain nombre d'opérations : sélection de champs, création de noeuds en fonction d'une sélection, enregistrement et impression des résultats. Les options standard d'enregistrement,

d'exportation et d'impression sont disponibles dans le menu Fichier, et celles d'édition dans le menu Edition. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Lorsque vous accédez pour la première fois à la sortie du nœud Statistiques, les résultats sont développés. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement située à gauche des résultats à masquer ou cliquez sur le bouton Réduire tout pour réduire tous les résultats. Pour afficher de nouveau les résultats, utilisez la commande de développement située à gauche des résultats à afficher ou cliquez sur le bouton Développer tout pour développer tous les résultats.

Figure 6-38
Navigateur de sortie du nœud Statistiques



La sortie comporte une section pour chaque champ *Examiner*, contenant un tableau des statistiques demandées.

- **Effectifs.** Indique le nombre d'enregistrements contenant des valeurs valides pour le champ.
- **Moyenne.** Indique la valeur moyenne du champ sur l'ensemble des enregistrements.
- **Somme :** Indique la somme des valeurs du champ sur l'ensemble des enregistrements.
- **Minimum** Indique la valeur minimale du champ.
- **Maximum** Indique la valeur maximale du champ.
- **Intervalle :** Indique la différence entre les valeurs minimale et maximale.
- **Variance.** Mesure de la variabilité des valeurs d'un champ. Pour la calculer, prenez la différence entre chaque valeur et la moyenne globale, élevez-la au carré, additionnez toutes les valeurs et divisez la somme par le nombre d'enregistrements.
- **Ecart-type.** Autre mesure de la variabilité des valeurs d'un champ, correspondant à la racine carrée de la variance.

- **Erreur standard de la moyenne.** Mesure de la part d'incertitude dans l'estimation de la moyenne d'un champ si cette moyenne est censée s'appliquer à de nouvelles données.
- **Médiane.** Valeur médiane du champ, c'est-à-dire valeur qui sépare la première moitié des données de la seconde moitié (sur la base des valeurs du champ).
- **Mode.** Valeur unique la plus courante dans les données.

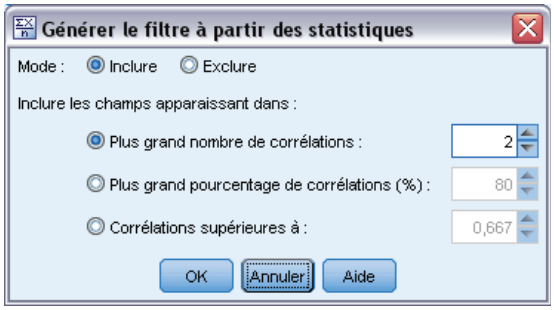
Corrélations : Si vous avez spécifié des champs de corrélation, la sortie contient également une section indiquant la corrélation de Pearson entre le champ Examiner et chaque champ de corrélation, ainsi que des étiquettes descriptives facultatives pour les valeurs de corrélation. Pour plus d'informations, reportez-vous à la section Paramètres de corrélation sur p. 442.

Menu Générer. Le menu Générer contient des options permettant de générer des noeuds.

- **Filtrer.** Permet de générer un nœud Filtrer, afin d'éliminer les champs non corrélés ou faiblement corrélés aux autres champs.

Génération d'un nœud Filtrer à partir de Statistics

Figure 6-39
Boîte de dialogue Générer le filtre à partir des statistiques



Le nœud Filtrer, généré à partir du navigateur de sortie du nœud Statistiques, filtre les champs en fonction de leurs corrélations avec d'autres champs. Il trie les corrélations dans l'ordre de leur valeur absolue, prend les corrélations les plus grandes (selon le critère défini dans la boîte de dialogue) et crée un filtre qui utilise tous les champs apparaissant dans l'une de ces corrélations.

Mode. Permet de définir le mode de sélection des corrélations. Enlever conserve les champs apparaissant dans les corrélations spécifiées. Isoler filtre les champs.

Inclure/exclure les champs apparaissant dans. Définissez le critère de sélection des corrélations.

- **Plus grand nombre de corrélations.** Sélectionne le nombre indiqué de corrélations et inclut/exclut les champs qui y apparaissent.
- **Plus grand pourcentage de corrélations (%).** Sélectionne le pourcentage spécifié (*n* %) de corrélations et inclut/exclut les champs qui apparaissent dans ces corrélations.
- **Corrélations supérieures à.** Sélectionne les corrélations supérieures, en valeur absolue, au seuil indiqué.

Noeud Moyennes

Le noeud Moyennes compare les moyennes de groupes indépendants ou de paires de champs associés, afin de détecter toute différence sensible. Par exemple, vous pouvez comparer les revenus moyens avant et après l'application d'une promotion, ou les revenus des clients qui ont et qui n'ont pas bénéficié de cette promotion.

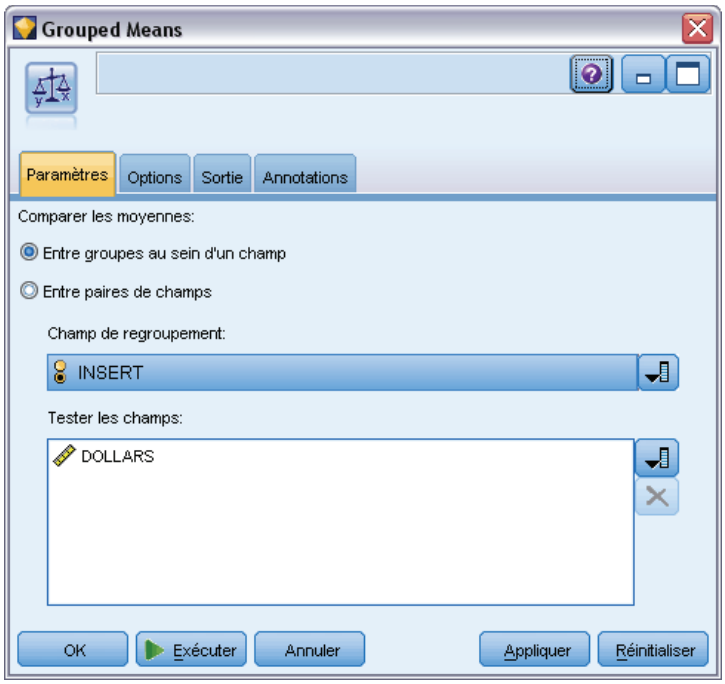
Deux méthodes de comparaison des moyennes s'offrent à vous, selon vos données :

- **Entre groupes au sein d'un champ.** Pour comparer des groupes indépendants, sélectionnez un champ de test et un champ de regroupement. Par exemple, vous pouvez exclure un échantillon de clients « représentatifs » dans le cas d'une offre de promotion et comparer la moyenne des revenus de ce groupe avec celle de tous les autres clients. Dans ce cas, vous spécifiez un champ de test unique indiquant les revenus de chaque client, ainsi qu'un champ booléen ou un champ nominal précisant si chacun a bénéficié de l'offre. Les échantillons sont indépendants dans le sens où chaque enregistrement est affecté à un groupe ou à un autre. Il est, en outre, impossible de lier un membre d'un groupe à un membre d'un autre groupe. Vous pouvez également définir un champ nominal comportant plus de deux valeurs pour comparer la moyenne de plusieurs groupes. Lorsqu'il est exécuté, le noeud effectue un test ANOVA unilatéral sur les champs sélectionnés. S'il n'existe que deux groupes de champs, les résultats du test ANOVA unilatéral sont globalement identiques à ceux d'un test *t* pour échantillons indépendants. Pour plus d'informations, reportez-vous à la section Comparaison des moyennes de groupes indépendants sur p. 446.
- **Entre paires de champs.** Lorsque vous comparez la moyenne de deux champs liés, vous devez réunir les groupes par paires pour que les résultats soient significatifs. Par exemple, vous pouvez comparer le revenu moyen d'un même groupe de clients avant et après l'application d'une promotion, ou bien encore les taux d'utilisation d'un service dans les paires époux-épouse pour voir s'il y a des différences. Chaque enregistrement contient deux mesures distinctes mais liées pouvant être comparées de manière significative. Lorsqu'il est exécuté, le noeud effectue un test *t* pour paires d'échantillons sur chaque paire de champs sélectionnée. Pour plus d'informations, reportez-vous à la section Comparaison de moyennes entre paires de champs sur p. 447.

Comparaison des moyennes de groupes indépendants

Sélectionnez Entre groupes au sein d'un champ dans le noeud Moyennes pour comparer la moyenne d'au moins deux groupes indépendants.

Figure 6-40
Comparaison des moyennes entre les groupes d'un champ



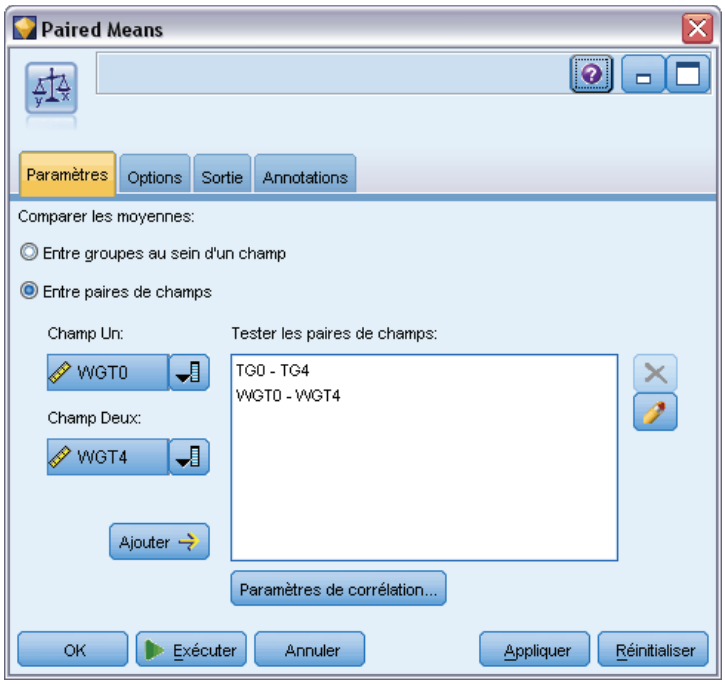
Champ de regroupement. Sélectionnez un champ booléen ou un champ nominal comportant au moins deux valeurs distinctes et répartissant les enregistrements entre les différents groupes à comparer (personnes qui ont bénéficié d'une offre et personnes qui n'en n'ont pas bénéficié, par exemple). Quel que soit le nombre de champs de test, vous ne pouvez sélectionner qu'un seul champ de regroupement.

Tester les champs. Sélectionnez un ou plusieurs champs numériques contenant les mesures à tester. Un test distinct est effectué pour chaque champ sélectionné. Par exemple, vous pouvez tester l'incidence que peut avoir une promotion sur l'utilisation, les revenus et l'attrition.

Comparaison de moyennes entre paires de champs

Sélectionnez Entre paires de champs dans le noeud Moyennes pour comparer la moyenne de différents champs. Ces champs doivent être liés d'une manière ou d'une autre pour que les résultats soient significatifs (revenus avant et après une promotion, par exemple). Vous pouvez également sélectionner plusieurs paires de champs.

Figure 6-41
Comparaison de moyennes entre des paires de champs



Champ Un. Sélectionnez un champ numérique contenant la première des mesures à comparer. Dans une étude de type « avant-après », il s'agit du champ Avant.

Champ Deux. Sélectionnez le second champ à comparer.

Ajouter. Ajoute la paire sélectionnée à la liste Tester les paires de champs.

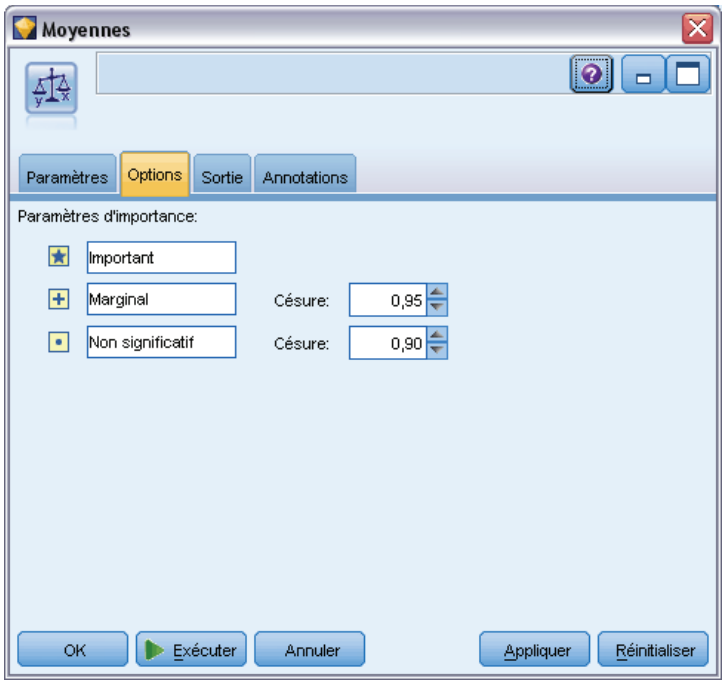
Si nécessaire, répétez les sélections de champ pour ajouter plusieurs paires à la liste.

Paramètres de corrélation. Permet de définir les options d'étiquetage de la force des corrélations. Pour plus d'informations, reportez-vous à la section Paramètres de corrélation sur p. 442.

Options du noeud Moyennes

L'onglet Options vous permet de définir les valeurs de seuil *p* utilisées pour étiqueter les résultats comme étant importants, marginaux ou non significatifs. Vous pouvez également éditer l'étiquette de chaque classement. L'importance est mesurée en termes de pourcentage et peut être définie globalement en soustrayant à 1 la probabilité d'obtention d'un résultat (la différence de moyenne entre deux champs, par exemple) aussi élevée ou plus élevée que le résultat généré de manière aléatoire. Par exemple, une valeur *p* supérieure à 0,95 indique une probabilité inférieure à 5% que le résultat soit dû exclusivement au hasard.

Figure 6-42
Paramètres d'importance



Etiquettes d'importance. Vous pouvez éditer les étiquettes servant à repérer chaque paire ou groupe de champs dans la sortie. Les étiquettes par défaut sont les suivantes : *Important*, *Marginal* et *Non significatif*.

Valeurs Césure. Indique le seuil de chaque rang. En général, les valeurs *p* supérieures à 0,95 sont considérées comme importantes et celles inférieures à 0,9 comme non significatives ; ces seuils peuvent être ajustés si nécessaire.

Remarque : Les mesures d'importance sont disponibles dans plusieurs noeuds. Les calculs possibles dépendent du noeud, ainsi que du type de la cible et des champs d'entrée utilisés, mais il est toujours possible de comparer les valeurs car toutes sont mesurées en pourcentage.

Navigateur de sortie du noeud Moyennes

Le navigateur de sortie du noeud Moyennes affiche les données sous forme de tableau croisé et vous permet d'exécuter des opérations standard : sélection et copie du tableau ligne par ligne, tri par colonne, et enregistrement et impression du tableau. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Les informations particulières contenues dans le tableau dépendent du type de comparaison (groupes faisant partie d'un même champ ou de champs distincts).

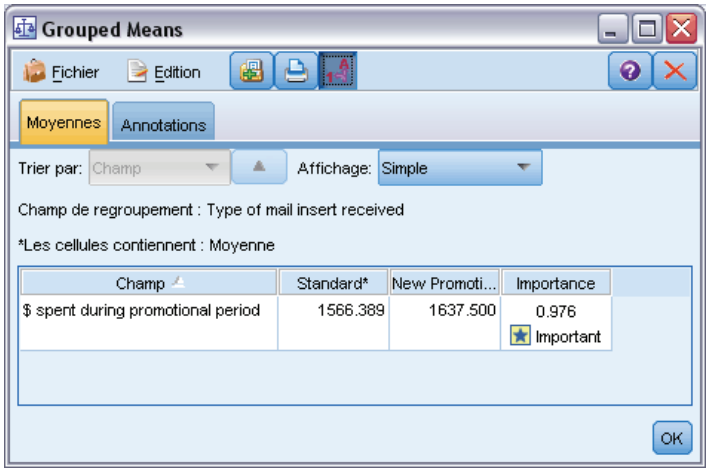
Trier par. Permet de trier la sortie en fonction d'une colonne donnée. Cliquez sur la flèche vers le haut ou vers le bas pour modifier le sens du tri. Vous pouvez également cliquer sur l'en-tête de la colonne en fonction de laquelle réaliser le tri. (Pour modifier le sens du tri dans la colonne, cliquez à nouveau sur son en-tête.)

Affichage. Vous pouvez également choisir Simple ou Options avancées pour contrôler le niveau de détail de l'affichage. La vue avancée inclut toutes les informations de la vue simple, auxquelles elle ajoute des informations supplémentaires.

Sortie du noeud Moyennes par comparaison des groupes d'un champ

Lorsque vous comparez les groupes d'un champ, le nom du champ de regroupement apparaît au-dessus du tableau de sortie, et les moyennes et les statistiques liées sont indiquées séparément pour chaque groupe. Le tableau inclut une ligne distincte pour chaque champ de test.

Figure 6-43
Comparaison des groupes d'un champ



Les colonnes suivantes apparaissent :

- **Champ.** Indique le nom des champs de test sélectionnés.
- **Moyennes par groupe.** Affiche la moyenne de chaque catégorie du champ de regroupement. Par exemple, vous pouvez comparer ceux qui ont bénéficié d'une offre spéciale (*Nouvelle promotion*) avec ceux qui n'en ont pas bénéficié (*Standard*). L'écart-type, l'erreur standard et le comptage sont également affichés dans la vue avancée.
- **Importance.** Affiche la valeur et l'étiquette d'importance. Pour plus d'informations, reportez-vous à la section Options du noeud Moyennes sur p. 448.

Sortie avancée

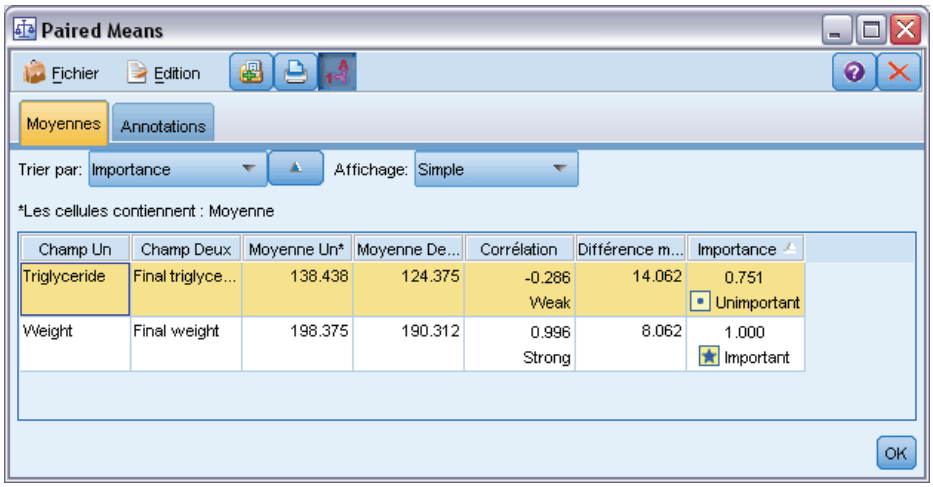
Dans la vue avancée, les colonnes supplémentaires suivantes apparaissent.

- **Test F.** Ce test est basé sur le rapport entre la variance entre les groupes et la variance au sein de chaque groupe. Si les moyennes sont identiques pour tous les groupes, le rapport *F* devrait être proche de 1, puisqu'il s'agit dans les deux cas de l'estimation de la même variance de population. Plus le rapport est élevé, plus la variation entre les groupes est importante et plus la probabilité d'une différence significative est forte.
- **ddl.** Affiche les degrés de liberté.

Sortie du noeud Moyennes par comparaison de paires de champs

Lorsque vous comparez des champs distincts, le tableau de sortie inclut une ligne pour chaque paire de champs sélectionnée.

Figure 6-44
Comparaison de paires de champs



- **Champ Un/Deux.** Affiche le nom des premier et second champs de chaque paire. L'écart-type, l'erreur standard et le comptage sont également affichés dans la vue avancée.
- **Moyenne Un/Deux.** Affiche la moyenne de chaque champ.
- **Corrélation.** Mesure la force de la relation entre deux champs continus (intervalle numérique). Les valeurs proches de 1 indiquent une association positive forte et les valeurs proches de –1., une association négative forte. Pour plus d'informations, reportez-vous à la section Paramètres de corrélation sur p. 442.
- **Différence moyenne.** Affiche la différence entre les deux moyennes de champ.
- **Importance.** Affiche la valeur et l'étiquette d'importance. Pour plus d'informations, reportez-vous à la section Options du noeud Moyennes sur p. 448.

Sortie avancée

La sortie avancée ajoute les colonnes suivantes :

Intervalle de confiance de 95 %. Limites inférieure et supérieure de l'intervalle où la moyenne réelle est susceptible de figurer dans 95 % des échantillons possibles de cette taille au sein de cette population.

Test T. La statistique *t* est obtenue en divisant la différence de moyenne par l'erreur standard correspondante. Plus la valeur absolue de cette statistique est élevée, plus la probabilité que les moyennes soient différentes est forte.

ddl. Affiche les degrés de liberté correspondant à la statistique.

Noeud Rapport

Ce noeud permet de créer des rapports formatés contenant du texte fixe et des données, ainsi que des expressions calculées à partir de ces données. Le format du rapport est déterminé par des modèles texte définissant la structure du texte fixe et de la sortie de données. Vous pouvez définir un formatage de texte personnalisé en utilisant des balises HTML dans le modèle et en définissant des options dans l'onglet Sortie. Les valeurs de données et autres sorties conditionnelles sont incluses dans le rapport à l'aide des expressions CLEM du modèle.

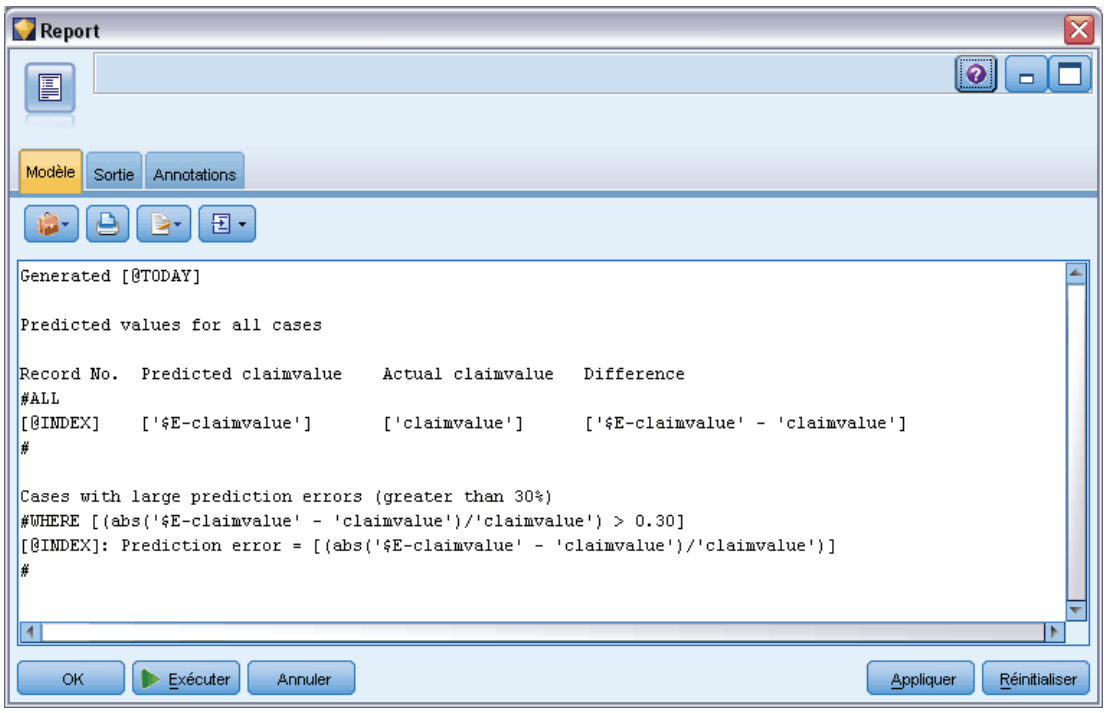
Alternatives au noeud Rapport

Le noeud Rapport est le plus souvent utilisé pour répertorier une sortie d'enregistrements ou d'observations d'un flux (par exemple, tous les enregistrements répondant à une certaine condition). A cet égard, il peut être considéré comme une alternative moins structurée au noeud Table.

- Si vous souhaitez un rapport répertoriant les informations de champ ou tout autre élément défini dans le flux plutôt que les données elles-mêmes (par exemple, des définitions de champ indiquées dans un noeud Typer), vous pouvez alors utiliser un script à la place.
- Pour générer un rapport incluant plusieurs objets de sortie (par exemple un ensemble de modèles, de tableaux et de graphiques générés par un ou plusieurs flux) et pouvant être créé sous plusieurs formats (texte, HTML et Microsoft Word/Office), vous pouvez utiliser un projet IBM® SPSS® Modeler.
- Pour produire une liste de noms de champ sans utiliser la génération de scripts, vous pouvez utiliser un noeud Table précédé d'un noeud Echantillonner qui supprime tous les enregistrements. Cette opération entraîne la création d'un tableau sans lignes, qui peut être transposé lors de l'exportation afin de produire une liste de noms de champ dans une seule colonne. (Pour ce faire, sélectionnez Transposer les données dans l'onglet Sortie du noeud Table.)

Noeud Rapport - Onglet Modèle

Figure 6-45
Noeud Rapport : onglet Modèle



Création d'un modèle. Pour définir le contenu du rapport, vous devez créer un modèle dans l'onglet Modèle du nœud Rapport. Ce modèle se compose de lignes de texte définissant chacune un aspect du contenu du rapport, et de lignes de balises indiquant la portée de chaque ligne. Encadrées de crochets ([]), les expressions CLEM des lignes de contenu sont évaluées avant l'écriture de la ligne dans le rapport. Les portées suivantes sont disponibles pour les lignes du modèle :

Fixe : Les lignes qui ne portent aucune indication sont considérées comme fixes. Les lignes fixes ne sont copiées qu'une fois dans le rapport, après l'évaluation des éventuelles expressions qu'elles contiennent. Par exemple, la ligne :

```
This is my report, printed on [@TODAY]
```

entraîne la copie dans le rapport d'une ligne contenant le texte indiqué et la date actuelle.

Global (Itérer TOUT). Les lignes comprises entre les balises spéciales #ALL et # sont copiées dans le rapport une fois pour chaque enregistrement de données d'entrée. Les expressions CLEM (entre crochets) sont évaluées en fonction de l'enregistrement actuel pour chaque ligne de sortie. Par exemple, les lignes :

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

copient une ligne par enregistrement, indiquant le numéro de l'enregistrement et l'âge.

Pour générer la liste de tous les enregistrements :

```
#ALL
[Age]   [Sex]   [Cholesterol]    [BP]
#
```

Conditionnel (Itérer SI). Les lignes figurant entre les balises spéciales #WHERE <*condition*> et # sont copiées dans le rapport une fois pour chaque enregistrement pour lequel la condition spécifiée a la valeur true (vrai). La condition est une expression CLEM. (Dans la condition WHERE, les crochets sont facultatifs.) Par exemple, les lignes :

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

copient dans le fichier une ligne pour chaque enregistrement dans lequel le sexe a pour valeur *M*. Le rapport complet contient les lignes fixes, globales et conditionnelles définies après l'application du modèle aux données d'entrée.

Dans l'onglet Sortie, vous pouvez spécifier les options d'affichage ou d'enregistrement de résultats, qui seront communes à plusieurs types de nœud de sortie. Pour plus d'informations, reportez-vous à la section Noeud de sortie - Onglet Sortie sur p. 406.

Sortie des données au format HTML ou XML

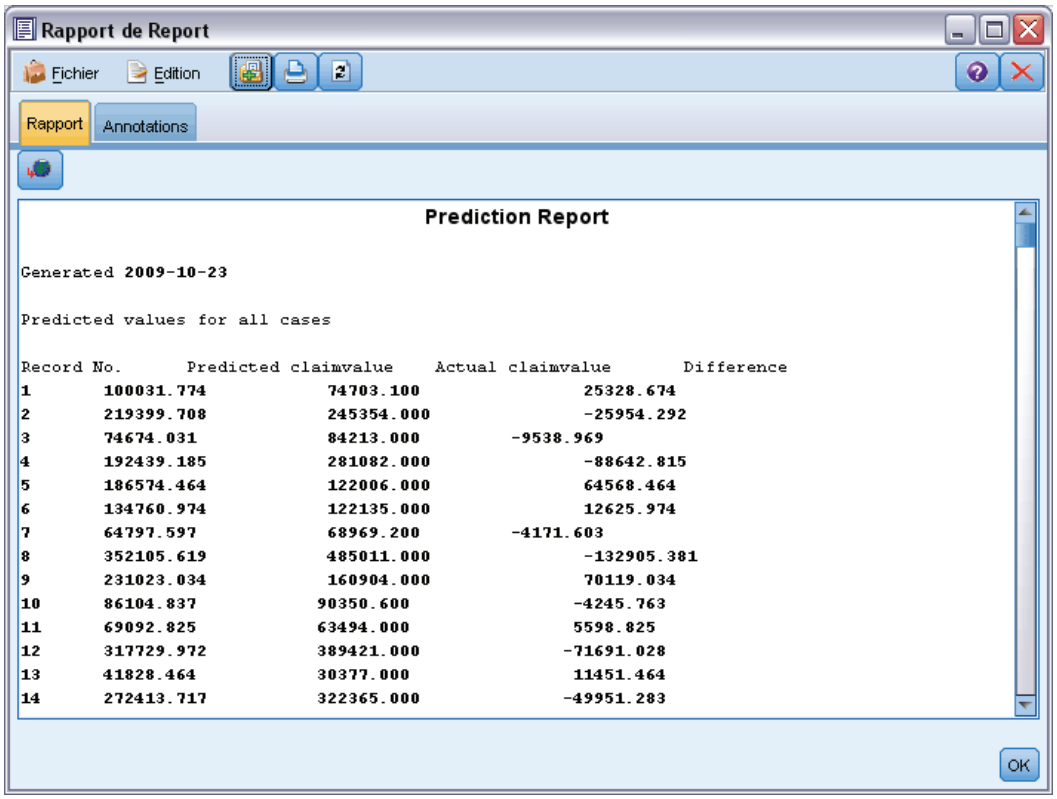
Vous pouvez inclure des balises HTML ou XML directement dans le modèle pour générer des rapports dans l'un ou l'autre des formats. Par exemple, le modèle suivant génère un tableau HTML.

```
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>
```

Navigateur de sortie du nœud Rapport

Ce navigateur affiche le contenu du rapport généré. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu Fichier, et celles d'édition dans le menu Edition. Pour plus d'informations, reportez-vous à la section Affichage des sorties sur p. 399.

Figure 6-46
Navigateur du nœud Rapport

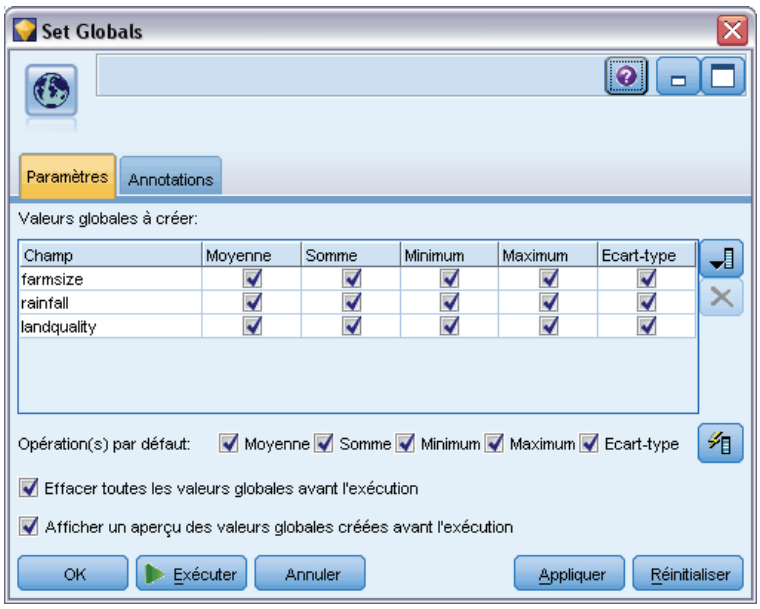


Noeud V. globales (Valeurs globales)

Le nœud V. globales (Valeurs globales) analyse les données et calcule des valeurs récapitulatives pouvant être utilisées dans des expressions CLEM. Par exemple, vous pouvez utiliser un noeud V. globales (Valeurs globales) pour calculer les statistiques d'un champ *âge*, puis utiliser la moyenne globale de ce champ dans des expressions CLEM en insérant la fonction @GLOBAL_MEAN(age).

Noeud V. globales (Valeurs globales) - Onglet Paramètres

Figure 6-47
Noeud V. globales (Valeurs globales) : onglet Paramètres



Valeurs globales à créer. Sélectionnez les champs pour lesquels vous souhaitez que des valeurs globales soient disponibles. Vous pouvez sélectionner plusieurs champs. Pour chaque champ, indiquez les statistiques à calculer en les sélectionnant dans les colonnes en regard du nom du champ.

- **Moyenne.** Indique la valeur moyenne du champ sur l'ensemble des enregistrements.
- **Somme.** Indique la somme des valeurs du champ sur l'ensemble des enregistrements.
- **Minimum.** Indique la valeur minimale du champ.
- **Maximum.** Indique la valeur maximale du champ.
- **Ecart-type.** L'écart-type est une mesure de variabilité des valeurs d'un champ ; il correspond à la racine carrée de la variance.

Opération(s) par défaut. Les options sélectionnées ici sont utilisées lorsque de nouveaux champs sont ajoutés à la liste des valeurs globales ci-dessus. Pour modifier l'ensemble de statistiques par défaut, sélectionnez ou désélectionnez les statistiques de votre choix. Vous pouvez également utiliser le bouton Appliquer pour appliquer les options par défaut à tous les champs de la liste.

Effacer toutes les valeurs globales avant l'exécution. Cette option permet de supprimer toutes les valeurs globales avant d'en calculer de nouvelles. Si cette option n'est pas sélectionnée, les valeurs recalculées remplacent les anciennes, mais les valeurs globales non recalculées restent disponibles.

Afficher un aperçu des valeurs globales créées avant l'exécution. Si vous sélectionnez cette option, l'onglet Valeurs globales de la boîte de dialogue des propriétés du flux apparaît à la fin de l'exécution pour afficher les valeurs globales calculées.

Programmes externes de IBM SPSS Statistics

Si une version compatible de IBM® SPSS® Statistics est installée sur votre ordinateur et que vous disposez d'une version sous licence, vous pouvez configurer IBM® SPSS® Modeler pour traiter les données avec la fonctionnalité SPSS Statistics à l'aide des noeuds Transformation Statistics, Modèle Statistics, Sortie Statistics, ou Export Statistics.

▶ Pour configurer SPSS Modeler afin de l'utiliser avec SPSS Statistics et d'autres applications, choisissez :

Outils > Options > Programmes externes

IBM SPSS Statistics Interactive. Saisissez le chemin d'accès et le nom complet de la commande (par exemple, *C:\Program Files\IBM\SPSS\Statistics\<nn\stats.exe*) à utiliser lorsque vous lancez SPSS Statistics directement sur un fichier de données produit par le noeud Exporter Statistics. Pour plus d'informations, reportez-vous à la section Noeud Exporter Statistics dans le chapitre 8 sur p. 511.

Connexion. Si le serveur SPSS Statistics est situé sur le même ordinateur que IBM® SPSS® Modeler Server, vous pouvez, à des fins d'efficacité, activer une connexion entre les deux applications. Lors de l'analyse, les données restent ainsi sur le serveur. Sélectionnez Serveur pour activer l'option Port ci-après. L'option par défaut est Local.

Port. Spécifiez le port du serveur SPSS Statistics.

Utilitaire d'emplacement de licence IBM SPSS Statistics. Pour permettre à SPSS Modeler d'utiliser les nœuds Transformation Statistics, Modèle Statistics et Sortie Statistics, vous devez disposer d'une copie de SPSS Statistics installée avec licence sur l'ordinateur où le flux est exécuté. De plus, dans le cas d'une exécution en mode réparti avec un serveur SPSS Modeler Server distant, il vous faut également une copie du client SPSS Statistics installée sur l'ordinateur client SPSS Modeler et pour laquelle vous devez disposer d'une licence.

- Si SPSS Modeler est exécuté en mode local (autonome), la copie avec licence de SPSS Statistics doit se trouver sur l'ordinateur local. Cliquez sur ce bouton pour indiquer l'emplacement de l'installation SPSS Statistics locale que vous souhaitez utiliser pour la licence.
- De plus, si l'exécution s'effectue en mode réparti avec un serveur SPSS Modeler Server distant, la version avec licence de SPSS Statistics doit également se trouver sur l'ordinateur serveur et la configuration de la licence doit être effectuée sur le serveur. Pour cela, à partir de l'invite de commande, passez au répertoire *bin* de SPSS Modeler Server et sur Windows, exécutez :

```
statisticsutility -location =<chemin vers le fichier de licence IBM SPSS Statistics Server>/bin
```

Sur Unix, exécutez :

```
./statisticsutility -location =<chemin vers le fichier de licence IBM SPSS Statistics Server>/bin
```

Où *<chemin vers le fichier de licence SPSS Statistics Server>* est le répertoire d'installation d'un serveur SPSS Statistics sous licence.

Si vous ne possédez pas de copie sous licence de SPSS Statistics sur votre ordinateur local, vous pouvez quand même exécuter le noeud Statistics sur un serveur SPSS Statistics sous licence, mais essayer d'exécuter d'autres noeuds de SPSS Statistics fera apparaître un message d'erreur.

Commentaires

Si vous ne parvenez pas à exécuter correctement un noeud Commande SPSS Statistics, suivez les conseils ci-dessous :

- Si les noms de champ utilisés dans SPSS Modeler dépassent huit caractères (pour les versions antérieures à SPSS Statistics 12.0), 64 caractères (pour SPSS Statistics 12.0 et les versions suivantes) ou contiennent des caractères incorrects, il est nécessaire de les renommer ou de les tronquer avant de les lire dans SPSS Statistics. Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics dans le chapitre 8 sur p. 513.
- Si SPSS Statistics a été installé après SPSS Modeler, il vous faudra indiquer l'emplacement de la licence SPSS Statistics. Voir ci-dessus.

Chapitre
7

Noeuds d'exportation

Présentation des noeuds d'exportation

Les noeuds d'exportation permettent d'exporter les données dans divers formats, afin de pouvoir les utiliser avec d'autres logiciels.

Les noeuds d'exportation disponibles sont les suivants :

Le noeud Export SGBD écrit des données dans une source de données relationnelles compatible ODBC. Pour que cette opération puisse être effectuée, la source de données ODBC doit exister et vous devez y avoir accès en écriture. Pour plus d'informations, reportez-vous à la section Noeud Export SGBD sur p. 460.

L'exportation à l'aide d'un fichier plat génère des données dans un fichier texte délimité. Elles peuvent ainsi être lues par d'autres logiciels d'analyse ou par des tableurs. Pour plus d'informations, reportez-vous à la section noeud Export Fichier plat sur p. 481.

Le noeud Exporter Statistics génère des données au format IBM® SPSS® Statistics.*sav*. Les fichiers *.sav* peuvent être lus par SPSS Statistics Base et d'autres produits . Ce format est également utilisé pour les fichiers cache IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section Noeud Exporter Statistics dans le chapitre 8 sur p. 511.

Le noeud Export IBM® SPSS® Data Collection génère des données au format utilisé par les logiciels d'étude de marché Data Collection. Pour pouvoir utiliser ce noeud, vous devez avoir installé avant la bibliothèque de données Data Collection. Pour plus d'informations, reportez-vous à la section Noeud d'exportation IBM SPSS Data Collection sur p. 483.

Le noeud Export SAS permet d'obtenir des données de sortie au format SAS afin qu'elles puissent être lues par SAS ou par un logiciel compatible. Trois formats de fichier SAS sont disponibles : SAS pour Windows/OS2, SAS pour UNIX ou SAS version 7/8. Pour plus d'informations, reportez-vous à la section Noeud Export SAS sur p. 489.

Le noeud Export Excel génère une sortie de données au format Microsoft Excel (*.xls*). Si vous le souhaitez, vous pouvez choisir de lancer Excel automatiquement et d'ouvrir le fichier exporté lors de l'exécution du noeud. Pour plus d'informations, reportez-vous à la section Noeud Export Excel sur p. 490.

Le noeud Export XML génère une sortie de données dans un fichier au format XML. Vous pouvez également créer un noeud source XML pour lire de nouveau les données exportées dans le flux. Pour plus d'informations, reportez-vous à la section Noeud Export XML sur p. 491.

© Copyright IBM Corporation 1994, 2012.

Noeud Export SGBD

Vous pouvez utiliser les noeuds SGBD pour écrire des données à des bases de données relationnelles compatibles ODBC, qui sont explicitées dans le noeud SGBD source. Pour plus d'informations, reportez-vous à la section Nœud Source de base de données dans le chapitre 2 sur p. 15.

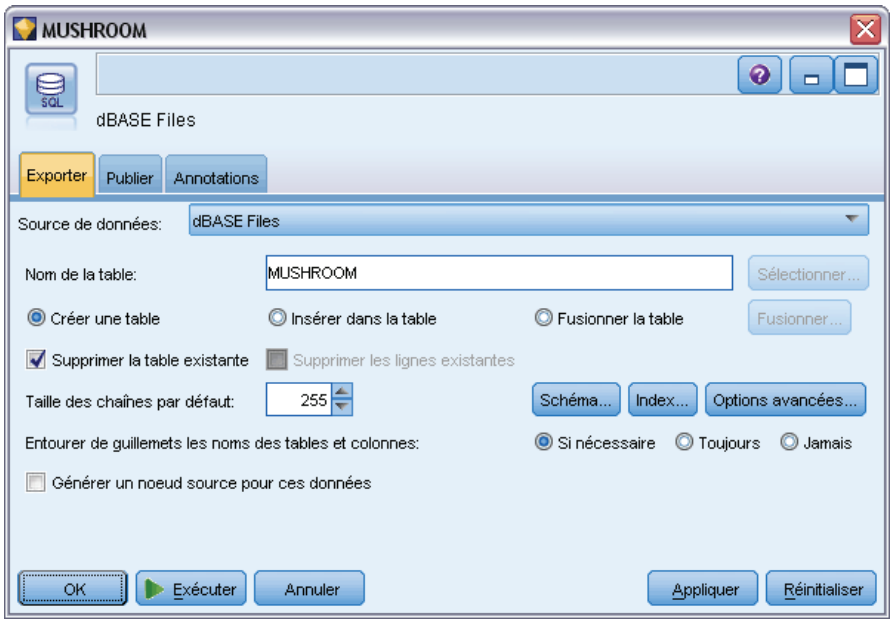
Pour écrire des données dans une base de données, utilisez la procédure générale suivante :

- Installez un pilote ODBC et configurez une source de données pour la base de données à utiliser.
- Dans l'onglet Exporter du noeud SGBD, indiquez la source de données et la table où écrire. Vous pouvez créer une table ou insérer des données dans une table existante.
- Indiquez d'autres options si nécessaire.

Cette procédure est détaillée dans les rubriques suivantes.

Noeud SGBD - Onglet Exporter

Figure 7-1
Noeud Export SGBD - Onglet Exporter



Source de données. Source de données sélectionnée. Entrez directement son nom ou sélectionnez-le dans la liste déroulante. Si la base de données souhaitée n'apparaît pas dans la liste, sélectionnez Ajouter une nouvelle connexion à la base de données et localisez votre base de données dans la boîte de dialogue Connexions de base de données. Pour plus d'informations, reportez-vous à la section Ajout d'une connexion à la base de données dans le chapitre 2 sur p. 18.

Nom de la table. Entrez le nom de la table vers laquelle envoyer les données. Si vous sélectionnez l'option Insérer dans la table, vous pouvez choisir une table existante dans la base de données en cliquant sur le bouton Sélectionner.

Créer une table. Sélectionnez cette option pour créer une nouvelle table de base de données ou écraser une table de base de données existante.

Insérer dans la table. Sélectionnez cette option pour insérer les données dans de nouvelles lignes d'une table de base de données existante.

Fusionner la table. (Le cas échéant) Sélectionnez cette option pour mettre à jour les colonnes de la base de données sélectionnées avec des valeurs de champs de données source correspondants. Sélectionner cette option active le bouton Fusionner, qui affiche une boîte de dialogue dans laquelle vous pouvez mapper les champs de données source sur les colonnes de la base de données.

Supprimer la table existante. Sélectionnez cette option pour supprimer, le cas échéant, une table existante du même nom que la table créée.

Supprimer les lignes existantes. Sélectionnez cette option pour supprimer les lignes existantes de la table avant l'exportation, lors de l'insertion dans une table.

Remarque : si vous sélectionnez l'une des deux options ci-dessus, vous recevez le message Avertissement d'écrasement lors de l'exécution du noeud. Pour que ces avertissements n'apparaissent plus, désélectionnez l'option Avertir lorsqu'un noeud écrase une table de base de données dans l'onglet Notifications de la boîte de dialogue Options utilisateur.

Taille des chaînes par défaut. Les champs marqués comme étant « sans type » dans un noeud Typer en amont sont écrits dans la base de données sous forme de champs de type chaîne. Indiquez la taille des chaînes à utiliser pour les champs sans type.

Cliquez sur Schéma pour ouvrir une boîte de dialogue dans laquelle vous pouvez définir diverses options d'exportation (pour les bases de données prenant en charge cette fonctionnalité), définir des types de données SQL pour vos champs et spécifier la clé primaire en vue de l'indexation de base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de la boîte de dialogue Schéma sur p. 464.

Cliquez sur Index pour définir les options d'indexation de la table exportée, afin d'améliorer les performances de la base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de l'index sur p. 468.

Cliquez sur Options avancées pour spécifier les options de chargement en masse et de validation de base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options avancées sur p. 470.

Entourer de guillemets les noms des tables et colonnes. Sélectionnez les options à utiliser lors de l'envoi d'une instruction CREATE TABLE à la base de données. Les tableaux ou colonnes comportant des espaces ou des caractères spéciaux doivent être mis entre guillemets.

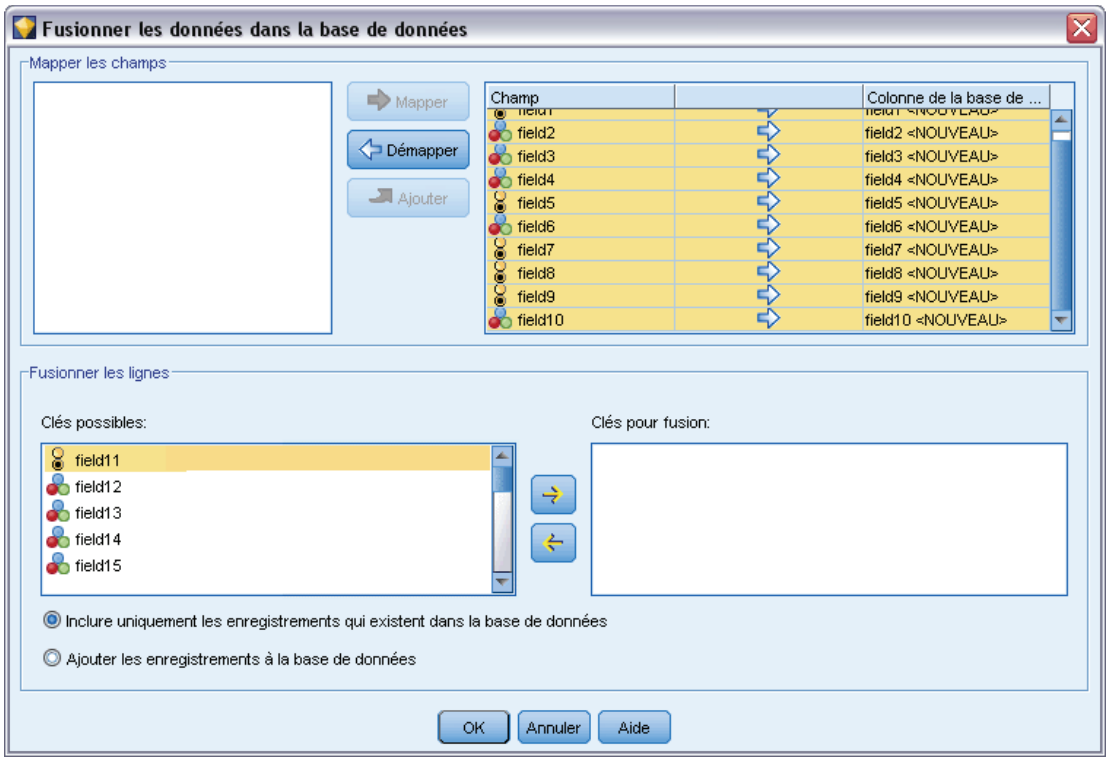
- **Si nécessaire.** Sélectionnez cette option pour que IBM® SPSS® Modeler détermine automatiquement, au cas par cas, la nécessité d'utiliser des guillemets.
- **Toujours.** Sélectionnez cette option pour que les noms de tableau et de colonne soient systématiquement mis entre guillemets.
- **Jamais.** Sélectionnez cette option pour désactiver l'utilisation des guillemets.

Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source SGBD pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès l'exécution, ce noeud est ajouté à l'espace de travail de flux.

Export SGBD - Options de fusion

Cette boîte de dialogue vous permet de mapper des champs à partir des données source dans des colonnes du tableau de la base de données cible. Lorsqu'un champ de données source est mappé sur une colonne de la base de données, la valeur de cette colonne est remplacée par la valeur des données source lorsque le flux est exécuté. Les champs source non mappés ne sont pas modifiés dans la base de données.

Figure 7-2
Mappage des champs de données source sur les colonnes de la base de données



Mapper les champs. C'est ici que vous indiquez le mappage entre les champs de données source et les colonnes de la base de données. Les champs de données source avec le même nom que les colonnes de la base de données sont automatiquement mappés.

- **Mapper.** Mappe un champ de données source sélectionné dans la liste des champs à la gauche du bouton sur une colonne de la base de données sélectionnée dans la liste de droite. Vous pouvez mapper plusieurs champs en même temps mais le nombre d'entrées sélectionnées dans les deux listes doit être le même.
- **Démapper.** Supprime le mappage pour une ou plusieurs colonnes sélectionnées de la base de données. Ce bouton est activé lorsque vous sélectionnez un champ ou une colonne de base de données dans la table située à droite de la boîte de dialogue.
- **Ajouter.** Ajoute un ou plusieurs champs de données source sélectionnés dans la liste des champs à gauche du bouton, à la liste de droite prête pour le mappage. Ce bouton est activé lorsque vous sélectionnez un champ dans la liste de gauche et qu'aucun champ portant ce nom n'existe dans la liste de droite. En cliquant sur ce bouton, vous mappez le champ sélectionné

sur une nouvelle colonne de la base de données portant le même nom. Le mot <NEW> est affiché après le nom de la colonne de base de données pour indiquer que le champ est nouveau.

Fusionner les lignes. Vous utilisez un champ-clé, tel que *ID transaction*, pour fusionner les enregistrements ayant une valeur identique dans ce champ. Cette opération est équivalente à une « équi-jointure » de base de données. Les valeurs des clés doivent être celles des clés principales ; c'est-à-dire qu'elles doivent être uniques et ne peuvent pas contenir de valeurs nulles.

- **Clés possibles.** Affiche tous les champs trouvés dans toutes les sources de données d'entrée. Sélectionnez un ou plusieurs champs de cette liste et utilisez la flèche pour les ajouter comme champs-clés pour la fusion des enregistrements. Tout champ de mappage avec une colonne de base de données mappée correspondante est disponible comme champ-clé, sauf que les champs ajoutés en tant que nouvelles colonnes de la base de données (indiqués par un <NEW> après le nom) ne sont pas disponibles.
- **Clés pour fusion.** Affiche tous les champs utilisés pour fusionner les enregistrements de toutes les sources de données d'entrée, sur la base des valeurs des champs-clés. Pour supprimer une clé de la liste, sélectionnez-la et utilisez la flèche pour la renvoyer dans la liste Clés possibles. Lorsque plusieurs champs-clés sont sélectionnés, l'option ci-dessous est activée.
- **Inclure uniquement les enregistrements qui existent dans la base de données.** Effectue une jointure partielle ; si l'enregistrement se trouve dans la base de données et dans le flux, les champs mappés seront mis à jour.
- **Ajouter les enregistrements à la base de données.** Effectue une jointure externe ; tous les enregistrements dans le flux seront fusionnés (si le même enregistrement existe dans la base de données) ou ajoutés (si l'enregistrement n'existe pas encore dans la base de données).

Pour mapper un champ de données source sur une nouvelle colonne de la base de données

▶ Cliquez sur le nom du champ source dans la liste de gauche, sous Mapper les champs.

▶ Cliquez sur le bouton Ajouter pour terminer le mappage.

Pour mapper un champ de données source sur une colonne existante de la base de données

▶ Cliquez sur le nom du champ source dans la liste de gauche, sous Mapper les champs.

▶ Cliquez sur le nom de la colonne sous Colonne de la base de données à droite.

▶ Cliquez sur le bouton Mapper pour terminer le mappage.

Pour supprimer un mappage

▶ Dans la liste de droite, dans Champ, cliquez sur le nom du champ dont vous souhaitez supprimer le mappage.

▶ Cliquez sur le bouton Démapper.

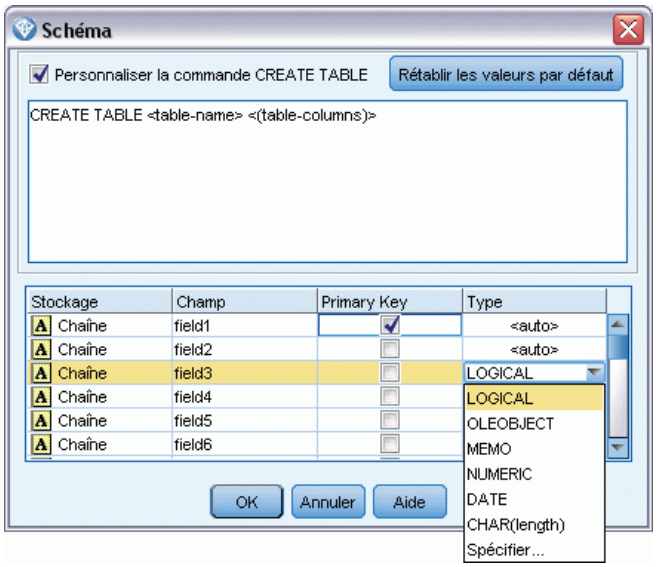
Pour annuler la sélection d'un champ de l'une des listes

▶ Maintenez enfoncée la touche CTRL et cliquez sur le nom du champ.

Export SGBD - Options de la boîte de dialogue Schéma

Dans la boîte de dialogue du schéma d'exportation de base de données, vous pouvez définir des options d'exportation de base de données (pour les bases de données prenant en charge ces options), définir des types de données SQL pour vos champs, indiquer les champs qui constituent des clés primaires et personnaliser l'instruction CREATE TABLE générée lors de l'exportation.

Figure 7-3
Exemple de boîte de dialogue Schéma d'exportation de base de données



Cette boîte de dialogue comprend plusieurs parties :

- La partie supérieure (si elle est visible) contient des options d'exportation vers une base de données prenant en charge ces options. Cette section n'apparaît pas si vous n'êtes pas connecté à une base de données de cette catégorie.
- Le champ de texte, dans la partie centrale, affiche le modèle utilisé pour générer la commande CREATE TABLE, qui suit par défaut le format ci-après :

  CREATE TABLE <table-name> <(table columns)>
- Le tableau, dans la partie inférieure, vous permet de définir le type de données SQL de chaque champ et d'indiquer les champs qui constituent des clés primaires, comme indiqué ci-dessous. La boîte de dialogue génère automatiquement les valeurs des paramètres <table-name> et <(table columns)> à partir des spécifications figurant dans le tableau.

Définition des options d'exportation de base de données

Si cette section apparaît, vous pouvez spécifier un certain nombre de paramètres pour l'exportation vers la base de données. Les bases de donnés prenant en charge cette fonctionnalité sont les suivantes.

- IBM InfoSphere Warehouse s'exécutant sur DB2 9.1 ou une version ultérieure. Pour plus d'informations, reportez-vous à la section Options pour IBM DB2 InfoSphere Warehouse sur p. 466.

- SQL Server 2008 ou des éditions Enterprise et Developer ultérieures. Pour plus d'informations, reportez-vous à la section Options pour SQL Server sur p. 466.
- Oracle 10g et 11gR1 ou des éditions Enterprise ou Personal ultérieures. Pour plus d'informations, reportez-vous à la section Options pour Oracle sur p. 466.

Personnalisation d'instructions CREATE TABLE

A l'aide du champ de texte de cette boîte de dialogue, vous pouvez ajouter d'autres options spécifiques aux bases de données à l'instruction CREATE TABLE.

▶ Cochez la case Personnaliser la commande CREATE TABLE pour activer la fenêtre de texte.

▶ Ajoutez des options de base de données à l'instruction. Veillez à conserver les paramètres de texte <table-name> et (<table-columns>), puisque IBM® SPSS® Modeler les remplace ensuite par les définitions de nom et de colonne réelles de la table.

Définition de types de données SQL

Par défaut, SPSS Modeler permet au serveur de base de données d'affecter automatiquement des types de données SQL. Pour remplacer le type affecté automatiquement à un champ, recherchez la ligne correspondant au champ et sélectionnez le type voulu dans la liste déroulante de la colonne *Type* du tableau de la boîte de dialogue Schéma. Vous pouvez utiliser Maj-clic pour sélectionner plusieurs lignes.

Dans le cas de types dotés d'un argument de longueur, de précision ou d'échelle (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC, and NUMBER), il vaut mieux spécifier une longueur plutôt que de laisser le serveur de base de données définir une longueur automatiquement. Par exemple, si vous spécifiez une valeur probable, comme VARCHAR(25), pour la longueur, le type de stockage dans SPSS Modeler sera écrasé si telle est votre intention. Pour remplacer l'affectation automatique, sélectionnez Spécifier dans la liste déroulante Type et remplacez la définition du type par l'instruction de définition de type SQL souhaitée.

Figure 7-4
Boîte de dialogue Spécifier le type du noeud de sortie SGBD

La méthode la plus simple consiste à sélectionner d'abord le type le plus proche de la définition souhaitée, puis à choisir Spécifier afin de modifier cette définition. Par exemple, pour définir le type de données SQL sur VARCHAR(25), paramétrez d'abord le type sur VARCHAR(length) dans la liste déroulante Type, puis sélectionnez Spécifier et remplacez la longueur de texte par la valeur 25.

Clés primaires

Si une ou plusieurs colonnes de la table exportée doivent comporter une valeur ou une combinaison de valeurs unique pour chaque ligne, vous pouvez l'indiquer en cochant la case Clé primaire correspondant à chaque champ concerné. La plupart des bases de données n'autorisent aucune modification de la table qui invaliderait une contrainte de clé primaire et créent automatiquement un index en fonction de la clé primaire pour appliquer cette restriction. (Si vous le souhaitez, vous pouvez créer des index pour d'autres champs dans la boîte de dialogue Index. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de l'index sur p. 468.)

Options pour IBM DB2 InfoSphere Warehouse

Espace Table. L'espace Table utilisé pour l'exportation. Les administrateurs de la base de données peuvent créer ou configurer des espaces tables partitionnés. Nous vous recommandons de sélectionner un de ces espaces tables (plutôt que celui par défaut) pour l'exportation vers la base de données.

Partition des données par champ. Spécifie le champ d'entrée à utiliser pour la partition.

Utiliser la compression. Si cette option est sélectionnée, elle crée des tables compressées pour l'exportation (par exemple, l'équivalent de CREATE TABLE MYTABLE(…) COMPRESS YES; en SQL).

Options pour SQL Server

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Ligne.** Active la compression au niveau des lignes (par exemple, l'équivalent de CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW); en SQL).
- **Page.** Active la compression au niveau des pages (par exemple, CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE); en SQL).

Options pour Oracle

Paramètres d'Oracle 10g

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées. Pour cette version de la base de données, seule la compression basique est disponible (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL).

Paramètres d'Oracle 11gR1

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Opérations de chargement direct.
- **Opérations de chargement direct.** Active la compression des opérations d'insertion en masse (directes) uniquement (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR DIRECT_LOAD OPERATIONS; en SQL).
- **Toutes les opérations.** Active la compression de toutes les opérations (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR ALL OPERATIONS; en SQL).

Paramètres d'Oracle 11gR2 - option basique

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Basique.
- **Basique.** Active la compression basique (par exemple, CREATE TABLE MYTABLE(…) COMPRESS BASIC; en SQL).

Paramètres d'Oracle 11gR2 - option avancée

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

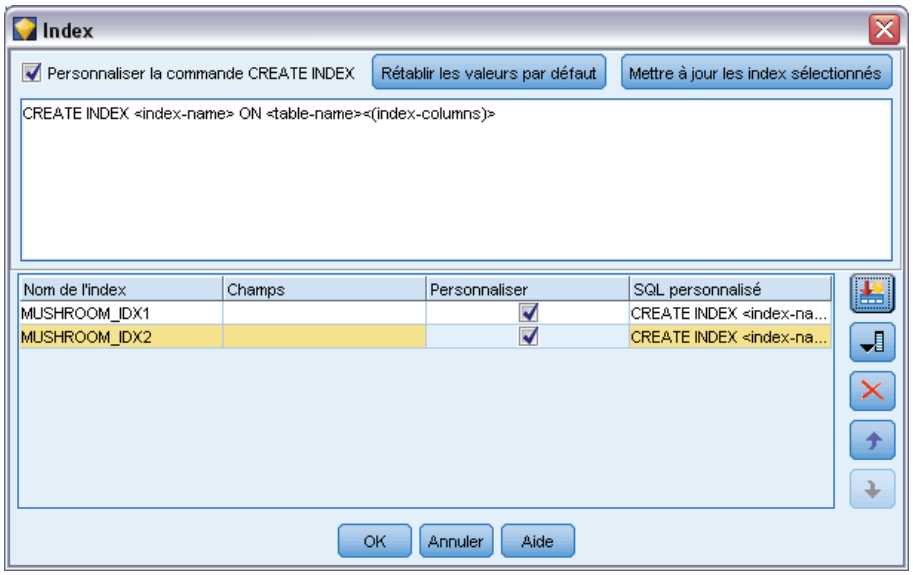
Compression pour. Choisissez le niveau de compression.

- **Défaut :** Active la compression par défaut (par exemple, CREATE TABLE MYTABLE(…) COMPRESS; en SQL). Dans ce cas, cela a le même effet que l'option Basique.
- **Basique.** Active la compression basique (par exemple, CREATE TABLE MYTABLE(…) COMPRESS BASIC; en SQL).
- **OLTP.** Active la compression OLTP (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP; en SQL).
- **Requête faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les requêtes (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW; or CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH; en SQL). La compression des requêtes est utile dans les environnements d'entreposage de données ; HIGH fournit un taux de compression plus grand que LOW.
- **Archive faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les archives (par exemple, CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW; ou CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH; en SQL). La compression des archives est utile pour compresser des données qui seront stockées pendant de longues périodes ; HIGH fournit un taux de compression plus grand que LOW.

Export SGBD - Options de l'index

La boîte de dialogue Index vous permet de créer des index sur les tables de base de données exportées depuis IBM® SPSS® Modeler. Vous pouvez indiquer les ensembles de champs à inclure et personnaliser la commande CREATE INDEX, selon les besoins.

Figure 7-5
Boîte de dialogue Index du noeud de sortie SGBD



Cette boîte de dialogue comprend deux parties :

- Le champ de texte figurant dans la partie supérieure affiche un modèle qui permet de générer une ou plusieurs commandes CREATE INDEX, qui suivent par défaut au format ci-après :

  CREATE INDEX <index-name> ON <table-name>

- Le tableau figurant dans la partie inférieure de la boîte de dialogue vous permet d'ajouter des spécifications pour chaque index à créer. Indiquez, pour chaque index, son nom ainsi que les champs ou les colonnes à inclure. La boîte de dialogue génère automatiquement les valeurs des paramètres <index-name> et <table-name> en conséquence.

  Par exemple, le code SQL généré pour un index simple réalisé sur les champs *empid* et *deptid* peut utiliser la syntaxe suivante :

  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)

  Vous pouvez ajouter plusieurs lignes pour créer plusieurs index. Une autre commande CREATE INDEX est générée pour chaque ligne.

Personnalisation de la commande CREATE INDEX

Si vous le souhaitez, vous pouvez personnaliser la commande CREATE INDEX pour tous les index ou pour un index précis uniquement. Vous disposez ainsi d'une marge de manoeuvre pour vous adapter à des options ou des exigences de base de données particulières et pour appliquer des personnalisations à tous les index ou à certains index uniquement, si nécessaire.

- Sélectionnez Personnaliser la commande CREATE INDEX en haut de la boîte de dialogue afin de modifier le modèle utilisé pour tous les index ajoutés dès à présent. Notez que ces modifications ne s'appliquent pas automatiquement aux index déjà ajoutés à la table.
- Sélectionnez une ou plusieurs lignes de la table, puis cliquez sur Mettre à jour les index sélectionnés en haut de la boîte de dialogue pour appliquer les personnalisations actuelles à toutes les lignes sélectionnées.
- Cochez la case Personnaliser sur chaque ligne pour modifier le modèle de commande de l'index uniquement.

La boîte de dialogue génère automatiquement les valeurs des paramètres <index-name> et <table-name> à partir des spécifications de la table ; ces valeurs ne peuvent pas être éditées directement.

Mot-clé BITMAP. Si vous utilisez une base de données Oracle, vous pouvez personnaliser le modèle afin de créer un index bitmap et non un index standard :

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

Les index bitmap peuvent s'avérer utiles pour indexer les colonnes contenant un nombre limité de valeurs distinctes. Le code SQL résultant peut être semblable à ce qui suit :

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

Mot-clé UNIQUE. La plupart des bases de données prennent en charge le mot clé UNIQUE dans la commande CREATE INDEX. Il est ainsi possible d'appliquer à la table sous-jacente une contrainte d'unicité semblable à une contrainte de clé primaire.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Pour les champs désignés comme clés primaires, cette spécification n'est pas nécessaire. La plupart des bases de données créent automatiquement un index pour les champs définis comme champs de clé primaire dans la commande CREATE TABLE. Par conséquent, la création explicite d'index sur ces champs est superflue. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de la boîte de dialogue Schéma sur p. 464.

Mot-clé FILLFACTOR. Certains paramètres physiques de l'index peuvent être affinés. Par exemple, SQL Server permet à l'utilisateur de compenser les coûts de maintenance par la taille de l'index (après sa création initiale), lors des modifications ultérieures apportées à la table.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Autres commentaires

- Si un index du même nom existe déjà, la création d'index échoue. Les échecs sont considérés initialement comme des avertissements (ce qui permet la création des index suivants), puis ils sont ensuite signalés comme des erreurs dans le journal des messages une fois que le système a essayé de créer tous les index.
- Pour optimiser les performances, les index doivent être créés une fois les données chargées dans la table. Les index doivent contenir au moins une colonne.

- Avant d'exécuter le noeud, vous pouvez prévisualiser le code SQL généré dans le journal des messages.
- Pour les tables temporaires écrites dans la base de données (c'est-à-dire lorsque la mise en cache des noeuds est activée), les options permettant de définir des clés primaires et des index ne sont pas disponibles. Toutefois, si cela s'avère nécessaire, le système peut créer des index à partir de la table temporaire en fonction du mode d'utilisation des données dans les noeuds en aval. Par exemple, si les données mises en cache sont ensuite liées par la colonne *DEPT*, il semble alors judicieux d'indexer sur cette colonne la table mise en cache.

Index et optimisation des requêtes

Dans certains systèmes de gestion de base de données, une fois la table de base de données créée, chargée et indexée, une autre étape est nécessaire pour que l'optimiseur puisse utiliser les index et accélérer l'exécution des requêtes sur la nouvelle table. Par exemple, dans Oracle, l'optimiseur de requêtes basé sur le coût exige qu'une table soit d'abord analysée avant que ses index puissent être utilisés pour l'optimisation des requêtes. Le fichier de propriétés ODBC interne pour Oracle (non visible par l'utilisateur) comporte une option pour que cela se produise, comme suit :

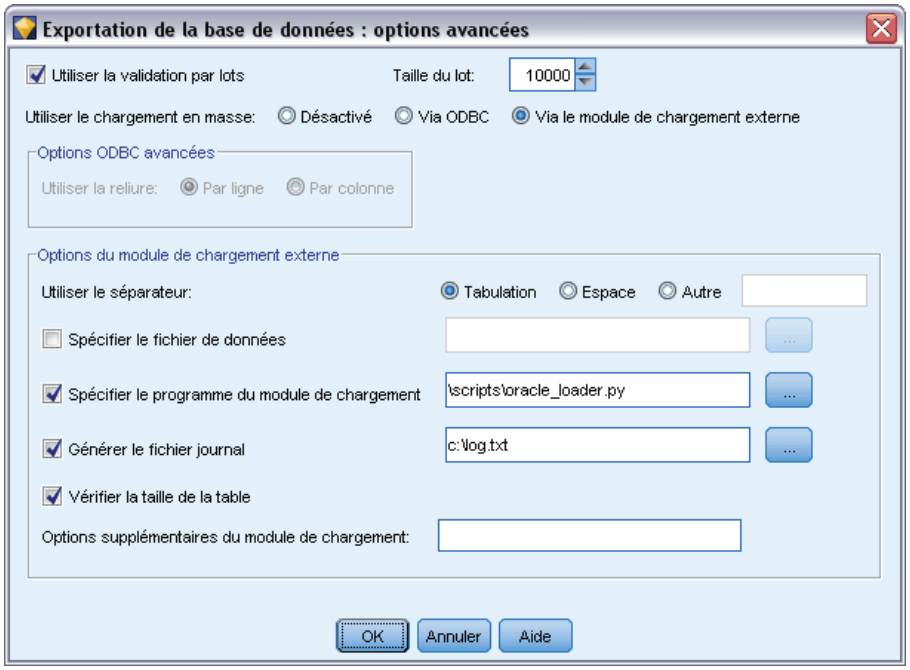
```
# Définit le code SQL à exécuter une fois qu'une table et les index associés
# ont été créés et renseignés
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

Cette étape est exécutée chaque fois qu'une table est créée dans Oracle (que des clés primaires ou des index soient définis ou non). Si nécessaire, le fichier de propriétés ODBC des bases de données supplémentaires peut être personnalisé de la même manière : contactez l'assistance technique.

Export SGBD - Options avancées

Lorsque vous cliquez sur Options avancées dans la boîte de dialogue du noeud d'exportation de la base de données, une nouvelle boîte de dialogue apparaît, qui vous permet de spécifier les détails techniques d'exportation des résultats dans une base de données.

Figure 7-6
Spécification des options avancées d'exportation dans une base de données



Utiliser la validation par lots. Sélectionnez cette option afin de désactiver la validation ligne à ligne dans la base de données.

Taille du lot. Indique le nombre d'enregistrements à envoyer à la base de données avant validation dans la mémoire. Si vous choisissez une valeur faible, l'intégrité des données est mieux préservée mais la vitesse de transfert moins rapide. Vous pouvez modifier cette valeur afin d'utiliser au mieux votre base de données.

Options de InfoSphere Warehouse. S'affiche uniquement si vous êtes connecté à une base de données InfoSphere Warehouse (IBM DB2 9.7 ou une version ultérieure). Ne pas enregistrer les mises à jour vous permet de ne pas enregistrer les événements lors de la création de tables et d'insertion de données.

Utiliser le chargement en masse. Spécifie la méthode de chargement en masse des données vers la base de données directement à partir de IBM® SPSS® Modeler. Vous devrez peut-être effectuer des tests pour sélectionner les options de chargement en masse adaptées à un scénario particulier.

- **Via ODBC.** Sélectionnez cette option afin d'utiliser l'API ODBC pour exécuter des insertions de plusieurs lignes. Cette méthode est plus efficace qu'une simple exportation vers la base de données. Parmi les options ci-après, optez pour un lien par ligne ou par colonne.
- **Via le module de chargement externe.** Sélectionnez cette option afin d'utiliser un programme de module de chargement en masse personnalisé propre à votre base de données. Les options ci-dessous sont alors automatiquement activées.

Options ODBC avancées. Ces options ne sont disponibles que lorsque vous sélectionnez Via ODBC. Notez que tous les pilotes ODBC ne prennent pas en charge ces fonctions.

- **Par ligne.** Sélectionnez le lien par ligne afin d'utiliser SQLBulkOperations pour charger les données dans la base de données. Le lien par ligne permet d'obtenir une vitesse plus importante que les insertions configurées qui insèrent les données de chaque enregistrement séparément.
- **Par colonne.** Sélectionnez cette option afin d'utiliser le lien par colonne pour charger les données dans la base de données. Le lien par colonne permet d'obtenir de meilleures performances car il relie chaque colonne de la base de données (dans une instruction INSERT configurée) à un ensemble de valeurs *N*. Si vous exécutez l'instruction INSERT une fois, *N* lignes sont insérées dans la base de données. Cette méthode permet d'obtenir de bien meilleures performances.

Options du module de chargement externe. Lorsque vous choisissez Via le module de chargement externe, de nombreuses options apparaissent : elles permettent d'exporter l'ensemble de données dans un fichier, et de spécifier et d'exécuter un programme de module de chargement personnalisé pour charger les données de ce fichier vers la base de données. SPSS Modeler fonctionne avec les chargeurs externes d'un grand nombre de systèmes de base de données connus. Plusieurs scripts ont été inclus dans le logiciel ; ils se trouvent, avec la documentation technique, dans le sous-répertoire *scripts*. Notez que pour utiliser cette fonctionnalité, Python 2.7 doit être installé sur le même ordinateur que SPSS Modeler ou IBM® SPSS® Modeler Server, et le paramètre python_exe_path doit être défini dans le fichier *options.cfg*. Pour plus d'informations, reportez-vous à la section Programmation de module de chargement en masse sur p. 473.

- **Utiliser le séparateur.** Spécifie le délimiteur à utiliser dans le fichier exporté. Sélectionnez Tabulation afin d'utiliser la tabulation comme délimiteur et Espace pour choisir l'espace. Sélectionnez Autre pour choisir un autre caractère, comme une virgule (,).
- **Spécifier le fichier de données.** Sélectionnez cette option afin de saisir l'emplacement de destination du fichier de données lors du chargement en masse. Par défaut, un fichier temporaire est créé dans le répertoire temporaire du serveur.
- **Spécifier le programme du module de chargement.** Sélectionnez cette option pour spécifier le programme de chargement en masse à utiliser. Par défaut, le logiciel recherche dans le sous-répertoire *scripts* du dossier d'installation SPSS Modeler, le script Python à exécuter pour une base de données spécifique. Plusieurs scripts ont été inclus dans le logiciel ; ils se trouvent, avec la documentation technique, dans le sous-répertoire *scripts*.
- **Générer le fichier journal.** Sélectionnez cette option afin de générer un fichier journal dans le répertoire spécifié. Ce fichier journal contient les informations relatives aux erreurs. Il est particulièrement utile en cas d'échec du chargement en masse.
- **Vérifier la taille de la table.** Sélectionnez cette option afin de vérifier les tables dans le but de garantir que l'augmentation de la taille des tables correspond au nombre de lignes exportées à partir de SPSS Modeler.
- **Options supplémentaires du module de chargement.** Spécifie les arguments supplémentaires servant au programme du module de chargement. Pour les arguments contenant des espaces, utilisez des guillemets doubles.

Pour utiliser des guillemets doubles dans les arguments facultatifs, accompagnez-les d'une barre oblique inverse. Par exemple, l'option spécifiée sous la forme -comment "This is a \"comment\"" contient à la fois le commutateur -comment et le commentaire lui-même sous la forme This is a "comment".

Vous pouvez utiliser une barre oblique inverse à condition de l'accompagner d'une autre barre oblique inverse. Par exemple, l'option spécifiée sous la forme -specialdir "C:\\Test Scripts\\" contient à la fois le commutateur -specialdir et le répertoire sous la forme C:\Test Scripts\.

Programmation de module de chargement en masse

Le noeud Export SGBD comporte des options de chargement en masse dans la boîte de dialogue Options avancées. Les programmes de module de chargement en masse permettent de charger les données d'un fichier texte dans une base de données.

L'option Utiliser le chargement en masse - Via le module de chargement externe configure l'application IBM® SPSS® Modeler de sorte qu'elle exécute les trois opérations suivantes :

- Création des tables de base de données requises.
- Exportation des données vers un fichier texte.
- Appel d'un programme de module de chargement en masse pour charger les données du fichier dans la table de base de données.

En général, le programme de module de chargement en masse ne correspond pas à l'utilitaire de chargement de base de données proprement dit (par exemple, l'utilitaire sqlldr d'Oracle) ; il s'agit en fait d'un petit script ou programme qui créé les arguments corrects et les fichiers auxiliaires propres aux bases de données (comme un fichier de contrôle), puis appelle l'utilitaire de chargement de base de données. Les sections suivantes vous expliquent comment éditer un module de chargement en masse existant.

Vous pouvez aussi écrire votre propre programme de chargement en masse. Pour plus d'informations, reportez-vous à la section Développement de programmes de chargement en masse sur p. 478.

Scripts destinés au chargement en masse.

SPSS Modeler est fourni avec plusieurs programmes de chargement en masse qui correspondent aux différentes bases de données implémentées utilisant les scripts Python. Lorsque vous exécutez un flux contenant un noeud Export SGBD et que l'option Via le module de chargement externe est sélectionnée, SPSS Modeler crée la table de base de données (si nécessaire) via ODBC, exporte les données vers un fichier temporaire sur l'hôte exécutant IBM® SPSS® Modeler Server, puis invoque le script de chargement en masse. Ensuite, ce script exécute des utilitaires fournis par le fournisseur SGBD afin de charger les données des fichiers temporaires vers la base de données.

Remarque : L'installation de SPSS Modeler ne comprend pas d'interpréteur d'exécution Python, par conséquent une installation distincte de Python est nécessaire. Pour plus d'informations, reportez-vous à la section Export SGBD - Options avancées sur p. 470.

Les scripts fournis (disponibles dans le dossier *\scripts* du répertoire d'installation SPSS Modeler) sont destinés aux bases de données suivantes.

Table 7-1
Scripts de chargement en masse fournis

SGBD	Nom du script	
IBM DB2	*db2_loader.py*	Pour plus d'informations, reportez-vous à la section Chargement en masse de données vers les bases de données IBM DB2 sur p. 474.
IBM Netezza	*netezza_loader.py*	Pour plus d'informations, reportez-vous à la section Chargement en masse de données vers les bases de données IBM Netezza sur p. 475.
Oracle	*oracle_loader.py*	Pour plus d'informations, reportez-vous à la section Chargement en masse de données vers les bases de données Oracle sur p. 476.
SQL Server	*mssql_loader.py*	Pour plus d'informations, reportez-vous à la section Chargement en masse de données vers les bases de données SQL Server sur p. 477.
Teradata	*teradata_loader.py*	Pour plus d'informations, reportez-vous à la section Chargement en masse de données vers les bases de données Teradata sur p. 477.

Chargement en masse de données vers les bases de données IBM DB2

Les points suivants peuvent vous aider à configurer le chargement en masse à partir de IBM® SPSS® Modeler vers une base de données IBM DB2 à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire du processeur de ligne de commande DB2 (CLP) est installé

Le script *db2_loader.py* invoque la commande DB2 LOAD. Vérifiez que le processeur de ligne de commande (*db2* sous UNIX, *db2cmd* sous Windows) est installé sur le serveur qui procède à l'exécution de *db2_loader.py* (généralement, l'hôte exécutant IBM® SPSS® Modeler Server).

Vérifiez que le nom d'alias de la base de données locale est le même que le nom réel de la base de données

Le nom d'alias de la base de données locale DB2 est le nom utilisé par le logiciel client DB2 pour faire référence à une base de données sur une instance DB2 locale ou distante. Si l'alias de la base de données locale est différent du nom de la base de données distante, utilisez l'option du module de chargement supplémentaire :

```
-alias <local_database_alias>
```

Par exemple, la base de données distante est nommée STARS sur l'hôte GALAXY mais l'alias de la base de données locale DB2 sur l'hôte exécutant SPSS Modeler Server est STARS_GALAXY. Utilisez l'option du module de chargement supplémentaire

```
-alias STARS_GALAXY
```

Codage des données de caractères non-ASCII

Si vous chargez en masse des données dont le format n'est pas ASCII, vous devez vous assurer que la variable codepage de la section de configuration de *db2_loader.py* est correctement définie sur votre système.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Chargement en masse de données vers les bases de données IBM Netezza

Les points suivants peuvent vous aider à configurer le chargement en masse à partir de IBM® SPSS® Modeler vers une base de données IBM Netezza à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Netezza nzload est installé

Le script *netezza_loader.py* invoque l'utilitaire Netezza *nzload*. Vérifiez que *nzload* est installé et correctement configuré sur le serveur qui va exécuter *netezza_loader.py*.

Exportation de données non-ASCII

Si votre exportation contient des données dont le format n'est pas ASCII, vous devrez peut-être ajouter -encoding UTF8 au champ Options supplémentaires du module de chargement de la boîte de dialogue Export SGBD - Options avancées. Ceci garantit que les données non-ASCII sont correctement chargées.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur JJ-MM-AAAA et le format d'heure sur HH:MM:SS.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Ordres des colonnes du flux et de la table cible différents lors de l'insertion de données dans une table existante

Si l'ordre des colonnes du flux est différent de celui de la table cible, les valeurs des données ne seront pas insérées dans les bonnes colonnes. Utilisez un noeud Re-trier pour garantir que l'ordre des colonnes du flux correspond à l'ordre de la table cible. Pour plus d'informations, reportez-vous à la section Noeud Re-trier dans le chapitre 4 sur p. 242.

Suivi de la progression de nzload

Lorsque SPSS Modeler est exécuté en mode local, ajoutez -sts au champ Options supplémentaires du module de chargement dans la boîte de dialogue Export SGBD - Option avancées, afin d'afficher des messages indiquant l'état de progression toutes les 1000 lignes dans la fenêtre de commande ouverte par l'utilitaire *nzload*.

Chargement en masse de données vers les bases de données Oracle

Les points suivants peuvent vous aider à configurer le chargement en masse à partir de IBM® SPSS® Modeler vers une base de données Oracle à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Oracle sqlldr est installé

Le script *oracle_loader.py* invoque l'utilitaire Oracle *sqlldr*. Remarque : l'utilitaire *sqlldr* n'est pas inclus de façon automatique dans le client Oracle. Vérifiez que *sqlldr* est installé sur le serveur qui va exécuter *oracle_loader.py*.

Indiquez le SID de la base de données ou le nom du service

Si vous exportez des données vers un serveur Oracle non local ou que votre serveur local Oracle comprend plusieurs bases de données, vous devrez spécifier les éléments suivants dans le champ Options supplémentaires du module de chargement situé dans la boîte de dialogue Export SGBD - Options avancées afin de transmettre le SID ou le nom du service :

-database <SID>

Modification de la section configuration dans oracle_loader.py

Sous UNIX (et éventuellement sous Windows), modifiez la section configuration située au début du script *oracle_loader.py*. Ici, les valeurs pour les variables d'environnement ORACLE_SID, NLS_LANG, TNS_ADMIN et ORACLE_HOME peuvent être spécifiées le cas échéant, de même que le chemin d'accès complet de l'utilitaire *sqlldr*.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur AAAA-MM-JJ et le format d'heure sur HH:MM:SS.

Si vous avez besoin d'utiliser un format de date et d'heure différent de celui indiqué ci-dessus, consultez votre documentation oracle et modifiez le fichier du script *oracle_loader.py*.

Codage des données de caractères non-ASCII

Si vous chargez en masse des données dont le format n'est pas ASCII, vous devez vous assurer que la variable d'environnement NLS_LANG est correctement définie sur votre système. Cette variable est lue par l'utilitaire Oracle de chargement *sqlldr*. Par exemple, la valeur

correcte de NLS_LANG pour Shift-JIS sous Windows est Japanese_Japan.JA16SJIS. Pour plus d'informations sur NLS_LANG, consultez votre documentation Oracle.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Chargement en masse de données vers les bases de données SQL Server

Les points suivants peuvent vous aider à configurer le chargement en masse à partir de IBM® SPSS® Modeler vers une base de données SQL Server à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire SQL Server bcp.exe est installé

Le script *mssql_loader.py* invoque l'utilitaire SQL Server *bcp.exe*. Vérifiez que *bcp.exe* est installé sur le serveur qui va exécuter *mssql_loader.py*.

L'utilisation d'espaces comme séparateur ne fonctionne pas

Evitez de choisir un espace comme séparateur dans la boîte de dialogue Export SGBD - Options avancées.

Option Vérifier la taille de la table recommandée

Nous vous recommandons d'activer l'option Vérifier la taille de la table dans la boîte de dialogue Export SGBD - Options avancées. Les échecs du processus de chargement en masse ne sont pas toujours détectés, et l'activation de cette option permet de procéder à une vérification supplémentaire indiquant si le nombre correct de lignes a été chargé.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Chargement en masse de données vers les bases de données Teradata

Les points suivants peuvent vous aider à configurer le chargement en masse à partir de IBM® SPSS® Modeler vers une base de données Teradata à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Teradata fastload est installé

Le script *teradata_loader.py* invoque l'utilitaire Teradata *fastload*. Vérifiez que *fastload* est installé et correctement configuré sur le serveur qui va exécuter *teradata_loader.py*.

Chargement en masse des donnés possible uniquement vers des tables vides

Seules des tables vides peuvent être utilisées comme cibles d'un chargement en masse. Si une table cible contient déjà des données avant le chargement en masse, l'opération échoue.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur AAAA-MM-JJ et le format d'heure sur HH:MM:SS.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

ID de processus Teradata (tdpid)

Par défaut, *fastload* exporte les données vers le système Teradata avec `tdpid=dbc`. Généralement, il existe une entrée dans le fichier HOSTS qui associe `dbccop1` à l'adresse IP du serveur Teradata. Pour utiliser un serveur différent, spécifiez les éléments suivants dans le champ Options supplémentaires du module de chargement de la boîte de dialogue Export SGBD - Options avancées, afin de transmettre le `tdpid` du serveur :

```
-tdpid <id>
```

Espaces dans les noms des tables et colonnes

Si les noms des tables et colonnes contiennent des espaces, l'opération de chargement en masse échoue. Si possible, renommez les tables ou colonnes pour supprimer les espaces.

Développement de programmes de chargement en masse

Cette section explique comment développer un programme de module de chargement en masse pouvant être exécuté dans IBM® SPSS® Modeler pour charger des données à partir d'un fichier texte vers une base de données.

Utilisation de Python pour créer des programmes de module de chargement en masse

Par défaut, SPSS Modeler recherche un programme de module de chargement en masse en fonction du type de la base de données. Consultez Table 7-1 sur p. 474.

Le script *test_loader.py* peut pour vous aider dans le développement de programmes de module de chargement en masse. Pour plus d'informations, reportez-vous à la section Test des programmes de module de chargement en masse sur p. 481.

Objets transmis au programme de module de chargement en masse

SPSS Modeler crée deux fichiers qui sont transmis au programme de module de chargement en masse.

- **Fichier de données.** Il contient les donnés à charger au format texte.
- **Fichier de schéma.** Ce fichier est au format XML. Il décrit les noms et types des colonnes et fournit des informations sur le format des données (par exemple, quel caractère sert de séparateur entre les champs).

En outre, SPSS Modeler transmet d'autres informations telles que le nom de la table, le nom et le mot de passe utilisateur sous forme d'arguments lors de l'invocation du programme de module de chargement en masse.

Remarque : pour indiquer la réussite de l'opération à SPSS Modeler, le programme de module de chargement en masse doit supprimer le fichier de schéma.

Arguments transmis au programme de module de chargement en masse

Les arguments transmis au programme sont les suivants.

Table 7-2
Arguments transmis au module de chargement en masse

Argument	Description
schemafile	Chemin du fichier de schéma.
data file	Chemin du fichier de données.
servername	Nom du serveur DBMS ; peut être vide.
databasename	Nom de la base de données sur le serveur DBMS ; peut être vide.
username	Nom d'utilisateur servant à la connexion à la base de données.
password	Mot de passe servant à la connexion à la base de données.
tablename	Nom de la table à charger.
ownername	Nom du propriétaire de la table (aussi appelé nom du schéma).
logfilename	Nom du fichier journal (s'il est laissé vide, aucun fichier journal n'est créé).
rowcount	Nombre de lignes dans l'ensemble de données.

Toutes les options spécifiées dans le champ Options supplémentaires du module de chargement de la boîte de dialogue Export SGBD - Options avancées, sont transmises au programme de module de chargement en masse après ces arguments standard.

Format du fichier de données.

Les données sont écrites dans le fichier de données au format texte, chaque champ étant séparé par un caractère de délimitation spécifié dans la boîte de dialogue Export SGBD - Options avancées. L'exemple suivant indique la manière dont un fichier de données séparé par des tabulations doit apparaître.

```
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
```

Le fichier est codé à l'aide du codage local utilisé par IBM® SPSS® Modeler Server (ou SPSS Modeler si le système n'est pas relié à SPSS Modeler Server). Une partie du format est contrôlée par les paramètres de flux SPSS Modeler.

Format du fichier de schéma.

Le fichier de schéma est un fichier XML qui décrit le fichier de données. Voici un exemple du fichier de schéma qui accompagnerait le fichier de données précédent.

```
<?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>
```

Les tableaux suivants répertorient les attributs des éléments <table> et <column> du fichier de schéma.

Table 7-3
Attributs de l'élément <table>

Attribut	Description
delimiter	Le caractère de délimitation des champs (TAB est représenté par \t).
commit_every	L'intervalle de taille du lot (tel qu'indiqué dans la boîte de dialogue Export SGBD - Options avancées).
date_format	Le format utilisé pour représenter les dates.
time_format	Le format utilisé pour représenter les heures.
append_existing	true si la table chargée contient déjà des données ; sinon false.
delete_datafile	true si le programme de chargement en masse doit supprimer le fichier de données après la fin du chargement.

Table 7-4
Attributs de l'élément <column>

Attribut	Description
name	Nom de la colonne.
encoded_name	Le nom de la colonne converti dans le même codage que celui du fichier de données et affiché sous la forme d'une série de nombres hexadécimaux à deux chiffres.
type	Le type de données de la colonne : parmi les types suivants :integer, real, char, time, date et datetime.
size	Pour le type de données char, la largeur maximale de la colonne en caractères.

Test des programmes de module de chargement en masse

Vous pouvez tester le chargement en masse à l'aide d'un script de test *test_loader.py* disponible dans le dossier *\scripts* du répertoire d'installation IBM® SPSS® Modeler. Il est utile de procéder à un test lors du développement, du débogage ou du dépannage de programmes de chargement en masse ou de scripts à utiliser avec SPSS Modeler.

Pour utiliser le script de test, suivez la procédure suivante.

- Exécutez le script *test_loader.py* pour copier les fichiers de schéma et de données vers les fichiers *schema.xml* et *data.txt*, et créez un fichier de commande Windows (*test.bat*).
- Modifiez le fichier *test.bat* pour sélectionner le programme de module de chargement en masse ou le script à tester.
- Exécutez *test.bat* depuis un shell de commande pour tester le programme de module de chargement en masse ou le script choisi.

Remarque : l'exécution de *test.bat* ne charge pas réellement les données vers la base de données.

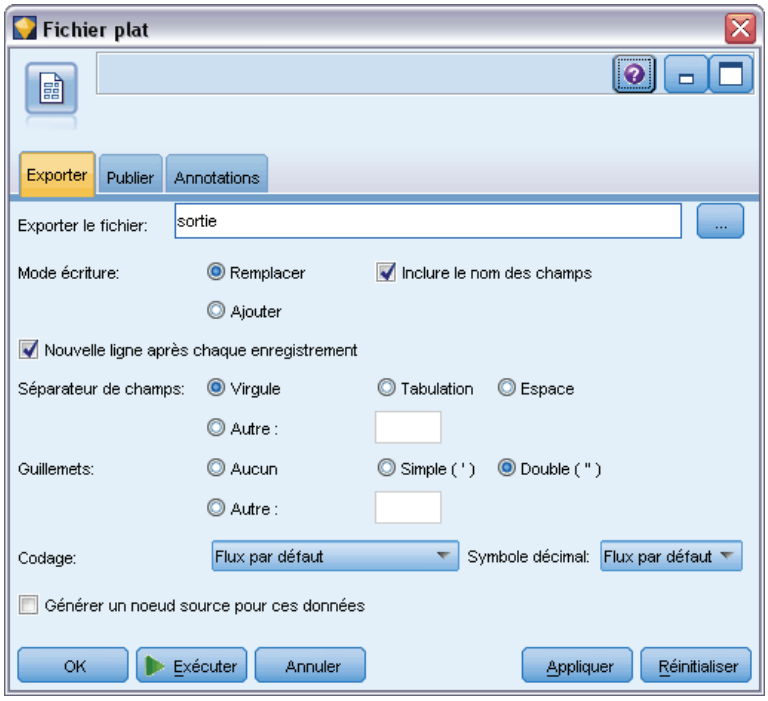
noeud Export Fichier plat

Le noeud Export Fichier plat permet d'écrire des données dans un fichier texte délimité. Elles peuvent ainsi être lues par d'autres logiciels d'analyse ou par des tableurs.

Remarque : vous ne pouvez pas écrire de fichiers dans l'ancien format cache, car IBM® SPSS® Modeler n'utilise plus ce format pour les fichiers cache. Les fichiers cache SPSS Modeler sont désormais enregistrés au format IBM® SPSS® Statistics*.sav*, à l'aide d'un noeud Export Statistics. Pour plus d'informations, reportez-vous à la section Noeud Exporter Statistics dans le chapitre 8 sur p. 511.

Noeud Fichier plat - Onglet Exporter

Figure 7-7
Noeud Fichier plat - Onglet Exporter



Exporter le fichier. Indique le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Mode écriture. Si vous sélectionnez Remplacer, les éventuelles données présentes dans le fichier indiqué seront écrasées. Si vous sélectionnez Ajouter, la sortie sera ajoutée à la fin du fichier et les données existantes conservées.

- **Inclure les noms des champs.** Si vous sélectionnez cette option, les noms des champs figurent sur la première ligne du fichier de sortie. Cette option est disponible uniquement avec le mode d'écriture Remplacer.

Nouvelle ligne après chaque enregistrement. Si vous sélectionnez cette option, chaque enregistrement est écrit sur une nouvelle ligne dans le fichier de sortie.

Séparateur de champs. Indique le caractère à insérer entre les valeurs des champs dans le fichier texte généré. Les options sont les suivantes : Virgule, Tabulation, Espace et Autre. Si vous sélectionnez Autre, entrez les caractères de séparation souhaités dans la zone de texte.

Guillemets. Indique le type de guillemet à utiliser pour les valeurs des champs symboliques. Les options disponibles sont les suivantes : Aucun (valeurs non accompagnées de guillemets), Simple (‘), Double (“) et Autre. Si vous sélectionnez Autre, entrez le type de guillemet souhaité dans la zone de texte.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal. Indique le mode de représentation des décimales dans les données.

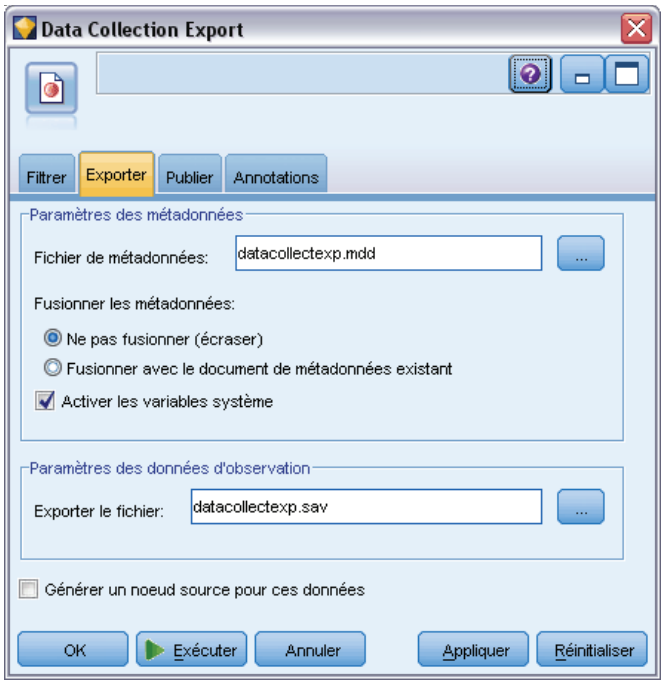
- **Flux par défaut.** Le séparateur décimal défini par défaut pour le flux actuel est utilisé. Il s'agit généralement du séparateur décimal défini dans les paramètres régionaux de l'ordinateur.
- **Point (.).** Le point est utilisé comme séparateur décimal.
- **Virgule (,).** La virgule est utilisée comme séparateur décimal.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source Délimité pour lire le fichier de données exporté. Pour plus d'informations, reportez-vous à la section Noeud Délimité dans le chapitre 2 sur p. 25.

Noeud d'exportation IBM SPSS Data Collection

Le noeud Export IBM® SPSS® Data Collection enregistre des données au format utilisé par les logiciels d'étude de marché Data Collection, en fonction du modèle Data Collection Data Model. Ce format fait la distinction entre les données d'observation (réponses réelles fournies à des questions et collectées au cours d'une enquête) et les métadonnées qui décrivent le mode de collecte et d'organisation des données d'observation. Les métadonnées consistent en des informations diverses : texte des questions, nom et description de variables, ensemble de réponses multiples, traduction des différents textes et définition de la structure des données d'observation. Pour plus d'informations, reportez-vous à la section Noeud Data Collection dans le chapitre 2 sur p. 36.

Figure 7-8
Noeud Export IBM SPSS Data Collection - Onglet Exporter



Remarque : Ce noeud requiert la version 4.0 ou supérieure de Data Collection Data Model, fournie avec les logiciels Data Collection. Pour plus d'informations, consultez la page Web Data Collection à l'adresse *http://www.ibm.com/software/analytics/spss/products/data-collection/*. Mise à part l'installation du modèle de données, aucune configuration supplémentaire n'est requise.

Fichier de métadonnées. Indique le nom du fichier de définition du questionnaire (*.mdd*) dans lequel les métadonnées exportées seront enregistrées. Un questionnaire par défaut est créé en fonction des informations de type de champ. Par exemple, un champ nominal (ensemble) peut être représenté sous la forme d'une question unique, avec la description du champ utilisée comme texte de la question et une case à cocher distincte pour chaque valeur définie.

Fusionner les métadonnées. Indique si les métadonnées remplaceront les versions existantes ou seront fusionnées avec les métadonnées existantes. Si l'option de fusion est sélectionnée, une nouvelle version est créée à chaque exécution du flux. Ceci permet le suivi des versions d'un questionnaire au fil des modifications. Chaque version peut être considérée comme un instantané des métadonnées utilisées pour collecter un ensemble précis de données d'observation.

Activer les variables système. Indique si les variables système sont incluses dans le fichier *.mdd* exporté. Il s'agit de variables telles que *Respondent.Serial*, *Respondent.Origin*, et *DataCollection.StartTime*.

Paramètres des données d'observation. Indique le fichier de données IBM® SPSS® Statistics (*.sav*) dans lequel les données d'observation sont exportées. Notez que toutes les restrictions sur les noms de variable et de valeur s'appliquent ici, vous pouvez avoir besoin de basculer vers l'onglet Filtrer et d'utiliser l'option "Renommer pour SPSS Statistics" du menu des options de filtrage pour corriger les caractères non valides dans les noms de champ.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un nœud source Data Collection pour lire le fichier de données exporté.

Ensembles de réponses multiples. Tous les ensembles de réponses multiples définis dans le flux seront automatiquement préservés lors de l'exportation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples dans le chapitre 4 sur p. 159.

Noeud Export IBM Cognos BI

Le noeud Export IBM Cognos BI permet d'exporter des données d'un flux IBM® SPSS® Modeler vers Cognos BI, au format UTF-8. Ainsi, Cognos BI peut utiliser des données transformées ou évaluées de SPSS Modeler. Par exemple, vous pouvez utiliser Cognos BI Report Studio pour créer un rapport basé sur les données exportées, qui contient les prévisions et les valeurs de confiance. Le rapport peut ensuite être enregistré sur le serveur Cognos BI et distribué aux utilisateurs de Cognos BI.

Remarque : Vous pouvez uniquement exporter des données relationnelles et pas de données OLAP.

Pour exporter des données vers Cognos BI, vous devez spécifier les paramètres suivants :

- Connexion Cognos - la connexion au serveur Cognos BI
- Connexion ODCB - la connexion au serveur de données Cognos que le serveur Cognos BI utilise

Dans la connexion Cognos, vous spécifiez une source de données Cognos à utiliser. Cette source de données doit utiliser la même connexion que la source de données ODBC.

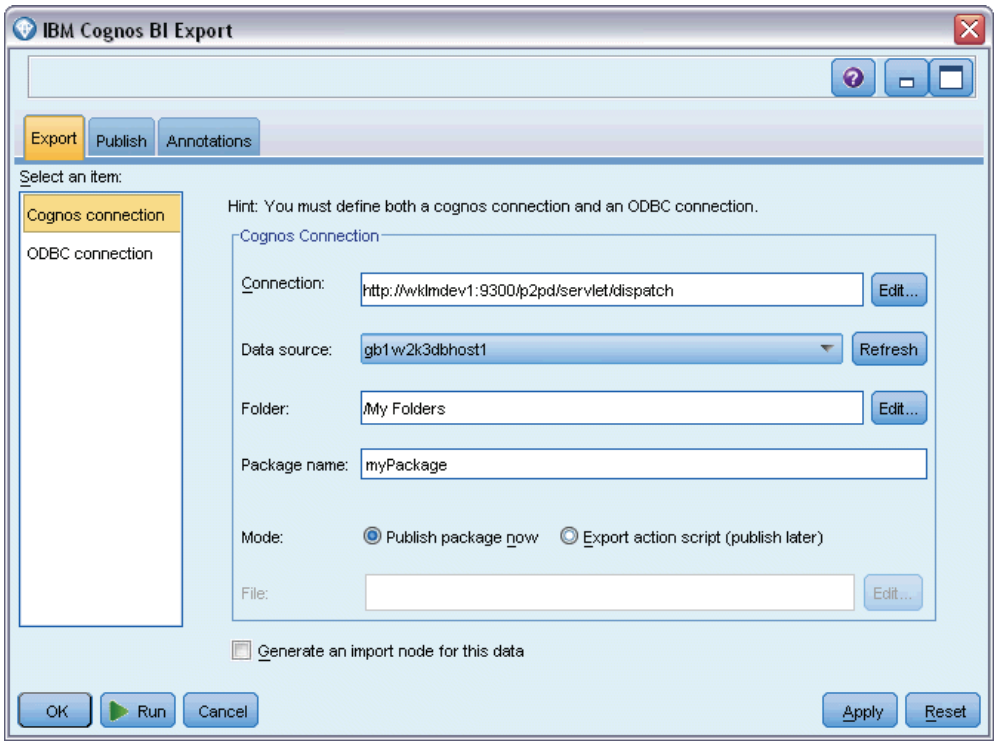
Vous exportez les données de flux vers le serveur de données et les métadonnées du package vers le serveur Cognos BI.

Comme avec n'importe quel autre noeud d'exportation, vous pouvez également utiliser l'onglet Publier de la boîte de dialogue du noeud pour publier le flux et le déployer à l'aide de IBM® SPSS® Modeler Solution Publisher.

Connexion Cognos

Vous spécifiez à cet endroit la connexion au serveur Cognos BI que vous souhaitez utiliser pour l'exportation. Cette procédure se compose de l'exportation des métadonnées vers un nouveau package sur le serveur Cognos BI pendant que les données de flux sont exportées vers le serveur de données Cognos.

Figure 7-9
Exportation des données Cognos



Connexion. Cliquez sur le bouton Modifier pour afficher une boîte de dialogue dans laquelle vous pourrez définir l'URL et les autres informations sur le serveur Cognos BI vers lequel vous souhaitez exporter les données. Si vous êtes déjà connecté à un serveur Cognos BI via IBM® SPSS® Modeler, vous pouvez également modifier les détails de la connexion actuelle. Pour plus d'informations, reportez-vous à la section Connexions Cognos dans le chapitre 2 sur p. 49.

Source de données. Le nom de la source de données Cognos (généralement une base de données) vers laquelle vous exportez les données. La liste déroulante indique toutes les sources de données Cognos auxquelles vous pouvez accéder à partir de la connexion actuelle. Cliquez sur le bouton Rafraîchir pour mettre à jour la liste.

Dossier. Le chemin d'accès et le nom du dossier du serveur Cognos BI où le package d'exportation doit être créé.

Nom du package. Le nom du package dans le dossier spécifié qui doit contenir les métadonnées exportées. Il doit s'agir d'un nouveau package avec un seul objet de requête ; l'exportation ne peut pas se faire vers un package existant.

Mode. Spécifie les modalités de l'exportation :

- **Publier le package maintenant.** (par défaut) Effectue l'exportation dès que vous cliquez sur Exécuter.
- **Exporter le script d'action.** Crée un script XML que vous pourrez exécuter ultérieurement (par exemple, à l'aide de Framework Manager) pour effectuer l'exportation. Saisissez le chemin d'accès et le nom du fichier du script dans le champ Fichier ou utilisez le bouton Modifier pour spécifier le nom et l'emplacement du fichier du script.

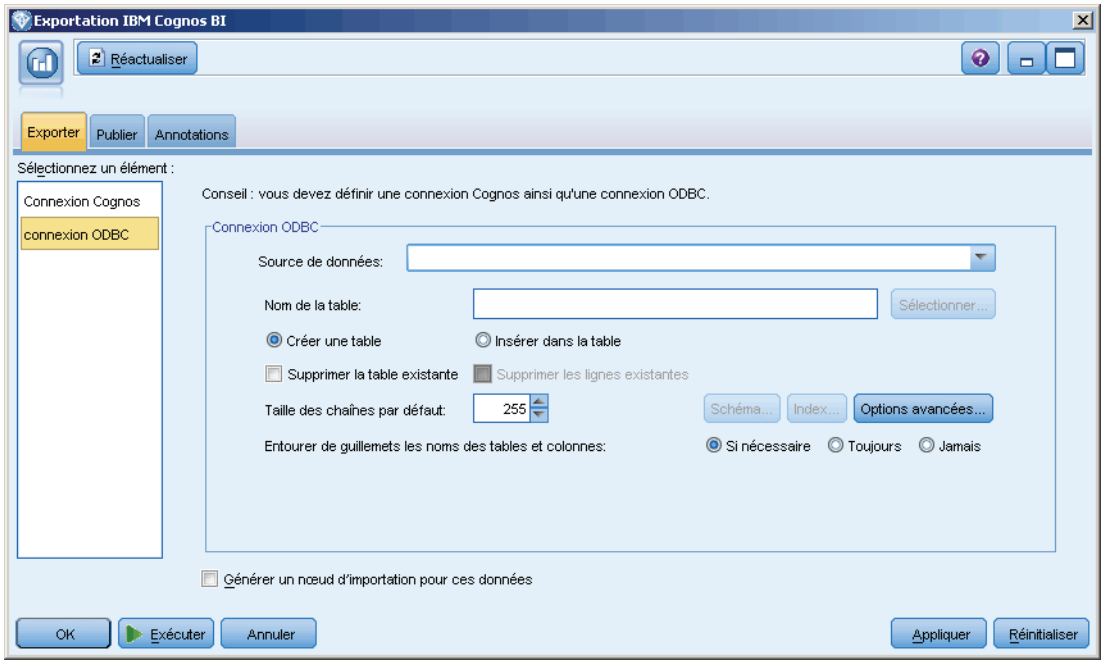
Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès que vous cliquez sur Exécuter, ce noeud est ajouté à l'espace de travail de flux.

connexion ODBC

C'est ici que vous spécifiez la connexion au serveur de données Cognos (c'est-à-dire la base de données) vers lequel les données de flux vont être exportées.

Remarque : Vous devez vous assurer que la source des données spécifiée ici pointe vers la même source que celle spécifiée dans le volet Connexions Cognos. Vous devez également vous assurer que la source de données de connexion Cognos utilise la même source de données que la source de données ODBC.

Figure 7-10
Exportation des données Cognos



Source de données. Source de données sélectionnée. Entrez directement son nom ou sélectionnez-le dans la liste déroulante. Si la base de données souhaitée n'apparaît pas dans la liste, sélectionnez Ajouter une nouvelle connexion à la base de données et localisez votre base de

données dans la boîte de dialogue Connexions de base de données. Pour plus d'informations, reportez-vous à la section Ajout d'une connexion à la base de données dans le chapitre 2 sur p. 18.

Nom de la table. Entrez le nom de la table vers laquelle envoyer les données. Si vous sélectionnez l'option Insérer dans la table, vous pouvez choisir une table existante dans la base de données en cliquant sur le bouton Sélectionner.

Créer une table. Sélectionnez cette option pour créer une nouvelle table de base de données ou écraser une table de base de données existante.

Insérer dans la table. Sélectionnez cette option pour insérer les données dans de nouvelles lignes d'une table de base de données existante.

Fusionner la table. (Le cas échéant) Sélectionnez cette option pour mettre à jour les colonnes de la base de données sélectionnées avec des valeurs de champs de données source correspondants. Sélectionner cette option active le bouton Fusionner, qui affiche une boîte de dialogue dans laquelle vous pouvez mapper les champs de données source sur les colonnes de la base de données.

Supprimer la table existante. Sélectionnez cette option pour supprimer, le cas échéant, une table existante du même nom que la table créée.

Supprimer les lignes existantes. Sélectionnez cette option pour supprimer les lignes existantes de la table avant l'exportation, lors de l'insertion dans une table.

Remarque : si vous sélectionnez l'une des deux options ci-dessus, vous recevez le message Avertissement d'écrasement lors de l'exécution du noeud. Pour que ces avertissements n'apparaissent plus, désélectionnez l'option Avertir lorsqu'un noeud écrase une table de base de données dans l'onglet Notifications de la boîte de dialogue Options utilisateur.

Taille des chaînes par défaut. Les champs marqués comme étant « sans type » dans un noeud Typer en amont sont écrits dans la base de données sous forme de champs de type chaîne. Indiquez la taille des chaînes à utiliser pour les champs sans type.

Cliquez sur Schéma pour ouvrir une boîte de dialogue dans laquelle vous pouvez définir diverses options d'exportation (pour les bases de données prenant en charge cette fonctionnalité), définir des types de données SQL pour vos champs et spécifier la clé primaire en vue de l'indexation de base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de la boîte de dialogue Schéma sur p. 464.

Cliquez sur Index pour définir les options d'indexation de la table exportée, afin d'améliorer les performances de la base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options de l'index sur p. 468.

Cliquez sur Options avancées pour spécifier les options de chargement en masse et de validation de base de données. Pour plus d'informations, reportez-vous à la section Export SGBD - Options avancées sur p. 470.

Entourer de guillemets les noms des tables et colonnes. Sélectionnez les options à utiliser lors de l'envoi d'une instruction CREATE TABLE à la base de données. Les tableaux ou colonnes comportant des espaces ou des caractères spéciaux doivent être mis entre guillemets.

- **Si nécessaire.** Sélectionnez cette option pour que IBM® SPSS® Modeler détermine automatiquement, au cas par cas, la nécessité d'utiliser des guillemets.

- **Toujours.** Sélectionnez cette option pour que les noms de tableau et de colonne soient systématiquement mis entre guillemets.
- **Jamais.** Sélectionnez cette option pour désactiver l'utilisation des guillemets.

Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès que vous cliquez sur Exécuter, ce noeud est ajouté à l'espace de travail de flux.

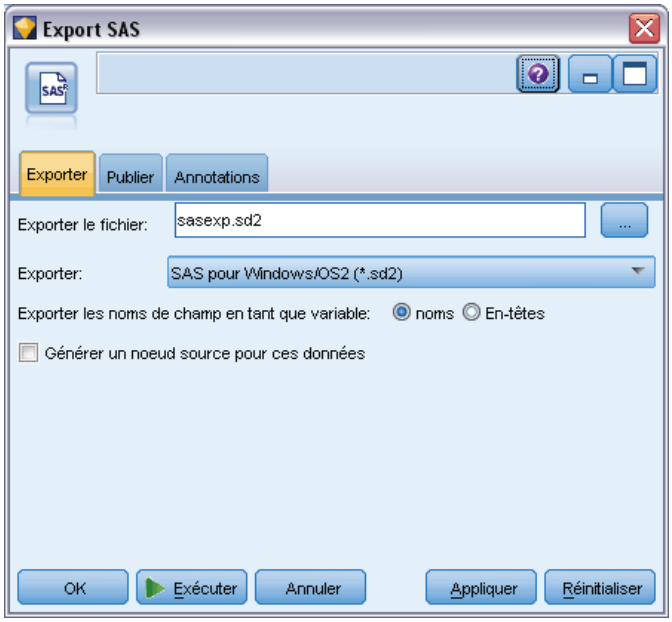
Noeud Export SAS

Remarque : cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud Export SAS permet d'écrire les données au format SAS afin qu'elles puissent être lues par SAS ou par un logiciel compatible. Vous pouvez utiliser trois formats de fichier SAS pour l'exportation : SAS pour Windows/OS2, SAS pour UNIX ou SAS version 7/8.

Noeud Export SAS - Onglet Exporter

Figure 7-11
Noeud Export SAS - Onglet Exporter



Exporter le fichier. Indiquez le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Exporter. Indiquez le format du fichier d'exportation. Les options disponibles sont les suivantes : SAS pour Windows/OS2, SAS pour UNIX ou SAS Version 7/8.

Exporter les noms de champ en tant que variable. Sélectionnez les options d'exportation des noms et des étiquettes de champ depuis IBM® SPSS® Modeler pour leur utilisation avec SAS.

- **Noms et étiquettes de variable.** Sélectionnez cette option pour exporter les noms et les étiquettes de champs SPSS Modeler. Les noms sont exportés en tant que noms de variable SAS, alors que les étiquettes le sont en tant qu'étiquettes de variable SAS.
- **Noms en tant qu'étiquettes de variable** Sélectionnez cette option pour utiliser les noms de champ SPSS Modeler en tant qu'étiquettes de variable dans SAS. Les noms de champ SPSS Modeler prennent en charge des caractères non valides dans les noms de variable SAS. Pour éviter de créer des noms SAS incorrects, sélectionnez noms à la place.

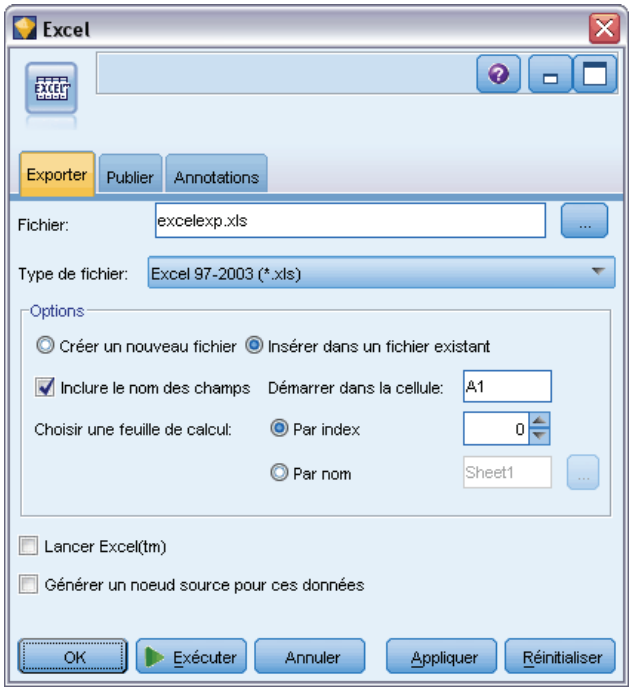
Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source SAS pour lire le fichier de données exporté. Pour plus d'informations, reportez-vous à la section Noeud source SAS dans le chapitre 2 sur p. 51.

Noeud Export Excel

Le noeud Export Excel génère une sortie de données au format Microsoft Excel (*.xls*). Si vous le souhaitez, vous pouvez choisir de lancer Excel automatiquement et d'ouvrir le fichier exporté lors de l'exécution du noeud.

Noeud Excel - Onglet Exporter

Figure 7-12
Noeud Export Excel - Onglet Exporter



Nom du fichier. Entrez directement le nom du fichier ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier. Le nom de fichier par défaut est *excelexp.xls*.

Type de fichier. Sélectionnez le fichier de type Excel que vous souhaitez exporter.

Créer un nouveau fichier. Crée un nouveau fichier Excel.

Insérer dans un fichier existant. Le contenu est remplacé dès le début de la cellule désignée par le champ Démarrer dans la cellule. Les autres cellules de la feuille de calcul sont laissées avec leur contenu d'origine.

Inclure les noms des champs. Indique si les noms de champ doivent être inclus dans la première ligne de la feuille de calcul.

Démarrer dans la cellule. L'emplacement des cellules est utilisé pour le premier enregistrement d'exportation (ou le premier nom de champ si Inclure les noms de champ est sélectionné). Les données sont remplies à droite et en bas de la cellule d'origine.

Choisissez une feuille de calcul. Spécifie la feuille de calcul dans laquelle vous souhaitez exporter les données. Vous pouvez identifier la feuille de calcul, via un index ou un nom.

- **Par index :** Si vous créez un nouveau fichier, spécifiez un nombre de 0 à 9 pour identifier la feuille de calcul dans laquelle vous souhaitez exporter, en commençant par 0 pour la première feuille de calcul, 1 pour la seconde feuille de calcul, etc. Vous pouvez utiliser des valeurs de 10 ou plus uniquement si une feuille de calcul existe déjà à cette position.
- **Par nom.** Si vous créez un nouveau fichier, spécifiez le nom utilisé pour la feuille de calcul. Si vous effectuez une insertion dans un fichier existant, les données sont insérées dans cette feuille de calcul si elle existe, ou une nouvelle feuille de calcul avec ce nom est créée.

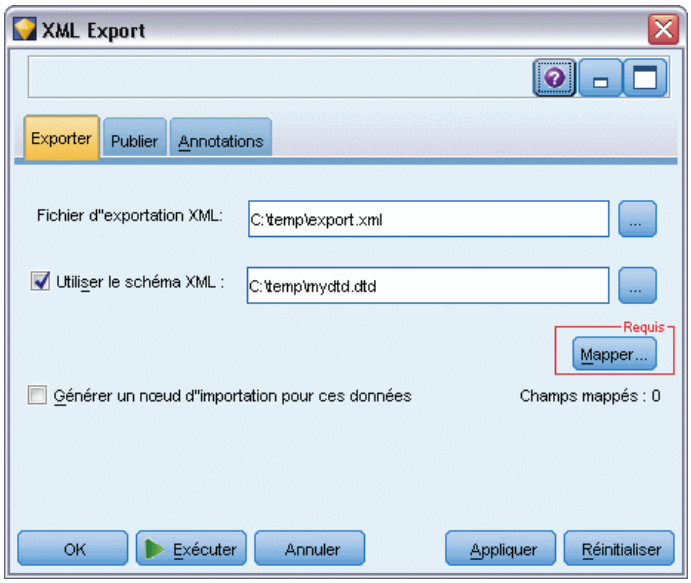
Lancer Excel. Indique si Excel est lancé automatiquement sur le fichier exporté lors de l'exécution du noeud. Dans le cas d'une exécution en mode réparti dans IBM® SPSS® Modeler Server, la sortie est enregistrée dans le système de fichiers du serveur et Excel est lancé sur le client avec une copie du fichier exporté.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un nœud source Excel pour lire le fichier de données exporté. Pour plus d'informations, reportez-vous à la section Noeud source Excel dans le chapitre 2 sur p. 53.

Noeud Export XML

Le noeud Export XML vous permet d'exporter des données au format XML à l'aide de l'encodage UTF-8. Vous pouvez également créer un noeud source XML pour lire de nouveau les données exportées dans le flux.

Figure 7-13
Exportation des données XML



Exporter le fichier XML. Le chemin complet et le nom du fichier XML dans lequel vous souhaitez exporter les données.

Utiliser un schéma XML. Cochez cette case si vous souhaitez utiliser un schéma ou DTD pour contrôler la structure des données exportées. Si vous la cochez, vous activez le bouton Mapper, décrit ci-dessous.

Si vous n'utilisez pas de schéma ou DTD, la structure par défaut suivante est utilisée pour les données exportées :

```
<records>
   <record>
      <fieldname1>value</fieldname1>
      <fieldname2>value</fieldname2>
   :
      <fieldnameN>value</fieldnameN>
   </record>
   <record>
   :
   :
   </record>
   :
   :
</records>
```

Les espaces dans un nom de champ sont remplacés par des caractères de soulignement ; par exemple, "Mon champ" devient <My_Field>.

Mapper. Si vous avez choisi d'utiliser un schéma XML, ce bouton ouvre une boîte de dialogue dans laquelle vous pouvez spécifier la partie de la structure XML à utiliser pour commencer chaque nouvel enregistrement. Pour plus d'informations, reportez-vous à la section Mappage XML - Options Enregistrements sur p. 493.

Champs mappés. Indique le nombre de champs ayant été mappés.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un nœud source XML pour lire le fichier de données exporté. Pour plus d'informations, reportez-vous à la section Noeud source XML dans le chapitre 2 sur p. 54.

Écrire des données XML

Lorsqu'un élément XML est spécifié, la valeur du champ est placée à l'intérieur de la balise de l'élément :

<element>value</element>

Lorsqu'un attribut est mappé, la valeur du champ est placée en tant que valeur pour l'attribut :

<element attribute="value">

Si un champ est mappé sur un élément au-dessus de l'élément <records>, le champ n'est écrit qu'une seule fois et représente une constante pour tous les enregistrements. La valeur de cet élément provient du premier enregistrement.

Si une valeur nulle doit être écrite, ceci est réalisé en spécifiant un contenu vide. Pour les éléments, il s'agit de :

<element></element>

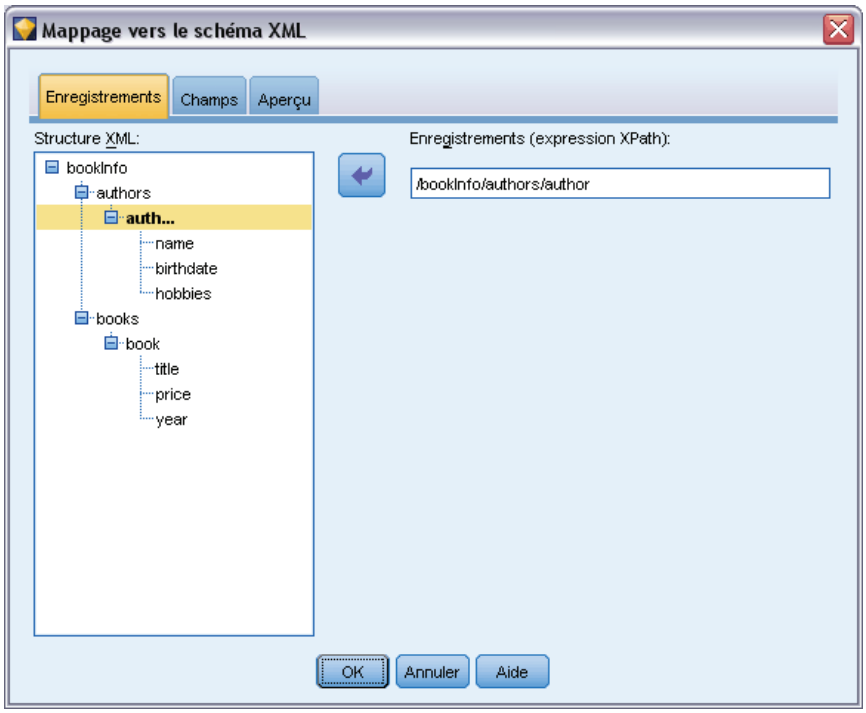
Pour les attributs, il s'agit de :

<element attribute="">

Mappage XML - Options Enregistrements

L'onglet Enregistrements vous permet de spécifier la partie de la structure XML à utiliser pour commencer chaque nouvel enregistrement. Afin de procéder correctement au mappage sur un schéma, vous devez spécifier le séparateur d'enregistrement.

Figure 7-14
Mappage XML - Enregistrements



Structure XML. Un arbre hiérarchique montrant la structure du schéma XML spécifié dans l'écran précédent.

Enregistrements (expression XPath). Pour définir le séparateur d'enregistrement, sélectionnez un élément dans la structure XML et cliquez sur le bouton de la flèche droite. À chaque fois que cet élément est rencontré dans les données source, un nouvel enregistrement est créé dans le fichier de résultat.

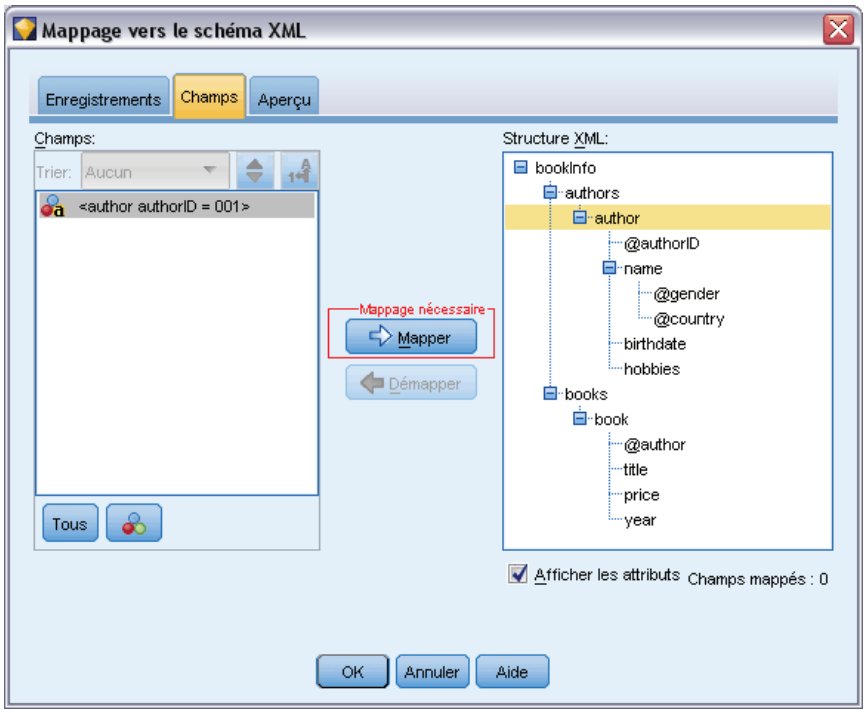
Remarque : Si vous sélectionnez l'élément racine de la structure XML, un seul enregistrement peut être écrit, et tous les autres enregistrements sont ignorés.

Mappage XML - Options Champs

L'onglet Champs permet de mapper des champs de l'ensemble de données sur des éléments ou des attributs dans la structure XML lorsqu'un fichier de schéma est utilisé.

Les noms de fichier qui correspondent à un nom d'élément ou d'attribut sont automatiquement mappés tant que le nom d'élément ou d'attribut est unique. Par conséquent, s'il y a un élément et un attribut portant le nom field1, il n'y a pas de mappage automatique. S'il n'y a qu'un seul élément de la structure nommé field1, un champ portant ce nom dans le flux est automatiquement mappé.

Figure 7-15
Champs de mappage XML



Champs. La liste des champs utilisés dans le modèle. Sélectionnez un ou plusieurs champs comme partie source du mappage. Vous pouvez utiliser les boutons situés en bas de la liste pour sélectionner tous les fichiers ou tous les champs ayant un niveau de mesure particulier.

Structure XML. Sélectionnez un élément de la structure XML en tant que cible de mappage. Pour créer le mappage, cliquez sur Mapper. Le mappage s'affiche alors. Le nombre de champs ayant été mappés de cette manière s'affiche en dessous de cette liste.

Pour supprimer un mappage, sélectionnez l'élément dans la liste de la structure XML et cliquez sur Démapper.

Afficher les attributs. Affiche ou masque les attributs, le cas échéant, des éléments XML dans la structure XML.

Mappage XML - Aperçu

Dans l'onglet Aperçu, cliquez sur Mettre à jour pour voir un aperçu du XML qui sera écrit.

Si le mappage n'est pas correct, revenez à l'onglet Enregistrements ou Champs pour corriger les erreurs et cliquez de nouveau sur Mettre à jour pour voir le résultat.

Chapitre
8

Noeuds IBM SPSS Statistics

Noeuds IBM SPSS Statistics - Présentation

Pour compléter IBM® SPSS® Modeler et ses capacités de Data mining, IBM® SPSS® Statistics vous permet d'effectuer des analyses statistiques et une gestion de données plus avancées.

Lorsque vous avez installé une copie compatible de SPSS Statistics avec sa licence, vous pouvez le connecter à partir de SPSS Modeler et effectuer une manipulation et des analyses de données complexes et en plusieurs étapes que SPSS Modeler ne prend habituellement pas en charge. Pour les utilisateurs avancés, il existe également une option permettant de modifier les analyses en utilisant la syntaxe de commande. Consulter les notes de version pour obtenir des informations concernant la compatibilité de version.

S'ils sont disponibles, les noeuds SPSS Statistics apparaissent dans une partie spécifique de la palette des noeuds.

Remarque : Nous vous recommandons d'instancier vos données dans un noeud Typer avant d'utiliser les noeuds Transformer, Modèle ou Sortie de SPSS Statistics. Ceci est aussi nécessaire lors de l'utilisation de la commande de syntaxe AUTORECODE.

La palette SPSS Statistics comporte les noeuds suivants :

Le noeud Statistics lit les données du format de fichier *.sav* utilisé par SPSS Statistics ainsi que des fichiers cache enregistrés dans SPSS Modeler, qui utilisent le même format. Pour plus d'informations, reportez-vous à la section Noeud Statistics sur p. 497.

Le noeud Transformation exécute une sélection de commandes de syntaxe SPSS Statistics en fonction des sources de données dans SPSS Modeler. Ce noeud requiert une copie avec licence de SPSS Statistics. Pour plus d'informations, reportez-vous à la section Noeud Transformation Statistics sur p. 499.

Le noeud Modèle Statistics vous permet d'analyser et de travailler avec vos données en exécutant des procédures SPSS Statistics qui produisent un PMML. Ce noeud requiert une copie avec licence de SPSS Statistics. Pour plus d'informations, reportez-vous à la section Noeud Modèle Statistics sur p. 503.

Le noeud Sortie Statistics vous permet d'appeler une procédure SPSS Statistics pour analyser les données SPSS Modeler. De nombreuses procédures d'analyses SPSS Statistics sont disponibles. Ce noeud requiert une copie avec licence de SPSS Statistics. Pour plus d'informations, reportez-vous à la section Nœud Sortie Statistics sur p. 507.

© Copyright IBM Corporation 1994, 2012.

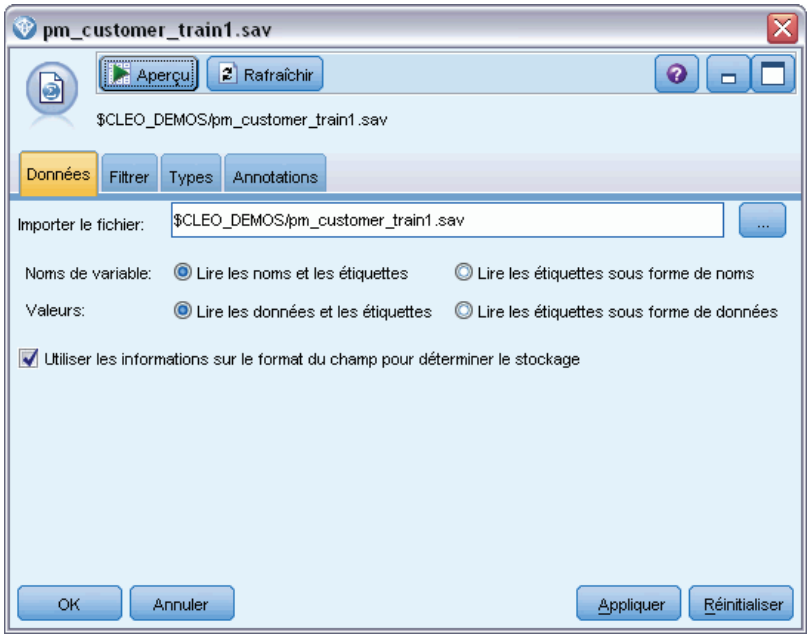
Le noeud Exporter Statistics génère des données au format SPSS Statistics.*sav*. Les fichiers *.sav* peuvent être lus par SPSS Statistics Base et d'autres produits . Ce format est également utilisé pour les fichiers cache SPSS Modeler. Pour plus d'informations, reportez-vous à la section Noeud Exporter Statistics sur p. 511.

Remarque : Si votre copie de SPSS Statistics possède une licence mono-utilisateur et que vous exécutez un flux avec deux branches ou plus, chacune d'elles contenant un noeud SPSS Statistics, vous pourriez recevoir une erreur de licence. Cette erreur survient lorsque la session SPSS Statistics d'une branche n'est pas terminée alors que la session d'une autre branche tente de démarrer. Si possible, revoyez la conception du flux de manière à ce qu'il ne présente pas plusieurs branches contenant des noeuds SPSS Statistics et s'exécutant en parallèle.

Noeud Statistics

Vous pouvez utiliser le noeud Statistics pour lire des données directement à partir d'un fichier IBM® SPSS® Statistics enregistré (*.sav*). Ce format remplace le format de fichier cache qui était utilisé dans les versions précédentes de IBM® SPSS® Modeler. Si vous souhaitez importer un fichier cache enregistré, vous devez utiliser un noeud SPSS Statistics.

Figure 8-1
Importation d'un fichier .sav



Importer le fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (...) pour sélectionner un fichier. Le chemin d'accès du fichier apparaît une fois le fichier sélectionné.

Noms des variables : Sélectionnez une méthode de gestion des noms et étiquettes de variable lors de l'importation d'un fichier SPSS Statistics *.sav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans SPSS Statistics.

- **Lire les noms et les étiquettes.** Sélectionnez cette option afin de lire les noms et les étiquettes de variable dans SPSS Modeler. Par défaut, cette option est sélectionnée et les noms de variable affichés dans le nœud Typer. Les étiquettes peuvent apparaître dans les graphiques, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux. Par défaut, l'affichage des étiquettes de la sortie est désactivé.
- **Lire les étiquettes sous forme de noms.** Sélectionnez cette option pour lire les étiquettes de variable descriptives du fichier SPSS Statistics *.sav* SPSS au lieu des noms de champ abrégés, puis utilisez ces étiquettes en tant que noms de variable dans SPSS Modeler.

Valeurs : Sélectionnez une méthode de gestion des noms et étiquettes lors de l'importation d'un fichier SPSS Statistics *.sav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans SPSS Statistics.

- **Lire les données et les étiquettes.** Choisissez cette option pour les valeurs réelles et les étiquettes de valeur dans SPSS Modeler. Par défaut, cette option est sélectionnée et les valeurs proprement dites apparaissent dans le nœud Typer. Les étiquettes de valeur peuvent être affichées dans le Générateur de formules, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux.
- **Lire les étiquettes sous forme de données.** Choisissez cette option si vous préférez utiliser les étiquettes de valeurs du fichier *.sav* plutôt que les codes numériques ou symboliques utilisés pour représenter les valeurs. Par exemple, si vous sélectionnez cette option pour les données dont le champ indiquant le genre a pour valeur 1 et 2 (représentant respectivement *masculin* et *féminin*), le champ sera converti en chaîne, et importera « *masculin* » et « *féminin* » comme valeurs réelles.

  Il est important de prendre en compte les valeurs manquantes dans vos données SPSS Statistics avant de choisir cette option. Par exemple, si un champ numérique n'utilise des étiquettes que pour les valeurs manquantes (0 = *Aucune réponse*, –99 = *Inconnu*), le fait de choisir l'option ci-dessus importe uniquement les étiquettes de valeur *Aucune réponse* et *Inconnu*, et convertit ce champ en chaîne. Dans ce cas, vous devez importer les valeurs elles-mêmes et définir les valeurs manquantes dans un nœud Typer.

Utilisez les informations de formats de champ pour déterminer le stockage. Si cette case est cochée, les valeurs de champs formatées dans le fichier *.sav* en tant qu'entiers (c'est à dire des champs spécifiés comme F*n*.0 dans l'Affichage des variables de SPSS Statistics) sont importées à l'aide du stockage d'entier. Toutes les autres valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels.

Si cette case n'est pas cochée (par défaut), toutes les valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels, qu'elles soient formatées dans le fichier *.sav* sous la forme d'entiers ou non.

Ensembles de réponses multiples. Tout ensemble de réponses multiples défini dans le fichier SPSS Statistics est automatiquement conservé lors de l'importation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples dans le chapitre 4 sur p. 159.

Noeud Transformation Statistics

Le noeud Transformation Statistics vous permet de procéder à des transformations de données grâce à la syntaxe des commandes IBM® SPSS® Statistics. Il est ainsi possible de procéder à certaines transformations non prises en charge par IBM® SPSS® Modeler et d'automatiser des transformations complexes en plusieurs étapes, y compris la création d'un certain nombre de champs à partir d'un noeud unique. Ce noeud ressemble au noeud Sortie Statistics à ceci près que les données sont renvoyées à SPSS Modeler pour analyse complémentaire, alors que, dans le noeud Sortie SPSS, les données sont renvoyées sous forme d'objets de sortie requis, tels que des graphiques ou des tableaux.

Vous devez disposer d'une version compatible de SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, reportez-vous à la section Programmes externes de IBM SPSS Statistics dans le chapitre 6 sur p. 457. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Si nécessaire, vous pouvez utiliser l'onglet Filtrer pour filtrer ou renommer des champs afin qu'ils soient conformes aux conventions de dénomination SPSS Statistics. Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics sur p. 513.

Référence de syntaxe. Pour plus d'informations sur les procédures SPSS Statistics spécifiques, consultez le guide de *référence de la syntaxe des commandes SPSS Statistics*, fourni avec votre copie du logiciel SPSS Statistics. Pour consulter le guide dans l'onglet Syntaxe, choisissez l'option Editeur de syntaxe et cliquez sur le bouton Lancer l'aide syntaxe SPSS Statistics.

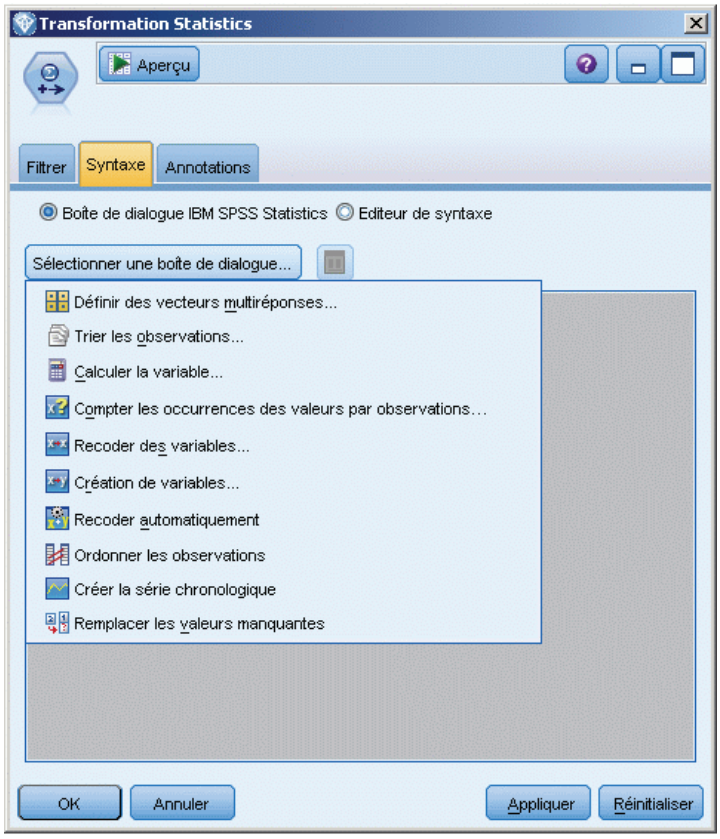
Remarque : ce noeud ne prend pas en charge la totalité de la syntaxe SPSS Statistics. Pour plus d'informations, reportez-vous à la section Syntaxe autorisée sur p. 501.

Noeud Transformation Statistics - Onglet Syntaxe

Option de la boîte de dialogue IBM SPSS Statistics

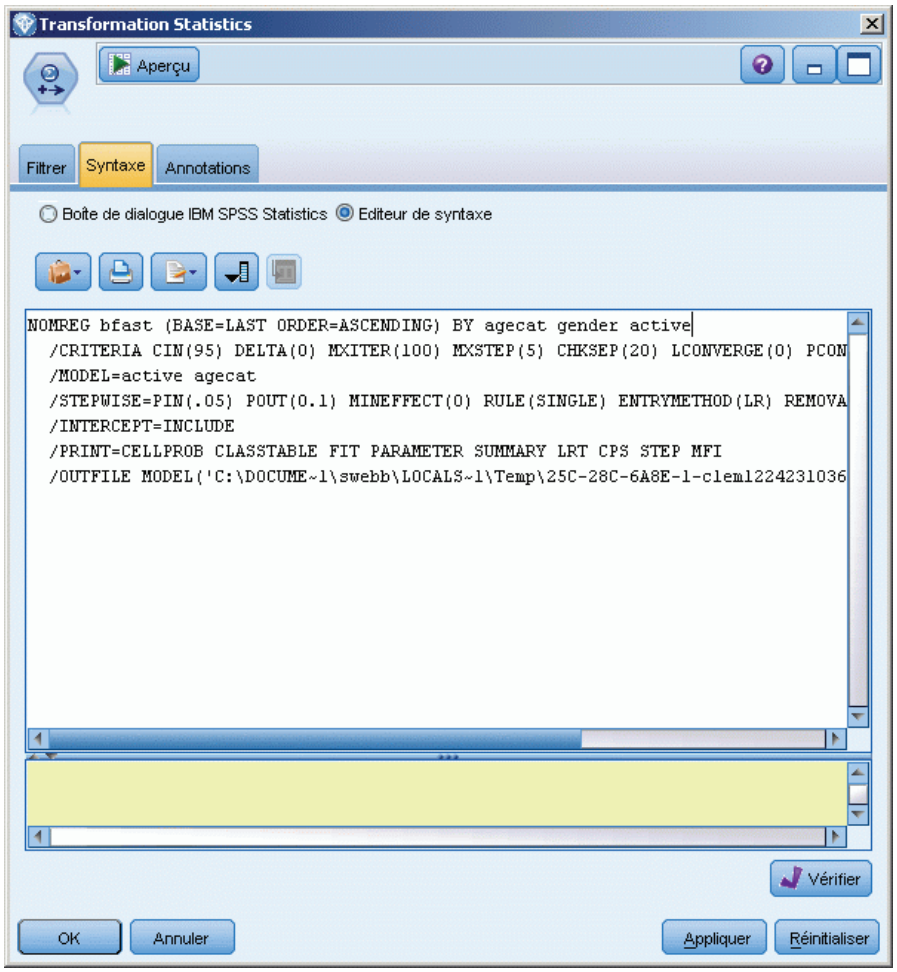
Si vous n'êtes pas habitué à la syntaxe IBM® SPSS® Statistics d'une procédure, la façon la plus simple de créer une syntaxe dans IBM® SPSS® Modeler est de choisir l'option Boîte de dialogue IBM SPSS Statistics, de sélectionner la boîte de dialogue de la procédure, suivre ses instructions et cliquer sur OK. Cela vous permet de placer la syntaxe dans l'onglet Syntaxe du noeud SPSS Statistics utilisé dans SPSS Modeler. Vous pouvez ensuite exécuter le flux afin d'obtenir les résultats de la procédure.

Figure 8-2
Noeud Transformation Statistics - Sélection de boîte de dialogue



Option de l'éditeur de syntaxe IBM SPSS Statistics

Figure 8-3
Noeud Transformation Statistics - Éditeur de syntaxe



Vérifier. Une fois que vous avez saisi vos commandes de syntaxe dans la partie supérieure de la boîte de dialogue, utilisez ce bouton pour valider vos entrées. Toute syntaxe incorrecte est mise en évidence dans la partie inférieure de la boîte de dialogue.

Pour garantir que le processus de vérification n'est pas trop long, lorsque vous validez la syntaxe, une comparaison est effectuée avec un échantillon représentatif de vos données, plutôt qu'avec la totalité de l'ensemble de données, afin d'assurer la validité de vos entrées.

Syntaxe autorisée

Si votre syntaxe est en grande partie héritée de IBM® SPSS® Statistics ou si vous connaissez les fonctions de préparation des données de SPSS Statistics, vous pouvez utiliser le noeud Transformation Statistics pour exécuter un grand nombre des transformations existantes. En tant qu'instruction, le noeud vous permet de transformer les données de façon prévisible, par

exemple en exécutant des commandes en boucle ou en modifiant, ajoutant, triant, filtrant ou sélectionnant des données.

Vous trouverez ci-dessous des exemples de commandes pouvant être exécutées :

- Calculer des nombres aléatoires d'après une loi binomiale :

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- Recoder une variable en une nouvelle variable :

```
RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
```

- Remplacer des valeurs manquantes :

```
RMV Age_1=SMEAN(Age)
```

La syntaxe SPSS Statistics prise en charge par le noeud Transformation Statistics est répertoriée dans le tableau suivant :

Nom de la commande

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

Nom de la commande

```
INSERT
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
SORT CASES
STRING
SUBTITLE
TEMPORARY
TITLE
UPDATE
V2C
VALIDATEDATA
VALUE LABELS
VARIABLE ATTRIBUTE
VARSTOCASES
VECTOR
```

Noeud Modèle Statistics

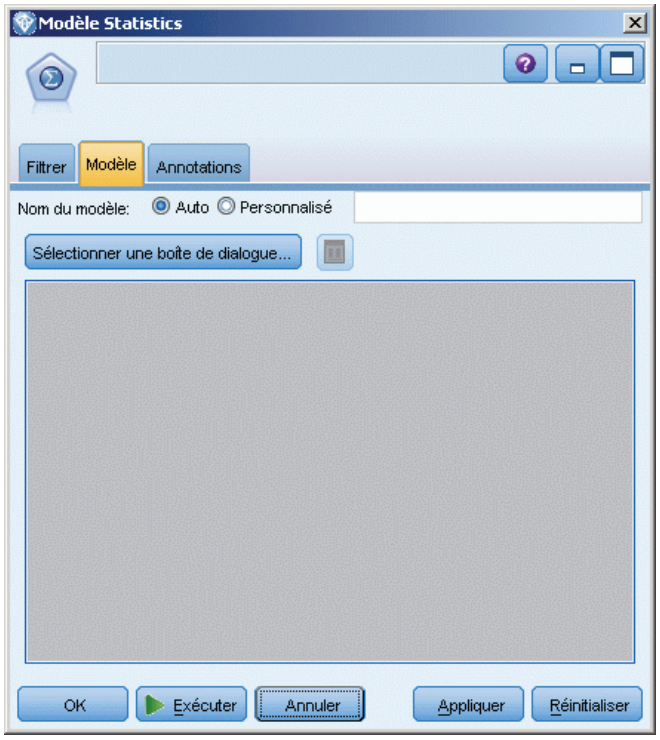
Le noeud Modèle Statistics vous permet d'analyser et de travailler avec vos données en exécutant des procédures IBM® SPSS® Statistics qui produisent PMML. Les nuggets de modèle que vous créez peuvent ensuite être utilisés de la façon habituelle dans les flux IBM® SPSS® Modeler pour le scoring, etc.

Vous devez disposer d'une version compatible de SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, reportez-vous à la section Programmes externes de IBM SPSS Statistics dans le chapitre 6 sur p. 457. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Les procédures d'analyse SPSS Statistics disponibles dépendent du type de licence que vous possédez.

Noeud Modèle Statistics - Onglet Modèle

Figure 8-4
Noeud Modèle Statistics, Onglet Modèle

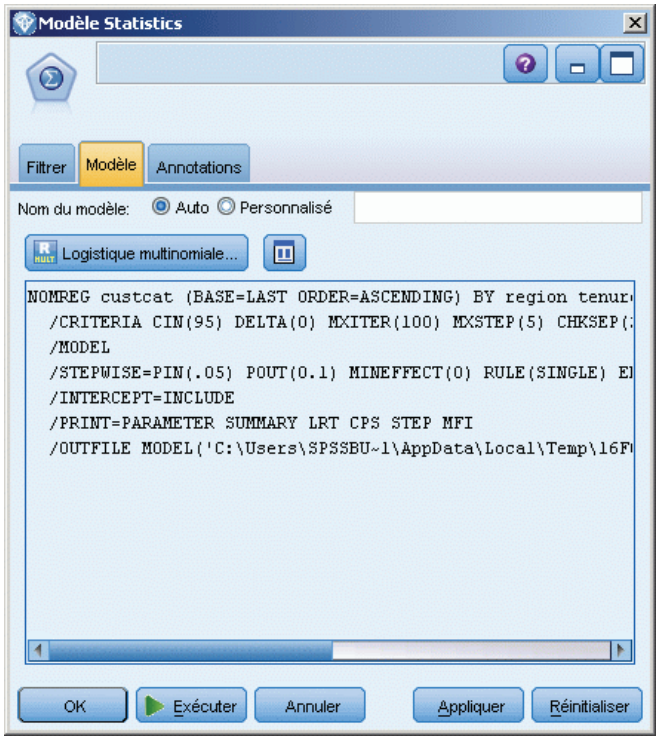


Nom du modèle. Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

Sélectionnez une boîte de dialogue. Cliquez pour afficher une liste des procédures IBM® SPSS® Statistics disponibles que vous pouvez sélectionner et exécuter. Cette liste ne contient que les procédures qui produisent PMML et pour laquelle vous disposez d'une licence, et ne contient pas de procédures écrites par l'utilisateur.

- Cliquez sur la procédure requise ; la boîte de dialogue SPSS Statistics correspondante s'affiche.
- Dans la boîte de dialogue SPSS Statistics , saisissez les détails de la procédure.
- Cliquez sur OK pour revenir au noeud Modèle Statistics; la syntaxe SPSS Statistics apparaît dans l'onglet Modèle.

Figure 8-5
Syntaxe affichée dans l'onglet Modèle

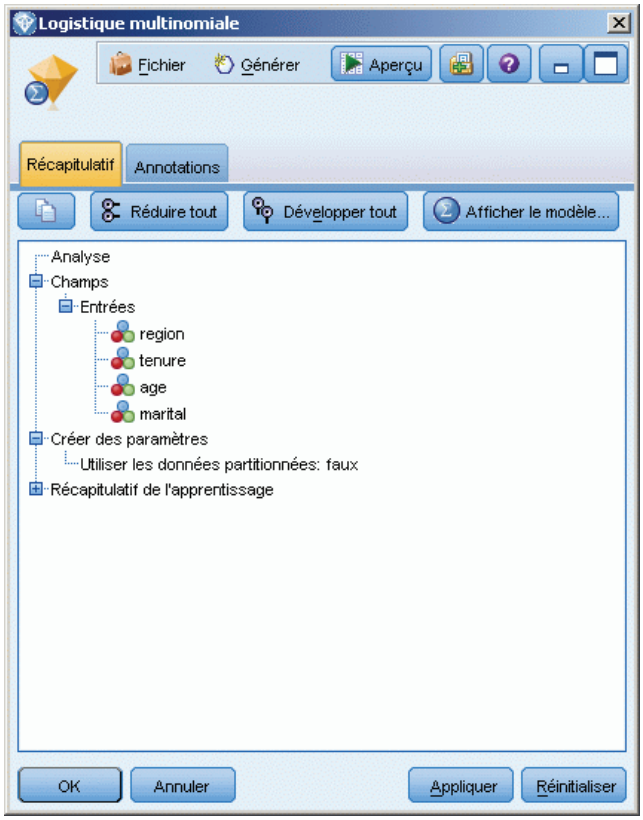


- Pour revenir à la boîte de dialogue SPSS Statistics à tout moment, par exemple pour modifier votre demande, cliquez sur le bouton d'affichage de la boîte de dialogue SPSS Statistics à droite du bouton de sélection des procédures.

Noeud de modèle Statistics - Récapitulatif du nugget de modèle

Lorsque vous exécutez le noeud Modèle Statistics, il exécute la procédure IBM® SPSS® Statistics associée et crée un nugget de modèle que vous pouvez utiliser dans les flux IBM® SPSS® Modeler pour le scoring.

Figure 8-6
Nugget du modèle Statistics, onglet Récapitulatif

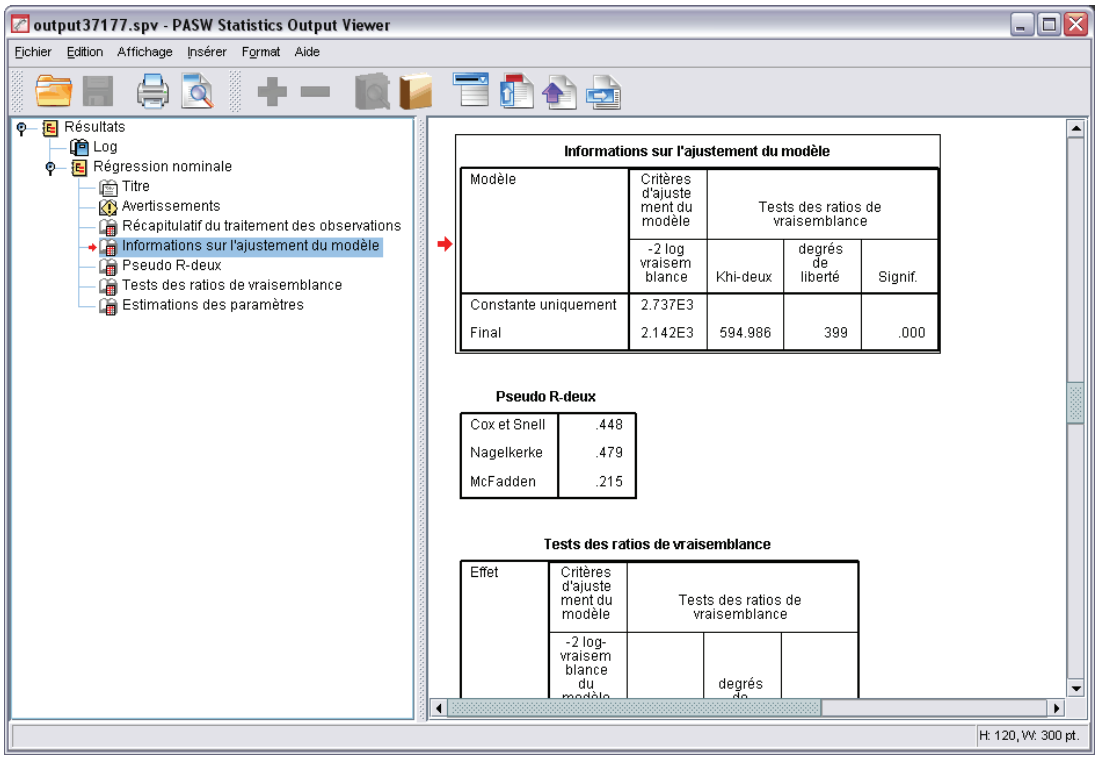


L'onglet Récapitulatif d'un nugget de modèle affiche des informations sur les champs, les paramètres de création et le processus d'estimation du modèle. Les résultats sont présentés dans un arbre que vous pouvez développer ou réduire en cliquant sur des éléments précis.

Le bouton Afficher le modèle affiche les résultats sous une forme modifiée de l'afficheur de sortie Output Viewer de SPSS Statistics. Pour des informations supplémentaires sur cet afficheur, consultez la documentation de SPSS Statistics.

Les options standard d'exportation et d'impression sont disponibles dans le menu Fichier. Pour plus d'informations, reportez-vous à la section Affichage des sorties dans le chapitre 6 sur p. 399.

Figure 8-7
Nugget du modèle Statistics, onglet Options avancées



Nœud Sortie Statistics

Le noeud Sortie Statistics vous permet d'appeler une procédure IBM® SPSS® Statistics pour analyser les données IBM® SPSS® Modeler. Vous pouvez visualiser les résultats dans une fenêtre de navigateur ou les enregistrer au format de fichier de sortie SPSS Statistics. SPSS Modeler permet d'accéder à de nombreuses procédures d'analyses SPSS Statistics.

Vous devez disposer d'une version compatible de SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, reportez-vous à la section Programmes externes de IBM SPSS Statistics dans le chapitre 6 sur p. 457. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Si nécessaire, vous pouvez utiliser l'onglet Filtrer pour filtrer ou renommer des champs afin qu'ils soient conformes aux conventions de dénomination SPSS Statistics. Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics sur p. 513.

Référence de syntaxe. Pour plus d'informations sur les procédures SPSS Statistics spécifiques, consultez le guide de *référence de la syntaxe des commandes SPSS Statistics*, fourni avec votre copie du logiciel SPSS Statistics. Pour consulter le guide dans l'onglet Syntaxe, choisissez l'option Editeur de syntaxe et cliquez sur le bouton Lancer l'aide syntaxe SPSS Statistics.

Noeud Sortie Statistics - Onglet Syntaxe

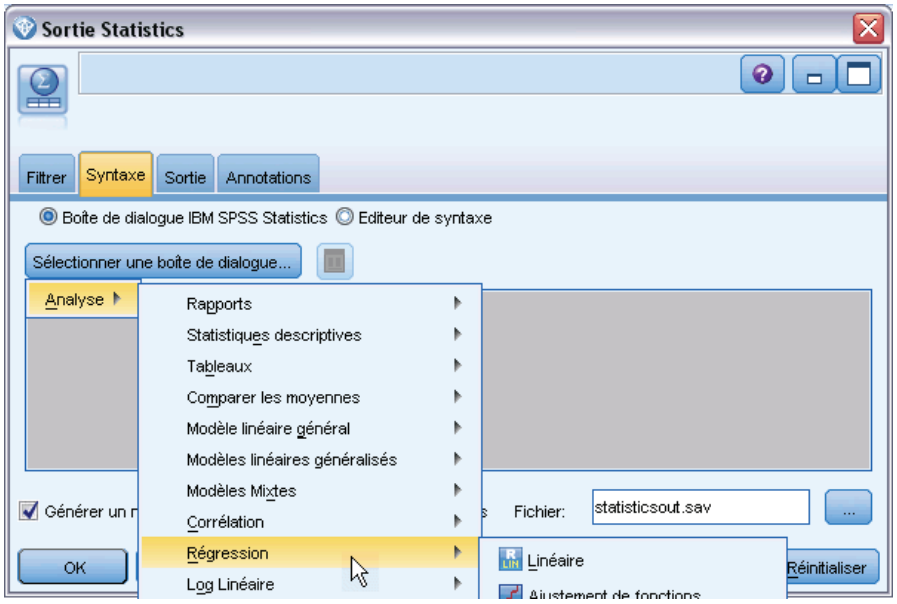
Utilisez cet onglet pour créer une syntaxe pour la procédure IBM® SPSS® Statistics que vous souhaitez utiliser pour analyser vos données. La syntaxe se décompose en deux parties : une **instruction** et des **options** associées. L'instruction indique l'analyse ou l'opération à effectuer, et les champs à utiliser. Les options décrivent les autres aspects de l'analyse, tels que les statistiques à afficher, les champs calculés à enregistrer, etc.

Option de la boîte de dialogue IBM SPSS Statistics

Si vous n'êtes pas habitué à la syntaxe SPSS Statistics d'une procédure, la façon la plus simple de créer une syntaxe dans IBM® SPSS® Modeler est de choisir l'option Boîte de dialogue IBM SPSS Statistics, de sélectionner la boîte de dialogue de la procédure, suivre ses instructions et cliquer sur OK. Cela vous permet de placer la syntaxe dans l'onglet Syntaxe du noeud SPSS Statistics utilisé dans SPSS Modeler. Vous pouvez ensuite exécuter le flux afin d'obtenir les résultats de la procédure.

Vous avez la possibilité de générer un noeud source Statistics pour importer les données obtenues. Cela peut être utile, par exemple, si une procédure écrit des champs tels que des scores dans l'ensemble de données actif en plus d'afficher les résultats.

Figure 8-8
Noeud Sortie Statistics, sélection de boîte de dialogue



Pour créer la syntaxe :

- Cliquez sur le bouton Sélectionner une boîte de dialogue.

- Choisissez une de ces options :
  - **Analyse.** Répertorie le contenu du menu Analyse de SPSS Statistics ; sélectionnez la procédure que vous souhaitez utiliser.
  - **Autre.** Si elle apparaît, répertorie les boîtes de dialogue créées par Custom Dialog Builder dans SPSS Statistics, ainsi que toutes les autres boîtes de dialogue de SPSS Statistics qui n'apparaissent pas dans le menue Analyse et pour lesquelles vous disposez d'une licence. Si aucune boîte de dialogue n'est concernée, cette option n'apparaît pas.

  Remarque : Les boîtes de dialogue Préparation automatique des données n'apparaissent pas.

  Si vous disposez d'une boîte de dialogue personnalisée SPSS Statistics qui crée de nouveaux champs, ces champs ne peuvent pas être utilisés dans SPSS Modeler parce que le noeud Sortie Statistics est un noeud terminal.

- Vous pouvez aussi cocher la case Générer un noeud d'importation pour les données obtenues pour créer un noeud source Statistics à utiliser pour importer les données obtenues vers un autre flux. Le noeud est placé sur l'espace de travail, avec les données contenues dans le fichier *.sav* spécifié dans le champ Fichier (l'emplacement par défaut est le répertoire d'installation SPSS Modeler).

Option de l'éditeur de syntaxe

Pour enregistrer la syntaxe créée pour une procédure fréquemment utilisée :

- Cliquez sur le bouton Options du fichier (le premier de la barre d'outils).
- Sélectionnez Enregistrer ou Enregistrer sous dans le menu.
- Enregistrez le fichier en tant que fichier *.sps*.

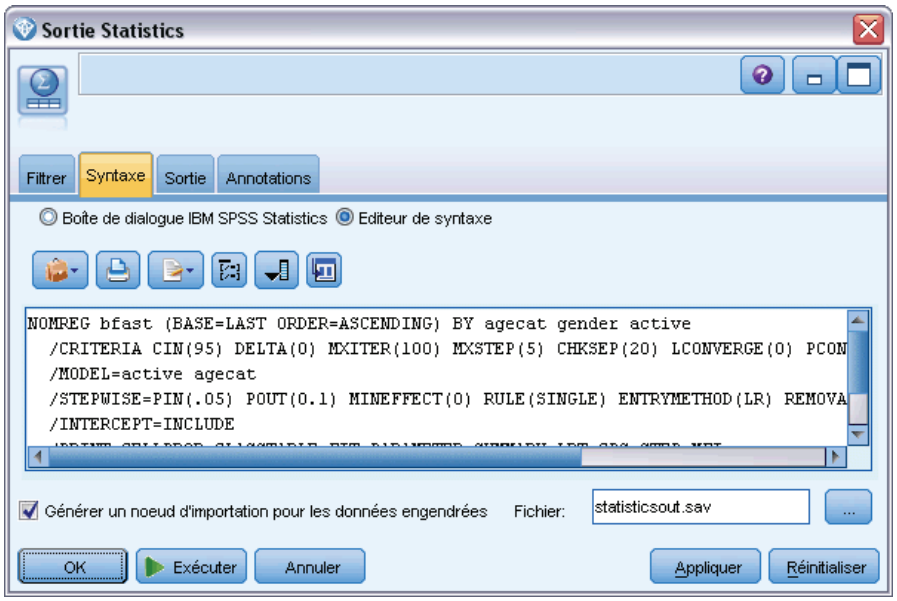
Pour utiliser des fichiers de syntaxe créés préalablement, en remplaçant le contenu actuel, le cas échéant, de l'éditeur de syntaxe :

- Cliquez sur le bouton Options du fichier (le premier de la barre d'outils).
- Dans le menu, sélectionnez Ouvrir.
- Sélectionnez un fichier *.sps* afin de coller son contenu dans l'onglet Syntaxe du noeud Sortie.

Pour insérer une syntaxe préalablement enregistrée sans remplacer le contenu actuel :

- Cliquez sur le bouton Options du fichier (le premier de la barre d'outils).
- Dans le menu, sélectionnez Insérer.
- Sélectionnez un fichier *.sps* afin de coller son contenu dans le noeud Sortie au point spécifié par le curseur.

Figure 8-9
Noeud Sortie Statistics, éditeur de syntaxe



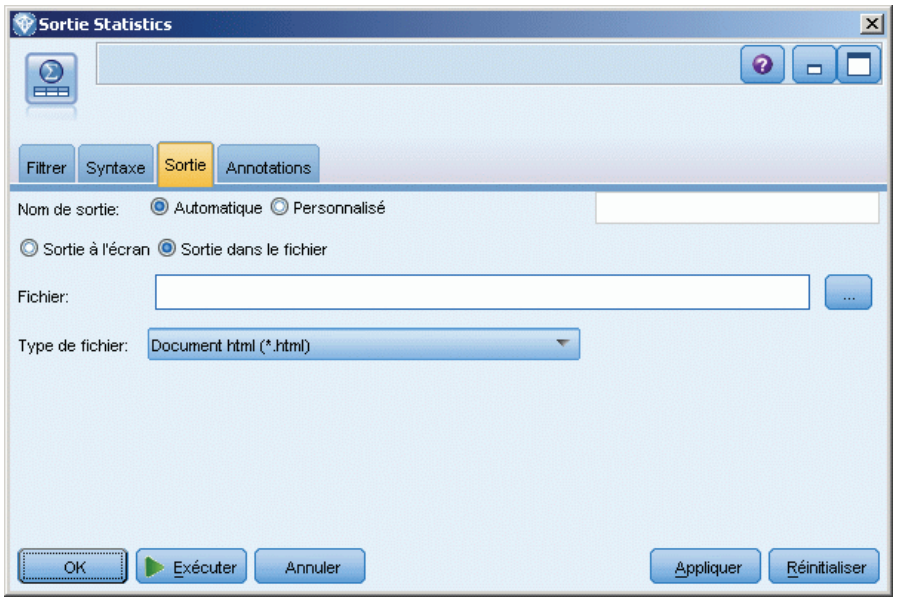
- Vous pouvez aussi cocher la case Générer un noeud d'importation pour les données obtenues pour créer un noeud source Statistics à utiliser pour importer les données obtenues vers un autre flux. Le noeud est placé sur l'espace de travail, avec les données contenues dans le fichier *.sav* spécifié dans le champ Fichier (l'emplacement par défaut est le répertoire d'installation SPSS Modeler).

  Lorsque vous cliquez sur Exécuter, les résultats sont affichés dans l'Afficheur de résultats SPSS Statistics. Pour plus d'informations sur l'afficheur, consultez la documentation de SPSS Statistics.

Noeud Sortie Statistics - Onglet Sortie

L'onglet Sortie vous permet de préciser le format et l'emplacement de la sortie. Vous pouvez choisir d'afficher les résultats à l'écran ou de les envoyer vers un des types de fichier disponibles.

Figure 8-10
Noeud Sortie Statistics, onglet Sortie



Nom de sortie. Spécifie le nom de la sortie générée lorsque le noeud est exécuté. L'option Automatique sélectionne un nom en fonction du nœud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir Personnalisé pour indiquer un autre nom.

Sortie à l'écran (option par défaut)**.** Crée un objet de sortie à afficher en ligne. L'objet de sortie apparaît dans l'onglet Sorties de la fenêtre du gestionnaire lors de l'exécution du noeud de sortie.

Sortie dans le fichier. Enregistre la sortie dans un fichier lorsque vous exécutez le noeud. Si vous choisissez cette option, entrez un nom de fichier dans le champ Nom de fichier (ou parcourez l'arborescence et indiquez un nom de fichier à l'aide du sélecteur de fichiers), puis sélectionnez un type de fichier.

Type de fichier. Sélectionnez le type de fichier auquel vous souhaitez envoyer le résultat.

- **Document HTML (*.html).** Écrit le résultat au format HTML.
- **SPSS Statistics Afficheur de fichiers (*.spv).** Écrit le résultat dans un format qui peut être lu par l'afficheur de résultat de IBM® SPSS® Statistics.
- **SPSS Statistics Fichier Web Reports (*.spw).** Écrit le résultat au format Web Reports de SPSS Statistics qui peut être publié sur un référentiel IBM SPSS Collaboration and Deployment Services et consulté ultérieurement dans un navigateur Web. Pour plus d'informations, reportez-vous à la section Publication sur le Web dans le chapitre 6 sur p. 399.

Noeud Exporter Statistics

Le noeud Exporter Statistics vous permet d'exporter les données au format IBM® SPSS® Statistics*.sav*. SPSS Statistics Les fichiers *.sav* peuvent être lus par SPSS Statistics Base et d'autres modules. Ce format est également utilisé pour les fichiers cache IBM® SPSS® Modeler.

Il est possible que le mappage des noms de champ SPSS Modeler à des noms de variable SPSS Statistics génère des erreurs, car ces derniers sont limités à 64 caractères et ne peuvent pas inclure certains caractères, comme l'espace, le signe dollar ($), et le tiret (SPSS Statistics). Vous pouvez ajuster ces restrictions de deux façons :

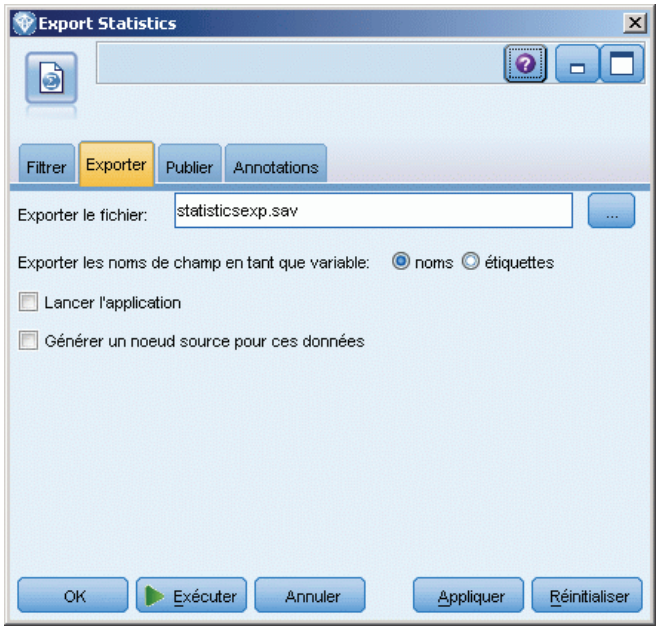
- Vous pouvez renommer les champs en respectant les conventions de dénomination des variables SPSS Statistics en cliquant sur l'onglet Filtrer. Pour plus d'informations, reportez-vous à la section Changement du nom ou filtrage des champs pour IBM SPSS Statistics sur p. 513.
- Exportez les noms et étiquettes de champ à partir de SPSS Modeler.

Remarque : SPSS Modeler écrit les fichiers *.sav* au format Unicode UTF-8. SPSS Statistics prend en charge uniquement les fichiers au format Unicode UTF-8 de la version 16.0 et versions supérieures. Pour limiter les risques de corruption des données, les fichiers *.sav* enregistrés avec le codage Unicode ne doivent pas être utilisés avec les versions de SPSS Statistics antérieures à 16.0. Pour plus d'informations, reportez-vous à l'aide de SPSS Statistics.

Ensembles de réponses multiples. Tous les ensembles de réponses multiples définis dans le flux seront automatiquement préservés lors de l'exportation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, reportez-vous à la section Modification des ensembles de réponses multiples dans le chapitre 4 sur p. 159.

Noeud Exporter Statistics - Onglet Exporter

Figure 8-11
Noeud Exporter Statistics - Onglet Exporter



Exporter le fichier. Indique le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Exporter les noms de champ en tant que variable. Indique une méthode de gestion des noms et étiquettes de variable lors de l'exportation de données de IBM® SPSS® Modeler vers un fichier *.sav*IBM® SPSS® Statistics.

- **Noms et étiquettes de variable.** Sélectionnez cette option pour exporter les noms et les étiquettes de champs SPSS Modeler. Les noms sont exportés en tant que noms de variable SPSS Statistics, alors que les étiquettes le sont en tant qu'étiquettes de variable SPSS Statistics.
- **Noms en tant qu'étiquettes de variable** Sélectionnez cette option pour utiliser les noms de champ SPSS Modeler en tant qu'étiquettes de variable dans SPSS Statistics. Les noms de champ SPSS Modeler prennent en charge des caractères non valides dans les noms de variable SPSS Statistics. Pour éviter de créer des noms SPSS Statistics incorrects, sélectionnez Etiquettes ou utilisez l'onglet Filtre pour ajuster les noms de champ.

Lancer l'application. Si SPSS Statistics est installé sur votre ordinateur, vous pouvez sélectionner cette option pour exécuter l'application directement sur le fichier de données enregistré. Les options de lancement de cette application doivent être indiquées dans la boîte de dialogue Programmes externes. Pour plus d'informations, reportez-vous à la section Programmes externes de IBM SPSS Statistics dans le chapitre 6 sur p. 457. Pour créer simplement un fichier SPSS Statistics*.sav* sans ouvrir de programme externe, désélectionnez cette option.

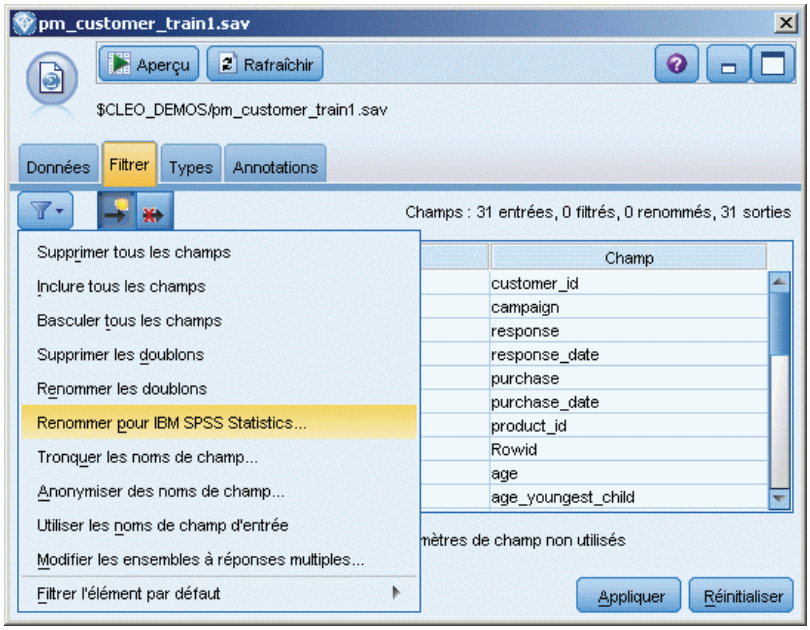
Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source Statistics pour lire le fichier de données exporté. Pour plus d'informations, reportez-vous à la section Noeud Statistics sur p. 497.

Changement du nom ou filtrage des champs pour IBM SPSS Statistics

Avant d'exporter ou de déployer des données de IBM® SPSS® Modeler vers des applications externes telles que IBM® SPSS® Statistics, vous pouvez être amené à renommer ou à ajuster des noms de champ. Les boîtes de dialogue Transformation Statistics, Sortie Statistics et Exporter Statistics contiennent un onglet Filtrer pour faciliter ce processus.

Vous trouverez dans une autre rubrique une brève description de l'onglet Filtrer. Pour plus d'informations, reportez-vous à la section Paramétrage des options de filtrage dans le chapitre 4 sur p. 156. Cette rubrique fournit des astuces concernant la lecture des données dans SPSS Statistics.

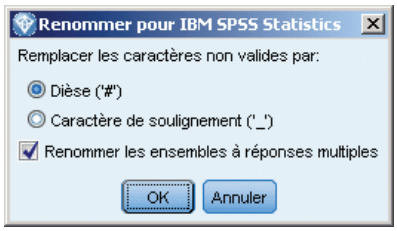
Figure 8-12
Utilisation de l'onglet Filtrer du noeud Statistics afin de renommer les champs pour IBM SPSS Statistics



Pour ajuster les noms de fichiers afin de se conformer aux conventions de dénomination de SPSS Statistics :

- Dans l'onglet Filtrer, cliquez sur le bouton de la barre d'outils du menu des options de filtrage (le premier de la barre d'outils).
- Sélectionnez Renommer pour SPSS Statistics.

Figure 8-13
Changement de nom des champs



- Dans la boîte de dialogue Renommer pour SPSS Statistics, vous pouvez choisir de remplacer les caractères non valides des noms de fichier soit par un caractère dièse (#), soit parc un caractère de soulignement (_).

Renommer des ensembles à réponses multiples. Sélectionnez cette option si vous souhaitez ajuster le nom de plusieurs ensembles a réponses multiples, lesquels peuvent être importés dans SPSS Modeler à l'aide d'un noeud source Statistics. Ils sont utilisés pour enregistrer des données qui peuvent comporter plus d'une valeur pour chaque cas, telles que les réponses à une enquête.

Chapitre
9

Super noeuds

Présentation des super noeuds

L'une des raisons pour laquelle l'interface de programmation visuelle de IBM® SPSS® Modeler est si facile à utiliser est que chaque noeud a une fonction clairement définie. Toutefois, un traitement complexe peut nécessiter une longue séquence de noeuds. Cela risque d'encombrer l'espace de travail de flux et de rendre difficile le suivi des diagrammes de flux. Vous pouvez éviter l'encombrement d'un flux long et complexe de deux manières :

- Vous pouvez partager une séquence de traitement en plusieurs flux qui s'auto-alimentent. Le premier flux, par exemple, crée un fichier de données que le deuxième utilise comme données d'entrée. Le deuxième crée un fichier que le troisième utilise également comme données d'entrée, et ainsi de suite. Vous pouvez gérer ces flux en les enregistrant dans un **projet**. Un projet permet d'organiser plusieurs flux, ainsi que leurs sorties. Cependant, un fichier de projet contient seulement une référence aux objets qu'il contient, et vous avez plusieurs fichiers de flux à gérer.
- Lorsque vous utilisez des processus de flux complexes, une alternative simple consiste à créer un **super noeud**.

Les super noeuds regroupent en un noeud unique plusieurs noeuds, en encapsulant les sections d'un flux de données. Le travail de Data mining est facilité grâce aux avantages suivants :

- Les flux sont plus nets et plus faciles à gérer.
- Les noeuds peuvent être regroupés en un super noeud propre à votre problème.
- Les super noeuds peuvent être exportés vers des bibliothèques pour être réutilisés dans plusieurs projets de Data mining.

Types de super noeuds

Les super noeuds sont représentés dans le flux de données par une icône en forme d'étoile. L'icône est partiellement hachurée pour indiquer le type de super noeud et le sens dans lequel le flux se déplace.

Il existe trois types de super noeuds :

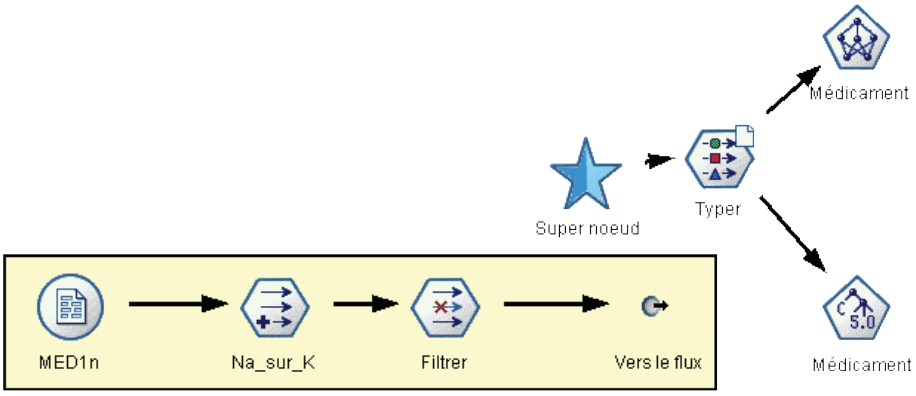
- Super noeuds source
- Super noeuds d'exécution
- Super noeuds terminaux

© Copyright IBM Corporation 1994, 2012.

Super noeuds source

A l'instar des noeuds source standard, les super noeuds source contiennent une source de données et peuvent être utilisés partout où le noeud source est utilisé. La partie gauche d'un super noeud source est hachurée pour indiquer qu'il n'est pas accessible à partir de la gauche et que les données doivent se déplacer en aval *à partir* d'un super noeud.

Figure 9-1
Super noeud source avec version zoom avant appliquée au flux

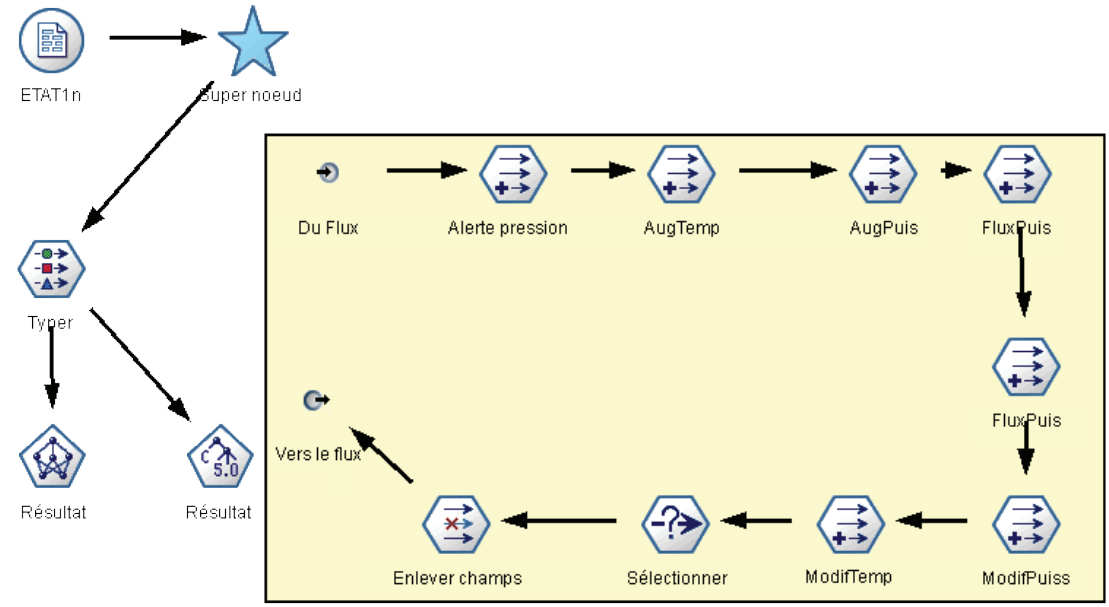


Les super noeuds source ne disposent que d'un seul point de connexion, à droite, indiquant que les données quittent le super noeud en direction du flux.

Super noeuds d'exécution

Les super noeuds d'exécution contiennent uniquement des noeuds d'exécution non hachurés pour indiquer que les données peuvent à la fois *entrer* et *sortir* de ce type de super noeud.

Figure 9-2
Super noeud d'exécution avec version zoom avant appliquée au flux



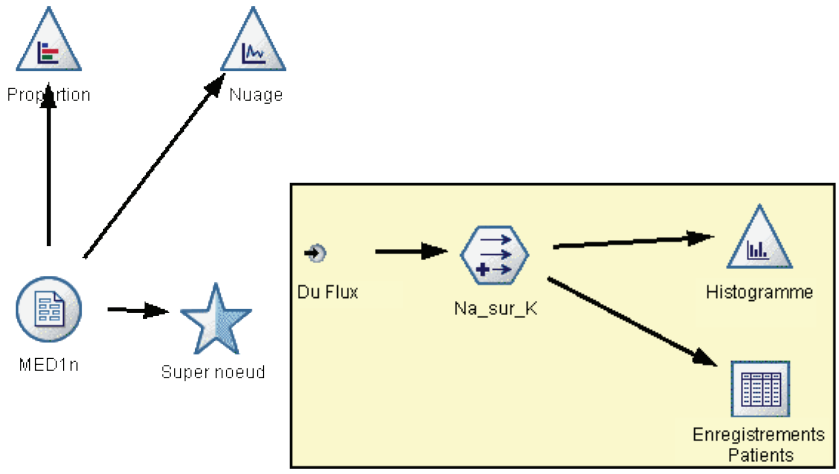
Les super noeuds d'exécution disposent de points de connexion à gauche et à droite, indiquant que les données pénètrent dans le super noeud et repartent dans le flux. Bien que les super noeuds puissent contenir des fragments de flux supplémentaires, et même des flux supplémentaires, les deux points de connexion doivent circuler via un chemin d'accès unique reliant les points *A partir du flux* et *Vers le flux* entre eux.

Remarque : Les super noeuds d'exécution sont parfois appelés « *super noeuds de manipulation* ».

Super noeuds terminaux

Les super noeuds terminaux contiennent un ou plusieurs noeuds terminaux (Nuage, Table, etc.) et peuvent être utilisés de la même façon que des noeuds terminaux. La partie droite d'un super noeud terminal est hachurée pour indiquer qu'il n'est pas accessible à droite et que les données ne peuvent se déplacer que *vers* un super noeud terminal.

Figure 9-3
Super noeud terminal avec version zoom avant appliquée au flux



Les super noeuds terminaux ne disposent que d'un seul point de connexion, à gauche, indiquant que les données pénètrent dans le super noeud depuis le flux et finissent à l'intérieur du super noeud.

Les super noeuds terminaux peuvent également contenir des scripts utilisés pour indiquer l'ordre d'exécution de tous les noeuds terminaux au sein du super noeud. Pour plus d'informations, reportez-vous à la section Super noeuds et génération de scripts sur p. 534.

Création de super noeuds

La création d'un super noeud entraîne le « rétrécissement » du flux de données puisque plusieurs noeuds sont encapsulés dans un seul. Une fois que vous avez créé ou chargé un flux dans l'espace de travail, vous pouvez créer un super noeud de différentes manières.

Sélection multiple

La méthode la plus simple pour créer un super noeud consiste à sélectionner tous les noeuds que vous souhaitez encapsuler :

- Utilisez la souris pour sélectionner plusieurs noeuds dans l'espace de travail du flux. Vous pouvez également utiliser la méthode Maj+clic pour sélectionner un flux ou la section d'un flux. *Remarque* : Vous devez sélectionner des noeuds provenant d'un flux continu ou bifurqué. Vous ne pouvez pas sélectionner des noeuds qui ne sont pas adjacents ou connectés.
- Ensuite, encapsulez les noeuds sélectionnés en exécutant l'une des trois méthodes suivantes :
  - Cliquez sur l'icône du super noeud (en forme d'étoile) dans la barre d'outils.

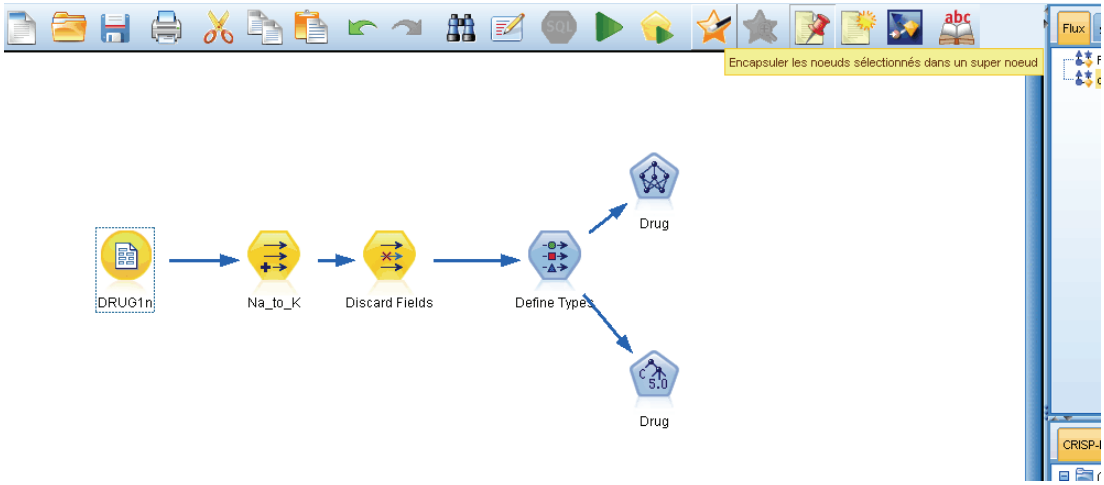
- Cliquez avec le bouton droit de la souris sur le super noeud et, dans le menu contextuel, sélectionnez :

Créer un super noeud > A partir de la sélection

- Dans le menu Super noeud, sélectionnez :

Créer un super noeud > A partir de la sélection

Figure 9-4
Création d'un super noeud à l'aide de la sélection multiple



Ces trois options encapsulent les noeuds dans un super noeud hachuré afin de refléter son type (source, exécution ou terminal) en fonction de son contenu.

Sélection unique

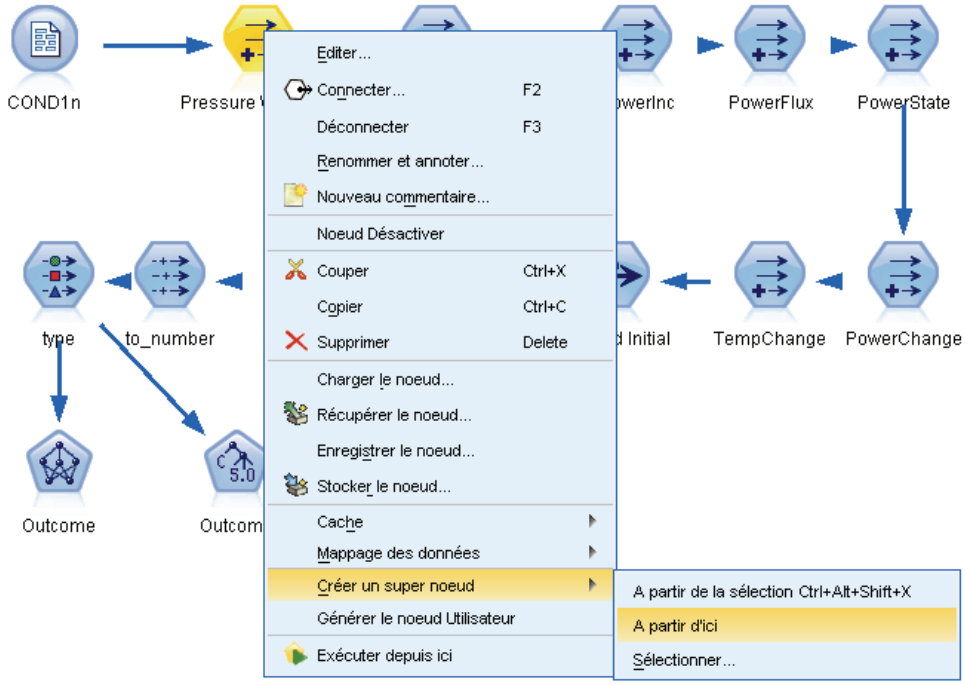
Vous pouvez également créer un super noeud en ne sélectionnant qu'un seul noeud et en utilisant les options du menu pour déterminer le début et la fin du super noeud, ou en encapsulant tous les noeuds se trouvant en aval du noeud sélectionné.

▶ Cliquez sur le noeud qui détermine le départ de l'encapsulation.

▶ Dans le menu Super noeud, sélectionnez :

Créer un super noeud > A partir d'ici

Figure 9-5
Création d'un super noeud à l'aide du menu contextuel pour la sélection



Vous pouvez également créer des super noeuds de manière plus interactive, en sélectionnant le début et la fin de la section du flux pour encapsuler les noeuds :

- Cliquez sur le premier ou le dernier noeud à ajouter au super noeud.
- Dans le menu Super noeud, sélectionnez :
  Créer un super noeud > SÃ©lectionner...
- Vous pouvez également utiliser les options du menu contextuel en cliquant avec le bouton droit de la souris sur le noeud souhaité.
- Le curseur prend la forme de l'icône de super noeud indiquant que vous devez sélectionner un autre endroit du flux. Déplacez-le vers le haut ou vers le bas, en direction de l'autre extrémité du fragment de super noeud et cliquez sur un noeud. Cette action remplace tous les noeuds situés entre les deux par l'icône en étoile du super noeud.

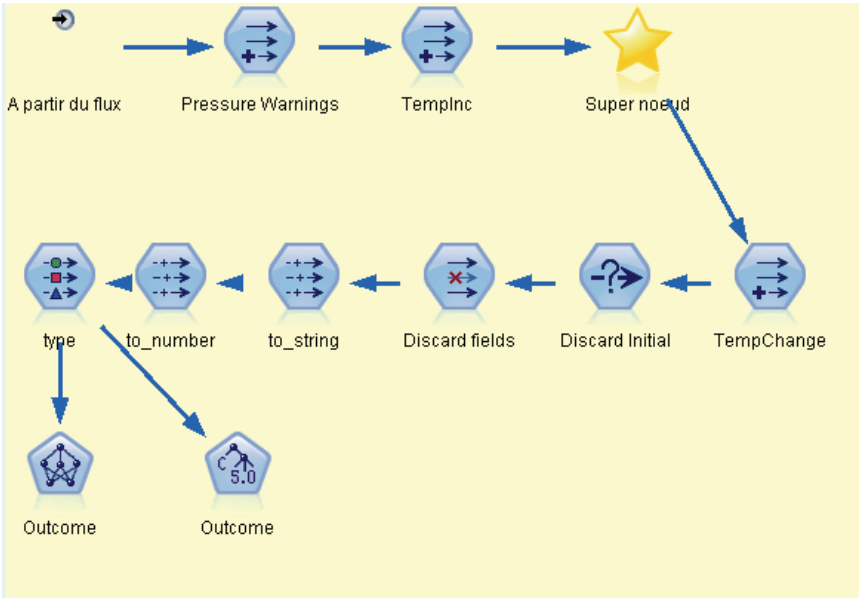
  Remarque : Vous devez sélectionner des noeuds provenant d'un flux continu ou bifurqué. Vous ne pouvez pas sélectionner des noeuds qui ne sont pas adjacents ou connectés.

Imbrication des super noeuds

Les super noeuds peuvent s'imbriquer dans d'autres super noeuds. Les mêmes règles pour chaque type de super noeud (source, exécution et terminal) s'appliquent aux super noeuds imbriqués. Par exemple, un super noeud d'exécution avec imbrication doit comporter un flux de données continu à

travers tous les super noeuds imbriqués afin de rester le super noeud d'exécution. Si l'un des super noeuds imbriqués est un super noeud terminal, les données ne circulent plus dans la hiérarchie.

Figure 9-6
Super noeud d'exécution imbriqué avec un autre Super noeud



Les super noeuds source et terminaux peuvent contenir d'autres types de super noeud imbriqué, mais les mêmes règles de base s'appliquent pour la création de super noeuds.

Exemples de super noeuds valides

La quasi-totalité des éléments que vous créez dans IBM® SPSS® Modeler peut être encapsulée dans un super noeud. Les exemples suivant représentent des super noeuds valides :

Figure 9-7
Super noeud d'exécution valide comportant deux connexions dans un flux valide

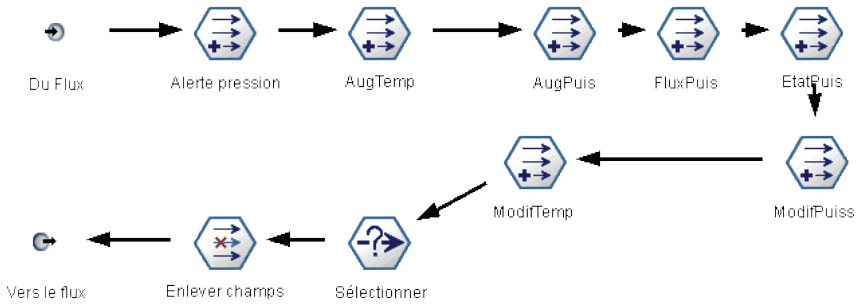


Figure 9-8
Super noeud terminal valide comprenant un flux séparé utilisé pour tester les modèles générés

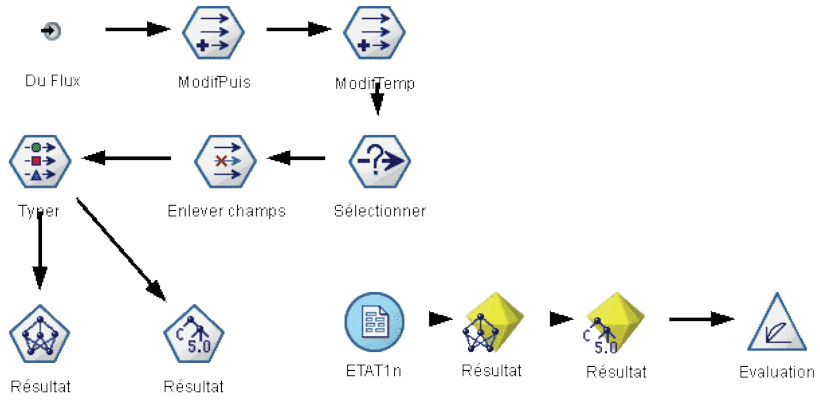
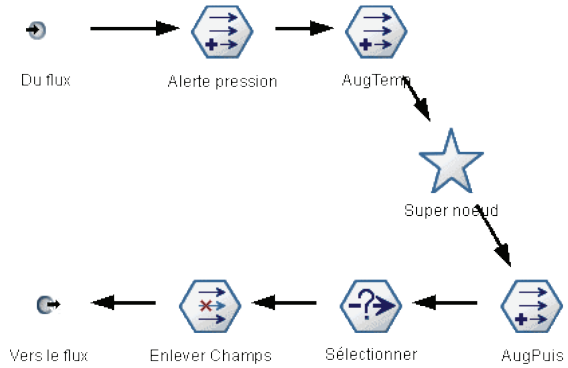


Figure 9-9
Super noeud d'exécution valide contenant un super noeud imbriqué



Exemples de super noeuds non valides

Lors de la création de super noeuds valides, il est très important de s'assurer que les données circulent de façon linéaire à travers les connexions du super noeud. Si deux connexions existent (un super noeud d'exécution), les données doivent circuler dans un flux, du connecteur de départ au connecteur d'arrivée. De la même manière, un super noeud source doit permettre aux données de circuler du noeud source au connecteur unique qui ramène les données vers le flux de zoom arrière.

Figure 9-10
Super noeud source non valide : Noeud source non connecté au chemin du flux de données

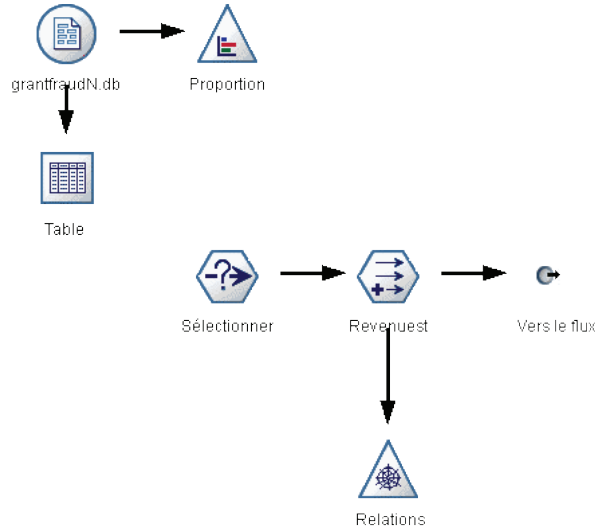
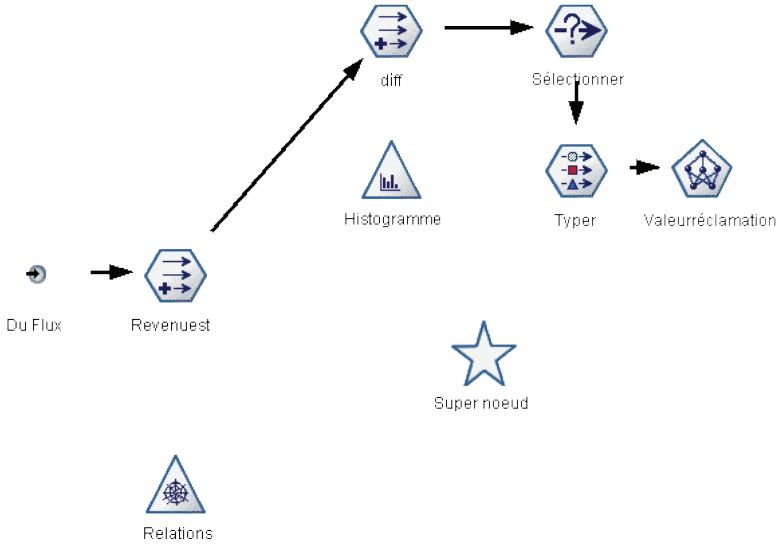


Figure 9-11
Super noeud terminal non valide : Super noeud imbriqué non connecté au chemin du flux de données



Verrouillage des super noeuds

Après avoir créé un super noeud, vous pouvez le verrouiller avec un mot de passe pour empêcher sa modification. Vous pouvez par exemple le faire si vous créez des flux, ou des parties de flux, en tant que modèles à valeur fixe destinés aux autres membres de votre organisation qui possèdent moins d'expérience dans la configuration d'enquêtes IBM® SPSS® Modeler.

Lorsqu'un super noeud est verrouillé, les utilisateurs peuvent toujours saisir des valeurs sur l'onglet Paramètres pour tous les paramètres qui ont été définis, et un super noeud verrouillé peut être exécuté sans saisir de mot de passe.

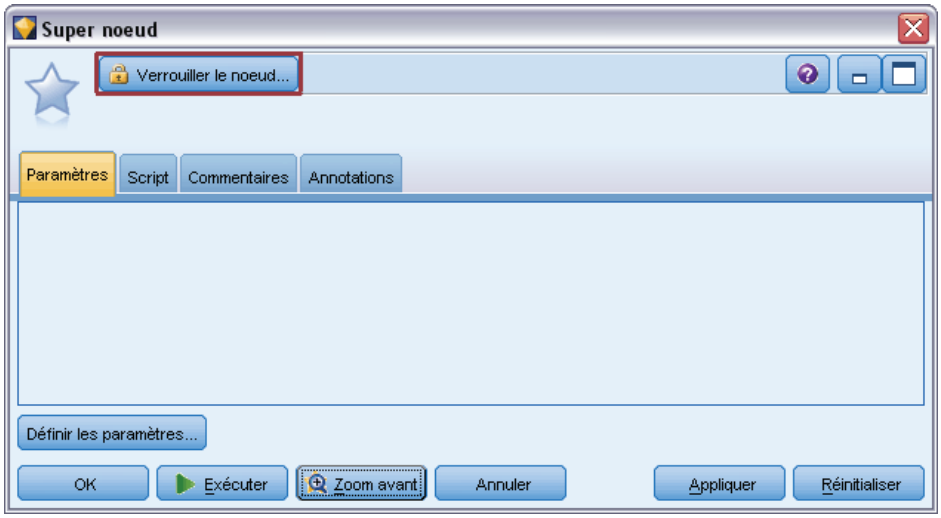
Remarque : Il est impossible d'effectuer un verrouillage ou un déverrouillage à l'aide de scripts.

Verrouillage et déverrouillage d'un super noeud

Avertissement : Les mots de passe oubliés ne peuvent pas être récupérés.

Vous pouvez verrouiller ou déverrouiller un super noeud sur n'importe lequel de ces trois onglets.

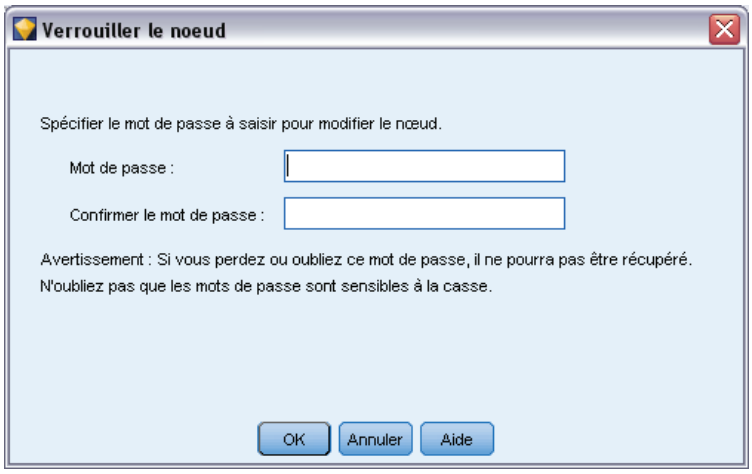
Figure 9-12
Verrouillage d'un super noeud



Cliquez sur Verrouiller un noeud.

Entrez et confirmez le mot de passe.

Figure 9-13
Entrez et confirmez le mot de passe du super noeud



- Cliquez sur OK.

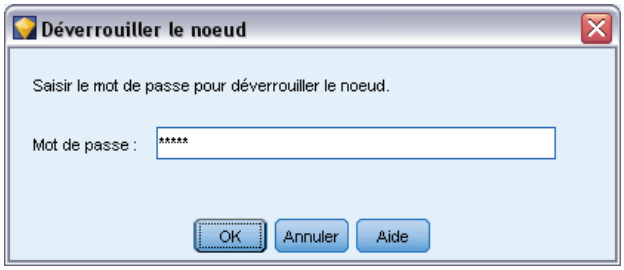
Un super noeud protégé par mot de passe est identifié sur l'espace de travail de flux par un petit symbole de cadenas en haut à gauche de l'icône du super noeud.

Figure 9-14
Super noeud source verrouillé dans un flux

Déverrouiller un super noeud

- Pour supprimer de manière permanente la protection par mot de passe, cliquez sur Déverrouiller le noeud ; vous êtes invité à saisir le mot de passe.

Figure 9-15
Saisie du mot de passe pour déverrouiller un super noeud

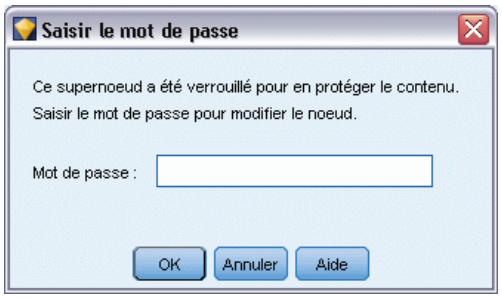


- Saisissez le mot de passe et cliquez sur OK ; le super noeud n'est plus protégé par mot de passe et le symbole de cadenas n'apparaît plus en regard de l'icône dans le flux.

Edition d'un super noeud verrouillé

Si vous essayez de définir des paramètres ou d'effectuer un zoom avant pour afficher un super noeud verrouillé, vous êtes invité à saisir le mot de passe.

Figure 9-16
Saisie du mot de passe pour effectuer un zoom avant ou éditer un super noeud



- Saisissez le mot de passe et cliquez sur OK.

Vous pouvez maintenant éditer les définitions de paramètre et effectuer un zoom avant ou arrière aussi souvent que nécessaire, jusqu'à ce que vous fermiez le flux dans lequel se trouve le super noeud.

Remarquez que cela ne supprime pas la protection par mot de passe, mais vous permet d'accéder au super noeud pour travailler dessus. Pour plus d'informations, reportez-vous à la section Verrouillage et déverrouillage d'un super noeud sur p. 524.

Edition de super noeuds

Après avoir créé un super noeud, vous pouvez l'analyser de plus près en effectuant un zoom avant ; si le super noeud est verrouillé, vous serez invité à saisir le mot de passe. Pour plus d'informations, reportez-vous à la section Edition d'un super noeud verrouillé sur p. 526.

Pour afficher le contenu d'un super noeud, vous pouvez utiliser l'icône de zoom avant située dans la barre d'outils IBM® SPSS® Modeler, ou la méthode suivante :

- Cliquez avec le bouton droit de la souris sur un super noeud.
- Dans le menu contextuel, choisissez Zoom avant.

Le contenu du super noeud sélectionné apparaît dans un environnement SPSS Modeler légèrement différent ; des connecteurs affichent le flot de données circulant dans le flux ou le fragment de flux. Sur ce niveau de l'espace de travail, vous pouvez effectuer différentes tâches :

- Modifier le type du super noeud (source, exécution ou terminal).
- Créer des paramètres ou éditer les valeurs d'un paramètre. Les paramètres sont utilisés dans la génération de scripts et les expressions CLEM.
- Indiquer des options de mise en cache pour le super noeud et ses sous-noeuds.
- Créer ou modifier le script d'un super noeud (super noeuds terminaux uniquement).

Modification des types de super noeud

Il peut être utile, dans certains cas, de modifier le type d'un super noeud. Cette option n'est disponible que si le zoom avant est activé dans un super noeud et elle ne s'applique au super noeud qu'à ce niveau. Il existe trois types de super noeuds :

Super noeud source	Une connexion en sortie
Super noeud d'exécution	Deux connexions : une en entrée et une en sortie
Super noeud terminal	Une connexion en entrée

Pour modifier le type d'un super noeud

- Assurez-vous que le zoom avant est activé dans le super noeud.
- Dans le menu Super noeud, sélectionnez Type de super noeud, puis choisissez le type voulu.

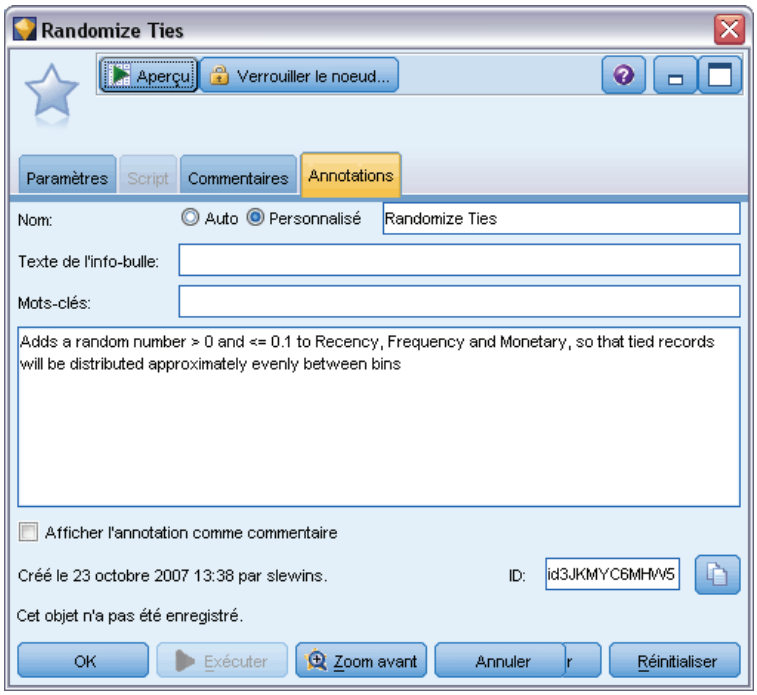
Annotation et changement de nom des super noeuds

Vous pouvez renommer un super noeud apparaissant dans le flux, et rédiger des annotations utilisées dans un projet ou un rapport. Pour accéder à ces propriétés :

- Cliquez avec le bouton droit de la souris sur un super noeud (zoom arrière activé) et sélectionnez Renommer et annoter.
- Vous pouvez également sélectionner Renommer et annoter dans le menu Super noeud. Cette option est disponible aussi bien en mode zoom avant qu'en mode zoom arrière.

  Dans les deux cas, une boîte de dialogue apparaît avec l'onglet Annotations sélectionné. Utilisez ces options pour personnaliser le nom affiché dans l'espace de travail du flux et fournir des informations concernant les opérations du super noeud.

Figure 9-17
Annotation d'un super noeud



Utilisation des commentaires avec les super noeuds

Si vous créez un super noeud à partir d'un noeud ou nugget commenté, vous devez inclure le commentaire dans la sélection pour créer le super noeud et que le commentaire y apparaisse. Si vous n'avez pas inclus le commentaire dans la sélection, il restera déconnecté dans le flux lors de la création du super noeud.

Lorsque vous développez un super noeud qui contenait des commentaires, ceux-ci sont restaurés à l'endroit où ils se trouvaient avant la création du super noeud.

Lorsque vous développez un super noeud qui contenait des objets commentés, mais que les commentaires n'étaient pas inclus dans le super noeud, les objets sont restaurés à l'endroit où ils se trouvaient mais les commentaires ne sont pas attachés à nouveau.

Paramètres du super noeud

Dans IBM® SPSS® Modeler, vous pouvez indiquer des variables définies par l'utilisateur, telles que *Minvalue*, dont les valeurs peuvent être spécifiées lorsqu'elles sont employées dans un script ou dans des expressions CLEM. Ces variables sont appelées des **paramètres**. Vous pouvez définir des paramètres pour les flux, les sessions et les super noeuds. Tous les paramètres définis pour un super noeud sont disponibles lors de la création d'expressions CLEM dans ce super noeud ou dans n'importe quel noeud imbriqué. Les paramètres définis pour les super noeuds imbriqués ne sont pas disponibles pour leur super noeud parent.

Vous devez suivre deux étapes pour créer et définir les paramètres des super noeuds :

- Définissez les paramètres du super noeud.
- Indiquez ensuite la valeur de chaque paramètre du super noeud.

Vous pouvez ensuite utiliser ces paramètres dans des expressions CLEM pour n'importe quel noeud encapsulé.

Définitions des paramètres de super noeud

Les paramètres d'un super noeud peuvent être définis en mode zoom avant comme en mode zoom arrière. Les paramètres définis s'appliquent à tous les noeuds encapsulés. Pour définir les paramètres d'un super noeud, vous devez d'abord accéder à l'onglet Paramètres de la boîte de dialogue du super noeud. Pour ouvrir la boîte de dialogue, utilisez l'une des méthodes suivantes :

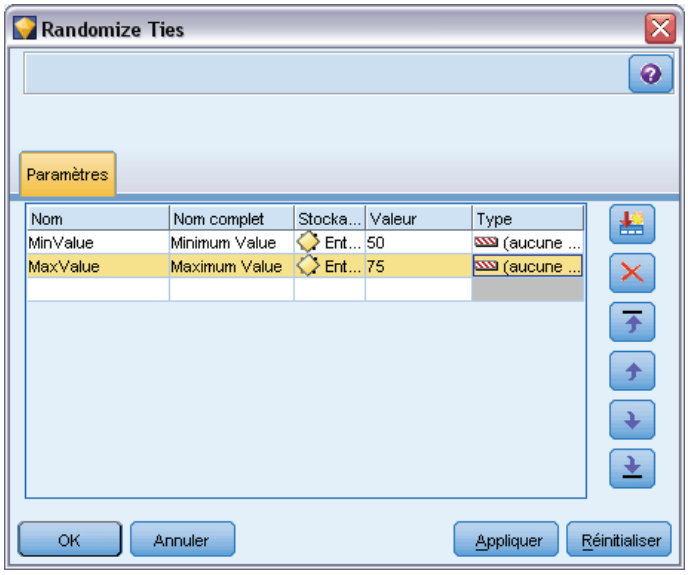
- Double-cliquez sur un super noeud du flux.
- Dans le menu Super noeud, sélectionnez Définir les paramètres.
- Si le zoom avant est activé dans le super noeud, vous pouvez également sélectionner Définir les paramètres dans le menu contextuel.

Dans la boîte de dialogue, l'onglet Paramètres affiche tous les paramètres définis précédemment.

Pour définir un nouveau paramètre

▶ Cliquez sur le bouton Définir les paramètres pour ouvrir la boîte de dialogue.

Figure 9-18
Définition les paramètres d'un super noeud



Nom. Les noms des paramètres sont répertoriés ici. Vous pouvez créer un paramètre en entrant un nom dans ce champ. Par exemple, pour créer un paramètre relatif à la température minimale, vous pouvez saisir Valeur min.. N'insérez pas le préfixe $P-, qui indique un paramètre dans

les expressions CLEM. Ce nom est également utilisé pour l'affichage dans le Générateur de formules de CLEM.

Nom complet. Répertorie le nom descriptif de chaque paramètre créé.

Stockage. Sélectionnez le type de stockage dans la liste. Le stockage indique le mode de stockage des valeurs de données dans le paramètre. Par exemple, si vous utilisez des valeurs commençant par des zéros à conserver (comme 008), vous devez sélectionner Chaîne comme type de stockage. Sinon, les zéros seront supprimés de la valeur. Les types de stockage disponibles sont les suivants : chaîne, entier, réel, temps, date et horodatage. Pour les paramètres de date, les valeurs doivent être définies à l'aide de la notation standard ISO telle qu'elle est présentée dans le paragraphe suivant.

Valeur. Indique la valeur actuelle du paramètre sélectionné. Modifiez ce paramètre selon les besoins. Pour les paramètres de date, les valeurs doivent être définies à l'aide de la notation standard ISO (soit, YYYY-MM-DD). Toute date définie dans un autre format est refusée.

Type (facultatif). Si vous prévoyez de déployer le flux vers une application externe, sélectionnez un niveau de mesure dans la liste. Sinon, il est conseillé de laisser la colonne *Type* en l'état. Si vous souhaitez spécifier des contraintes de valeur pour le paramètre, telles que des limites supérieures et inférieures d'un intervalle numérique, sélectionnez Spécifier dans la liste.

Vous ne pouvez définir les options de noms longs, de stockage et de type pour les paramètres que dans l'interface utilisateur. Il est impossible de définir ces options à l'aide de scripts.

Cliquez sur les flèches à droite pour déplacer le paramètre sélectionné vers le haut ou le bas de la liste des paramètres disponibles. Utilisez le bouton de suppression (indiqué par un *X*) pour supprimer le paramètre sélectionné.

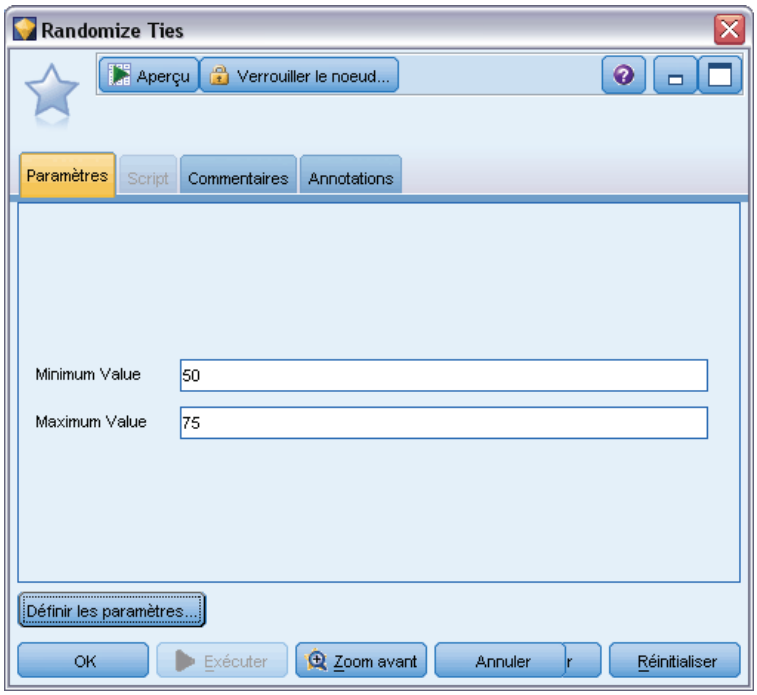
Définition des valeurs des paramètres de super noeud

Une fois que vous avez défini les paramètres d'un super noeud, vous pouvez spécifier les valeurs à l'aide des paramètres dans une expression CLEM ou un script.

Pour définir les paramètres d'un super noeud

- Double-cliquez sur l'icône Super noeud pour ouvrir la boîte de dialogue Super noeud.
- Vous pouvez également sélectionner Définir les paramètres dans le menu Super noeud.
- Cliquez sur l'onglet Paramètres. *Remarque* : les champs de cette boîte de dialogue sont ceux qui ont été définis à l'aide du bouton Définir les paramètres de cet onglet.
- Entrez une valeur dans la zone de texte pour chaque paramètre créé. Par exemple, vous pouvez définir la valeur *Valeur min.* sur un seuil d'intérêt particulier. Ce paramètre peut ensuite être utilisé dans de nombreuses opérations, telles que la sélection d'enregistrements au-delà ou en deçà de ce seuil pour une analyse plus approfondie.

Figure 9-19
Spécification des paramètres d'un super noeud



Utilisation des paramètres de super noeud pour accéder aux propriétés du noeud

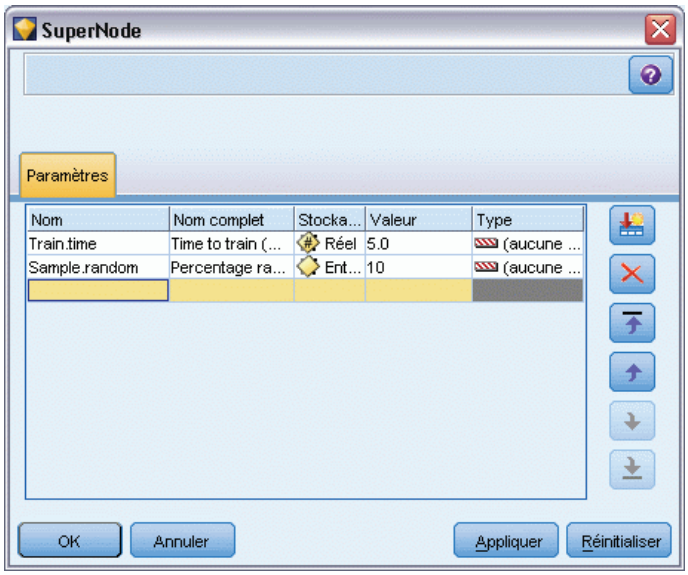
Vous pouvez également utiliser les paramètres de super noeud pour définir les propriétés de noeud (également appelées **paramètres de propriété**) pour les noeuds encapsulés. Par exemple, imaginons que vous souhaitiez spécifier qu'un super noeud forme un noeud R. neurones (Réseau de neurones) encapsulé pendant une durée déterminée en utilisant un échantillon aléatoire des données disponibles. Grâce aux paramètres, vous pouvez spécifier les valeurs concernant la durée et l'échantillon de pourcentage.

Figure 9-20
Fragment de flux encapsulé dans un super noeud

Le super noeud pris en exemple contient un noeud Echantillon appelé *Echantillonner*; et un noeud R. neurones (Réseau de neurones) appelé *Apprendre*. Vous pouvez utiliser les boîtes de dialogue du noeud pour définir le paramètre **Echantillonner** du noeud Echantillon sur % aléatoire et le paramètre **Critère d'arrêt** du noeud R. neurones (Réseau de neurones) sur Temps. Une fois ces options spécifiées, vous pouvez accéder aux propriétés du noeud avec les paramètres et indiquer

des valeurs particulières pour le super noeud. Dans la boîte de dialogue du super noeud, cliquez sur Définir les paramètres et créez les paramètres suivants :

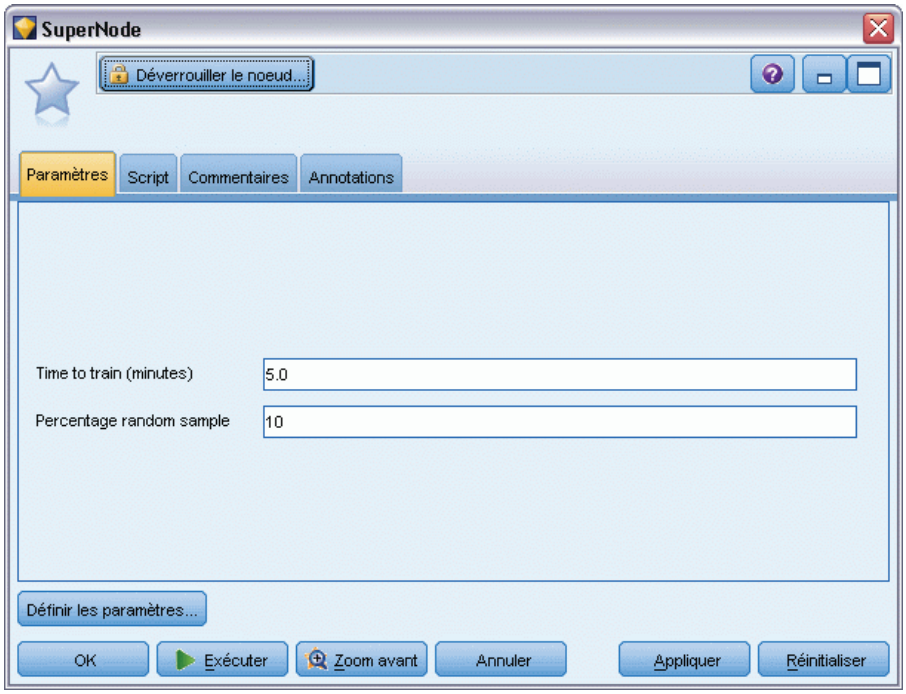
Figure 9-21
Définition des paramètres pour accéder aux propriétés du noeud



Remarque : les noms de paramètres tels que *Echantillonnage.aléatoire* utilisent une syntaxe correcte pour faire référence aux propriétés de noeud où *Echantillonnage* représente le nom du noeud et *aléatoire* une propriété de noeud.

Une fois que vous avez défini ces paramètres, vous pouvez facilement modifier les valeurs des propriétés des noeuds Echantillon et R. neurones (Réseau de neurones) sans rouvrir chaque boîte de dialogue. Au contraire, il vous suffit de sélectionner Définir les paramètres dans le menu du super noeud pour accéder à l'onglet Paramètres de la boîte de dialogue du super noeud, où vous pouvez choisir de nouvelles valeurs pour % aléatoire et Temps. Cette opération est particulièrement utile lorsque vous explorez les données pendant de nombreuses itérations de création de modèles.

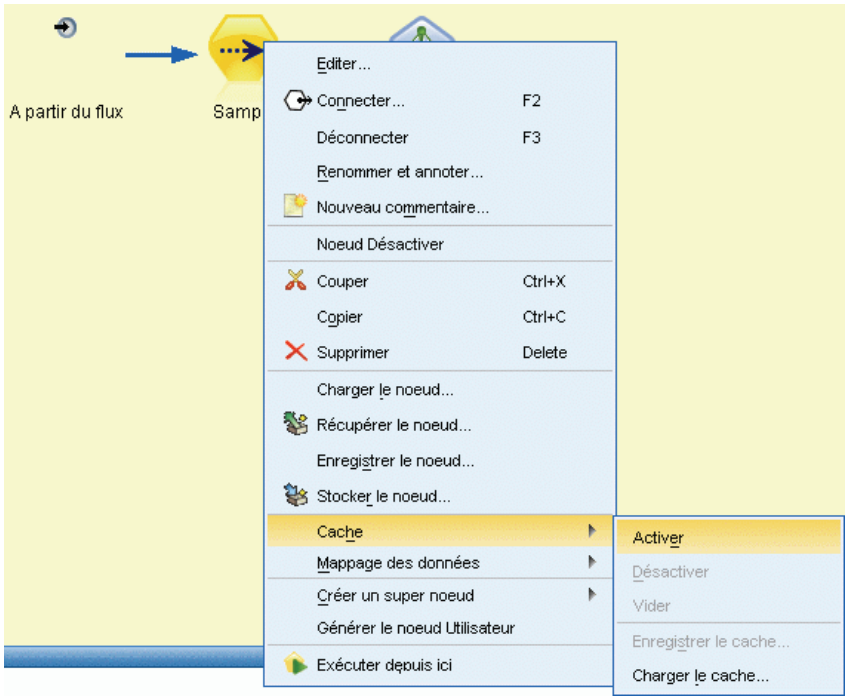
Figure 9-22
Spécifiez des valeurs pour les propriétés de noeud de l'onglet Paramètres de la boîte de dialogue du Super noeud



Super noeuds et mise en cache

A l'exception des noeuds terminaux, tous les noeuds peuvent être mis en cache depuis l'intérieur d'un super noeud. Pour effectuer une mise en cache, cliquez avec le bouton droit de la souris sur un noeud et sélectionnez une option dans le menu contextuel Cache. Cette option de menu est disponible depuis l'extérieur d'un super noeud, ainsi que pour les noeuds encapsulés au sein d'un super noeud.

Figure 9-23
Sélection des options de mise en cache pour un super noeud



Il existe plusieurs directives pour les caches de super noeud :

- Si la mise en cache est activée pour un noeud encapsulé, elle l'est également pour le super noeud.
- En désactivant le cache sur un super noeud, vous désactivez également le cache de *tous* les noeuds encapsulés.
- En activant le cache sur un super noeud, vous activez également le cache du dernier sous-noeud pouvant être mis en cache. En d'autres termes, si le dernier sous-noeud est un noeud Sélectionner, le cache sera activé pour ce noeud. Si le dernier sous-noeud est un noeud terminal (n'autorisant pas la mise en cache), le noeud suivant en amont prenant en charge la mise en cache est activé.
- Une fois que vous avez défini les caches pour les sous-noeuds d'un super noeud, toutes les activités en amont en provenance du noeud mis en cache, telles que l'ajout ou l'édition de noeuds, entraînent le vidage des caches.

Super noeuds et génération de scripts

Vous pouvez utiliser le langage de script IBM® SPSS® Modeler pour écrire des programmes simples servant à manipuler et à exécuter le contenu d'un super noeud. Vous pouvez, par exemple, indiquer l'ordre d'exécution d'un flux complexe. Par exemple, si un super noeud contient un noeud V. globales (Valeurs globales) qui doit être exécuté avant un noeud Nuage, vous pouvez créer un script qui exécute d'abord le noeud V. globales (Valeurs globales). Les valeurs calculées

par ce noeud, telles que la moyenne ou l'écart-type, peuvent être utilisées lorsque le noeud Nuage est exécuté.

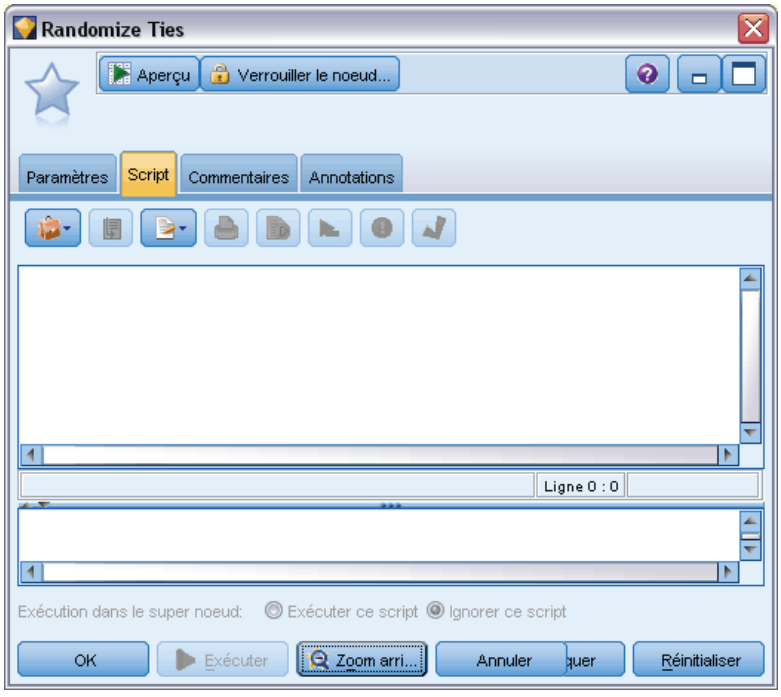
L'onglet Script de la boîte de dialogue du super noeud n'est disponible que pour les super noeuds terminaux.

Pour ouvrir la boîte de dialogue de script pour un super noeud terminal

- Cliquez avec le bouton droit de la souris sur l'espace de travail du super noeud et sélectionnez Script Super noeud.
- Sinon, que ce soit en mode zoom avant ou zoom arrière, vous pouvez sélectionner Script Super noeud dans le menu du super noeud.

Remarque : les scripts de super noeud sont exécutés uniquement avec le flux et le super noeud si vous avez sélectionné Exécuter ce script dans la boîte de dialogue.

Figure 9-24
Création d'un script pour un super noeud



Les options propres à la génération de scripts et son utilisation dans SPSS Modeler sont détaillées dans le *Guide de génération de scripts et d'automatisation* disponible dans le DVD de SPSS Modeler.

Enregistrement et chargement des super noeuds

Les super noeuds peuvent être enregistrés et réutilisés dans d'autres flux. L'extension utilisée pour l'enregistrement et le chargement des super noeuds est *.slb*.

Pour enregistrer un super noeud

- Effectuez un zoom avant dans le super noeud.
- Dans le menu Super noeud, sélectionnez Enregistrer le super noeud.
- Entrez un nom de fichier et un répertoire dans la boîte de dialogue.
- Indiquez si vous souhaitez ajouter le super noeud enregistré au projet en cours.
- Cliquez sur Enregistrer.

Pour charger un super noeud

- Dans le menu Insertion de la fenêtre IBM® SPSS® Modeler, sélectionnez Super noeud.
- Sélectionnez un fichier de super noeud (*.slb*) dans le répertoire ouvert ou accédez à un autre répertoire.
- Cliquez sur Charger.

  Remarque : les valeurs de tous les paramètres des super noeuds importés sont celles par défaut. Pour modifier les paramètres, double-cliquez sur un super noeud dans l'espace de travail.

Annexe

A

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

© Copyright IBM Corporation 1994, 2012.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

Toutes les données sur les performances contenues dans le présent document ont été obtenues dans un environnement contrôlé. Par conséquent, les résultats obtenus dans d'autres environnements d'exploitation peuvent varier de manière significative. Certaines mesures peuvent avoir été effectuées sur des systèmes en cours de développement et il est impossible de garantir que ces mesures seront les mêmes sur les systèmes commercialisés. De plus, certaines mesures peuvent avoir été estimées par extrapolation. Les résultats réels peuvent être différents. Les utilisateurs de ce document doivent vérifier les données applicables à leur environnement spécifique.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Toutes les déclarations concernant la direction ou les intentions futures d'IBM peuvent être modifiées ou retirées sans avertissement préalable et représentent uniquement des buts et des objectifs.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://www.ibm.com/legal/copytrade.shtml*.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Index

© Copyright IBM Corporation 1994, 2012.