POWERS: If you're a statistician, you know that statistical analysis systems aren't all created equal. Each one has their advantages and disadvantages. You might find yourself using SPSS for some tasks and R for other tasks. Sometimes you even have to mix and match the two to get your jobs done.

Welcome to This Week on developerWorks. I'm your host, Calvin Powers. My guest this week is Catherine Dalzell. She is a statistician and teacher at University of Ottawa. She's just published an article on developerWorks about how to integrate SPSS and R. Catherine, welcome to This Week on developerWorks. First, tell why you'd want to invoke R functions from your SPSS syntax file.

DALZELL: I think there could be several reasons. The main would be I think that you might want to use a function in R that doesn't currently exist in SPSS, because R is the development platform of choice for academic statisticians and others who are doing leading-edge work.

So, usually something will show up as an R function before it makes it to the commercial packages. And this way, the SPSS user can have access to all of that, that work. So, that would be certainly one reason; another reason would be if you're fundamentally an R programmer, an R user who wants to use SPSS and you're not really familiar with the SPSS syntax language.

So, this allows people who approach statistics from the R community to bring all that data management that they have, that they've already mastered, without having to deal with the SPSS syntax language, you know, which has a learning curve of its own. So, I think those would be the main reasons.

POWERS: Catherine, what's the most common way you see people trying to use R and SPSS data sets together?

DALZELL: Well, I think the most frequent way and kind of the simplest way is just to export your data from SPSS, and SPSS is a wonderful data vault, so quite often that's where you, if you're an analyst, you get your data from SPSS. You would export it using the .por format or .csv and read it into R using the appropriate R read function. And then you do your analysis in R and you get output from that and perhaps write an article on that basis.

POWERS: And what are some of the disadvantages of that approach?

DALZELL: Well, the disadvantages are that no data translation is really seamless and you lose things along the way. For the categorical data that are set up in SPSS with the label and the numeric code, you would have to choose whether you wanted to export the value name for the category or the numeric code, and then you would interpret it in R, set it up as categories there.

But you wouldn't actually have the same numeric code if you chose the label. And if you chose the label, if you chose the code, you would lose all that labeling information. So, that's messy. Times, time data is a bear. You would just get the number of seconds from the 10th of October or 1582. You'd lose a lot of that.

So, that would be a difficulty. And of course, it's a one-way trip because R has no ability to build an SPSS data set. So, if part of your analysis involved saving the results, residuals, predicted values, manipulating the data, there you are. You're stuck with basically and Excel format rather than SPSS.

POWERS: You've just published a developerWorks article called Calling R From SPSS, which documents a better way to do this integration. You want to tell us a little bit about the method you've documented in that article?

DALZELL: Well, this is a plug-in for SPSS that IBM has done to enable a much better transfer of the data to R. So, fundamentally R is called from within SPSS, and what the plugin basically does is give R access to a whole library of functions that enable you to transfer everything from SPSS, from the file, which is the data, the data dictionary, and the category information.

And you then read those into appropriate data frames in R, you would do your analysis, and you then have the ability with another set of functions within the plugin to build an SPSS database containing whatever you want so you don't have those problems of losing things in translation.
So, that’s what the plug-in does. And my article is taking people through the first steps, assuming that people have no familiarity with the plug-in, which I think is probably the case for many R users, just to get it up and running, talking to you, where do you get the plug-in and the first steps in getting the data talking into R, having R work with it, and then build a database in SPSS from that.

So, there's more you can do with the plug-in than I go through in the article, but this would at least get people going and they could actually do a lot just with those instructions.

POWERS: Thanks, Catherine. I really appreciate that. It does sound like a great how-to guide for people to get started, and it sounds like a great way for people to combine the best of both of those systems. That's all the time we have for This Week on developerWorks. Don't forget, we'll have a link to Catherine's article at ibm.com/developerworks/thisweek.

[ MUSIC ] [END OF SEGMENT]