SAS/SPECTRAVIEW® Software and Data Mining: A Case Study

Ronald Stogner and Aaron Hill, SAS Institute Inc., Cary NC

Abstract

Advances in information technology are allowing organizations to collect and store larger amounts of detailed data. Today's challenge is for companies to turn this information into a competitive advantage. Application areas, such as data mining, database marketing and decision support, hold the promise of turning this information into profits. One key component to this process is data visualization.

This paper presents a practical example of using data visualization technology to assist companies in the decision support process. SAS/SPECTRAVIEW® software is used to spot key relationships and trends in a large data set.

SAS/SPECTRAVIEW Software and Data Mining

Major companies today are faced with a growing dilemma. They have amassed large quantities of data related to their businesses and their customers. The challenge for these companies is to turn this data into profitable information. Today's desktop platforms possess the computing power to accomplish this task, but integrated software solutions that harness this power have trailed in development.

Past software solutions have primarily focused around single computing disciplines in their attempts to solve this problem. For example, past solutions have been purely statistical or purely graphical in nature. While useful, in and of themselves, these techniques are even more powerful when combined together as a part of a data mining solution. Data mining combines several disciplines in an attempt to provide a better solution for large data sets.

SAS® software, with its strong history of information delivery, broad range of integrated modules, and strong data warehousing capabilities, is uniquely positioned as a data mining solution. SAS Institute proposes a cost effective data mining solution referred to as SEMMA, or SAMPLE, EXPLORE, MODIFY, MODEL, and ASSESS. These terms are defined below.

SAMPLE: Extract a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly.

EXPLORE: Search speculatively for unanticipated trends and anomalies so as to gain understanding and ideas. This can reveal which subset of attributes will be the most productive to work with the modeling phase. Data visualization delivers intuitive tools for business professionals, while statistical techniques offer added detail for specialist.

MODIFY: The insights that are gained from exploration enable marketers to group the most productive subsets and clusters of data together for further analysis and exploration.

MODEL: This process involves searching automatically for a variable combination that reliably predicts a desired outcome. Data mining techniques such as neural networks, tree-based models and traditional statistical techniques can help reveal patterns in the data and provide a best fitting predictive model.

ASSESS: During this evaluation process, assessment of the results gained from modeling provides indications as to which results should be conveyed to senior management, how to model new questions that have been raised by the previous results and thus proceed back to the exploration phase.

Data visualization is one of the most versatile methods for performing the exploration phase of the data mining process. SAS/SPECTRAVIEW software is one of the tools available under The SAS System to perform data visualization. It allows the user to examine up to five variables at a time. Several new features have been added to assist in the data mining process.

SAS/SPECTRAVIEW Enhancements

Several new features have been added to the newest version of SAS/SPECTRAVIEW under release 6.12 of The SAS System. Features have been added to support data mining, some of which are discussed in more detail below.

Categorization

When reading data into SAS/SPECTRAVIEW there is a new option available, "Categorize". This allows the user to group independent variable values into data groups or "bins". This feature has two main uses. First, it allows the use of non-discrete values to be used as discrete or categorical data, a technique often applied to large data samples in data mining. Secondly, it also allows the user to reduce the amount of data that must be examined at one time. For example, if a customer's income is determined to be an independent predictor, categorization

can be used to reduce income into relevant ranges of income rather than examining individual income values. This not only makes the data more manageable to work with, but also decreases the complexity of the data.

Duplicate Value Handling

Another new option when loading data into SAS/SPECTRAVIEW is the handling of "duplicate" response values. Duplicate response values occur when multiple values of the dependent variable fall at the same independent variable location. For example, two people with the same income level, education and age have different debt levels. SAS/SPECTRAVIEW now allows the user to resolve these duplicates using simple statistics. Duplicate value handling is even more important when independent variables are categorized since duplicates are then more likely.

Visual Subsetting of Data

Another new feature for SAS/SPECTRAVIEW is the ability to visually subset data. This feature is useful during the exploration phase of data mining, where the user spots a key trend and would like to subset this data for further analysis.

Many of the visualization techniques used in the software such as cutting planes, point clouds, and volume rendering, now support the saving of data based on the visual displayed. For example, if a point cloud is displayed, a user can choose to save the data points that comprise the point cloud. Likewise, if a cutting plane were displayed, only points on that plane would be saved. Intersections of multiple visuals can also be saved. If for example, a cutting plane and a point cloud were displayed, then only points in the point cloud and on the cutting plane would be saved.

To visually subset, the user need only select the visual representation(s) of choice and once a trend of interest is found, select "save" and "data set". The data points that comprise the selected visuals are saved to a SAS data set and can be used for subsequent analysis in other SAS components.

Example Business Data

In this paper, we will examine a data set consisting of information on patient hospital stays in a southeastern hospital. Note that the actual city names will not be used here to protect the confidentiality of the company's data. The actual names have been replaced by the names of other unrelated cities. The data set contains information on the length of stay of the hospital patients, along with other related information.

In addition to the length of stay, the data set contains, the address of the patient, his or her home state and city, their age, race, and sex, as well as other information related to their visit.

For the purpose of this paper we will make the following assumptions. We will assume we work for a fictitious insurance company in a southeastern state. Since hospital stays tend to be expensive for insurance companies, our company will want to examine information concerning lengthy hospital visits. For this paper we will concentrate primarily on the exploration phase of the data mining process and will show how SAS/SPECTRAVIEW can be used to assist in this area. Finally, we will also assume that we are only concerned about information pertaining to those patients who reside in our home state, where our company provides medical coverage.

Exploring the Data

We start by applying a where clause to the data set when we load it into SAS/SPECTRAVIEW. We limit our exploration to only those individuals who reside within our state. This is easy to do in SAS/SPECTRAVIEW, we simply select the "Where" clause button, enter our where clause, and press enter.

We will then use the patient's home city, their age and race as our independent variables. We will examine their length of stay at the hospital as our dependent or response variable. Additionally, we will categorize the patient's age to place the patients into age groups and reduce the complexity of examining individual age values. Finally, we will take the mean of the length of stay, since we are primarily concerned with the average length of a patient's hospital stay.

To easily identify lengthy hospital visits, we will color code stays in excess of ten days as red and shorter stays as green. The brighter shades of green indicate shorter stays, while brighter shades of red indicate longer stays.

We start by examining particular cities that have high numbers of lengthy average hospital visits. That is, graphs with large amounts of red areas displayed. We use SAS/SPECTRAVIEW's cutting plane with a 3-D bar chart to hold a city constant and slice through the cities one at a time. Once we find a city that is different from the general trend (i.e. lot of red), we will use the new save to data set feature to save this information for further analysis later.

Note we could accomplish the former step by applying the appropriate where clause, however, this would require us to know something about our data set in advance. We will instead use SAS/SPECTRAVIEW to visually examine the data and decide based on our visual impression when a city is "of interest". This way we don't need to predefine some arbitrary "long" period of stay or some "high" number of instances of these long stays. We will subset the data visually based on the trends we see in the data rather than exact values.

After loading the data into SAS/SPECTRAVIEW and categorizing the patient age into 10 groups we have 6,600 data points from 17,657 observations. We select a bar chart to examine each city and then step through the various cities to examine general trends. If we find a city of particular interest we save that cities data to a data set via the “Save” button. Only values for that bar chart (city) will be written to the specified data set.

A few trends in the data set are immediately obvious. First, in general the lengthiest hospital stays generally correspond to the older age groups, as one might expect. Consider for example, figures 1 and 2 below, the average length of stay for a group of patients in the cities of Flint and Seattle, respectively.

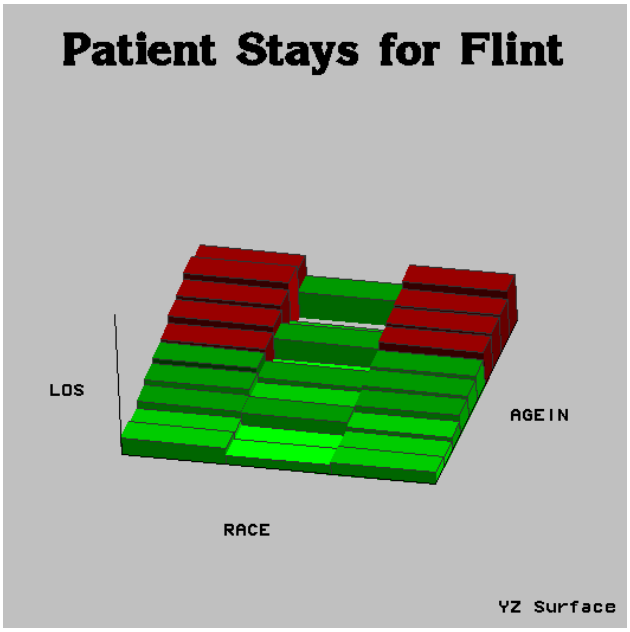

Figure 1 Average Patient Stays for Flint

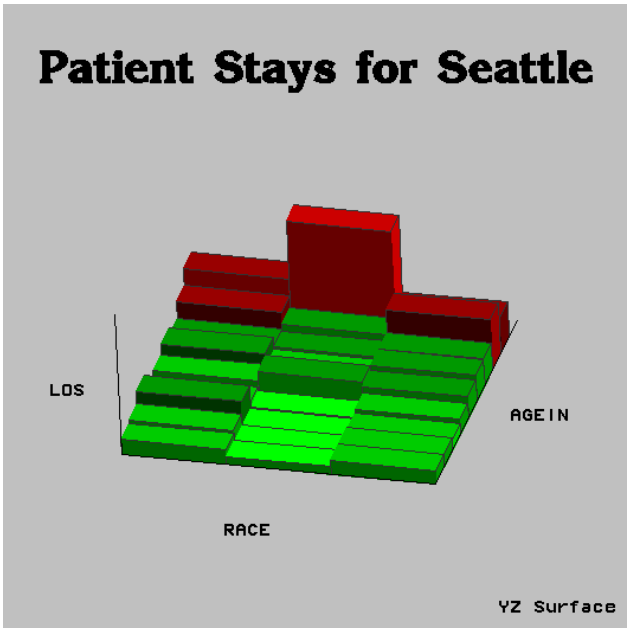

Figure 2 Average Patient Stays for Seattle

In both cites the longest hospital stays correspond to the older age groups. Note, however, that in Flint the difference between the age groups is less dramatic than that in Seattle. Also note that the red, or “long” stays occur at a much higher age in Seattle than in Flint. In Flint the first red bar occurs in the age group 53 to 64 years of age (determined by using SAS/SPECTRAVIEW’s Probe feature). In Seattle the first red bar occurs in the age group 74 to 85 years of age. Note also that in Seattle all races have long stays, while not all races have lengthy stays in Flint (middle race shown).

Note also that the ability to visualize the data is very important and can illuminate trends that might otherwise be missed. For example, assume we chose 52 years of age as a dividing age and calculated statistics on the two cities above. Using PROC MEANS and performing simple statistics on the two cities, once for each city for all persons 52 years of age and older and once for each city for those younger than 52 years of age yields the following statistics shown in table 1.

	Younger Than 52 Years of Age	52 Years of Age or Older
Flint	N = 5049 MEAN = 8.1 Days	N = 4282 MEAN = 13.5 Days
Seattle	N = 778 MEAN = 7.0 Days	N = 370 MEAN = 12.8 Days

Table 1 Average Patient Stays by Age for Flint and Seattle

The statistics for the two cities are quite similar. In both cities the mean length of stay for those younger than 52 years of age is approximately one week. While for those 52 years of age and older the average stay is approximately two weeks. Based on these numbers alone, one might assume the two cities are the same. Figures 1 and 2 above clearly indicate they are not.

In some cases there were differences based on the patients race, as in Hartford, shown in figure 3 below. Clearly the race shown on the left had longer average hospital stays than those to the right of the chart.

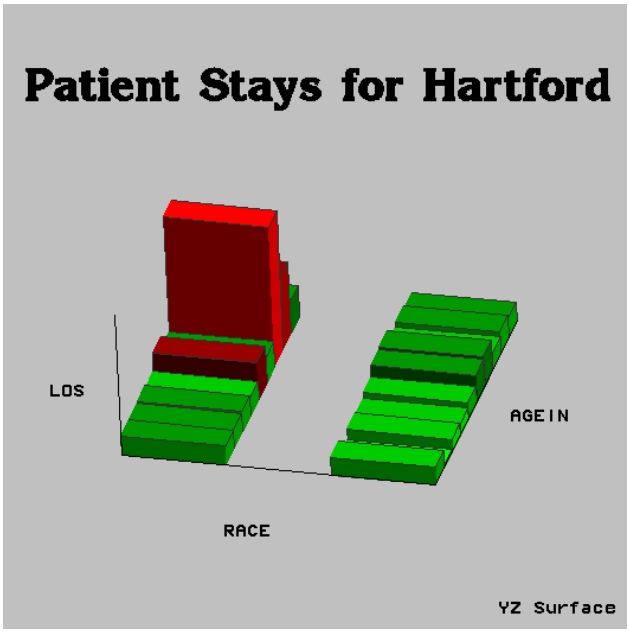

Figure 3 Average Patient Stays for Hartford

Using PROC MEANS, BY race, for the observations that comprise figure 3 yields table 2 shown below. Note that the average length of stay for the race shown on the left is nearly twice that of the patients shown on the right. It would be interesting to examine the factors that may have contributed to the longer hospital stays of those patients. By exploring the data in SAS/SPECTRAVIEW, other cities that exhibited similar trends could be examined along with those patients in Hartford.

Race	Average Stay
Left Side	N = 53 MEAN = 13.7 Days
Right Side	N = 135 MEAN = 8.5 Days

Table 2 Average Patient Stays by Race

Going Further

There are other ways to examine data in SAS/SPECTRAVIEW. The examples above, using cutting planes, are based on subsetting data by holding one independent variable constant. The effect of the two remaining variables can be then isolated. It is also possible to explore data based on the dependent or response variable. The point cloud is very useful for this type of analysis.

For example, in this data set, it is possible to isolate patients based on their length of stay. Using the point cloud, the user could select patients whose average stay fell between a range of two values. For example, the user could use a point cloud to display patients whose length of stay equaled or exceeded four weeks. The point cloud corresponding to this subset is shown in figure 4 below.

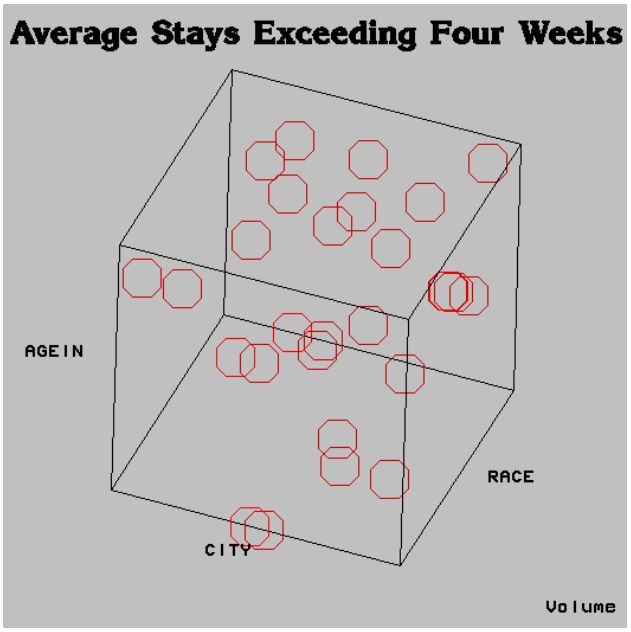

Figure 4 Patients Exceeding Four Week Stays

The user could then save this data to a SAS data set for further analysis. The advantage to subsetting visually is that here, the user could subset the data interactively until it was the desired size, for example only a handful of observations, and then save the data. An example partial listing of the above subset is shown below as written from SAS/SPECTRAVIEW, note race has been dropped from the listing.

City	Age In	Length of Stay
BLUEFIELD	90	28
DALLAS	58	28
CHICAGO	90	28
DETROIT	16	28
PEORIA	79	28
WINSTON	37	28
LONDON	5	30
PARIS	58	30
SALT LAKE	79	30
CARSON CITY	37	30
LOS ANGELES	90	31
ATLANTIC CITY	5	31
NEW YORK	69	32
WOODSTOCK	37	32

Table 3 Example Point Cloud Output Data Set

Conclusions

In this paper we have shown an example of how SAS/SPECTRAVIEW can be used during the exploration phase of the data mining process. The product was used

interactively and visually subset data based on visual analysis rather than predefined limits.

This paper has shown how two samples, similar in statistical makeup, can vary visually. Such attributes may be critical to determining key trends and correct samples in the data mining process. We have also demonstrated how many of the new data mining features in SAS/SPECTRAVIEW under release 6.12 of The SAS System can be used as part of a complete data mining solution.

New features like categorization and duplicate value handling allow a much broader range of data to be used within SAS/SPECTRAVIEW. Other features which were not discussed, like support for character variables, also enhance the use of SAS/SPECTRAVIEW as part of a data mining solution. Future enhancements to the product are planned to further the use of the product in the arena of data mining.

SAS/SPECTRAVIEW is a registered trademark of SAS Institute Inc. in the USA and other companies. ® indicates USA registration.

Authors

Ronald Stogner
Product Manager SAS/SPECTRAVIEW
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 677-8000
sasres@unx.sas.com

Aaron Hill
Marketing Analyst
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 677-8000
sascyh@unx.sas.com