Paper 173

Adding the Where to the Who

Jack Bulkley, Sanford Gale, Barry Hicks, Rob Stephens, SAS Institute Inc, Cary, NC

ABSTRACT

Part of knowing your customer, or prospect, is knowing where they live. The **where** can be used in reporting, data mining, and targeting your customers. A simple map showing your customers can be a powerful part of your reporting. Data mining can benefit at several stages, which are discussed in more detail below. Targeting customers can involve buying media. Many kinds of media, including newspapers, TV, and direct mail, can be used more effectively when you look at the areas covered by the media. This paper examines ways you can integrate spatial technology into your data warehousing and data mining strategies so that you can add the **where** to the who.

INTRODUCTION

Data mining is a valuable tool for looking at customer data and discovering who are your best customers. Enhancing this data to find out where they live can make your data even more valuable in several ways.

- Address standardization – enables better house-holding and cheaper mailing,
- Geocoding – allows mapping and spatial relationships,
- Demographic appending – describing a household or area.

You may have data that includes names and addresses. Because of differences in the way the values were given it may be tricky to decide when these values really represent the same person or people that are part of the same household. Standardizing the address can help in this process.

Geocoding adds geography values to your data. This allows you to see your data on a map and enables demographic appending.

Appending demographics gives you new ways to look at your data. Instead of seeing which counties had the most sales (it will usually be the counties with the largest population), you can look at the sales relative to the target population for your product. Just what your target population is varies depending on your product. It might be total population, adults, females, or something more specific like females 15 to 45 years of age.

SPATIALLY ENABLING THE WAREHOUSE

The warehousing process is a good place to add spatial values to your data. All of the normal advantages of warehousing would apply – wider availability, tracking time of last update, scheduled updates, etc. Some values in your data may already be spatially related, like zip code. Once you buy a zip code map, you would be ready to go. You can also add other geography values to data that has address values. This process is known as geocoding. Geocoding involves cleaning up your addresses – like changing Street, St. and Str. all to the standard St. Then the addresses are matched against a special database to add latitude, longitude and values that identify where area an address falls. Typical area values are county, MSA, census tract and block group.

Given an area value you can then add demographics for that area. We are running the data through a custom WA add-in to standardize the address, geocode, and append census block group demographics all in one pass. More information about this add-in will be available at SUGI 24. Standardized addresses are often used for reduced mailing rates, but they can also be useful for merging records together - in this case into households. The latitude and longitude generated during geocoding can be used to display points on a map. The geographic values (like census block group) can be used for summaries. These summaries can be used for reporting including maps and analysis.

The appended demographics like median income of census block group, can be used in analysis like that found in Enterprise Miner™.

The data feeding our effort is from product registration cards. Product registration cards are an imperfect source of data since important groups of customers may never use them. Still, they are an important customer contact and we want them captured and used in our warehouse. When bringing registration card data into our warehouse we want to tie it to any existing data for this customer. This is accomplished by matching the standardized addresses in a mapping that merges the registration data into the existing customers as represented in Figure 1.

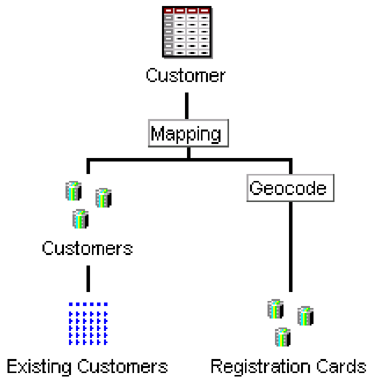

Figure 1: Geocoding Address Data

The geocoding step also adds values to identify the geographic area a customer falls in. Figure 2 represents a simple SAS/Warehouse Administrator® process for creating a summary table at the county level after geocoding the input data.

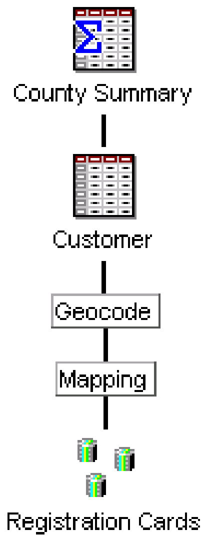

Figure 2: Creating a Summary

MINING WITH MAPS

SAS Institute defines data mining as the process used to reveal valuable information and complex relationships that exist in large amounts of data. Data mining is an iterative process-answers to one set of questions often lead to more interesting and more specific questions. To provide a methodology, in which the process can operate, SAS Institute further divides data mining into five stages that are represented by the acronym SEMMA. Beginning with a statistically representative sample of data, the SEMMA methodology- which stands for Sample, Explore, Modify, Model, and Access- makes it easy for business analysts to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. Here is an overview of each step in the SEMMA methodology:

- **Sample** the data by creating one or more data tables. The samples should be big enough to contain the significant information, yet small enough to process quickly.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- **Modify** the data by creating, selecting, and transforming the variables to focus the model selection process.
- **Model** the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.

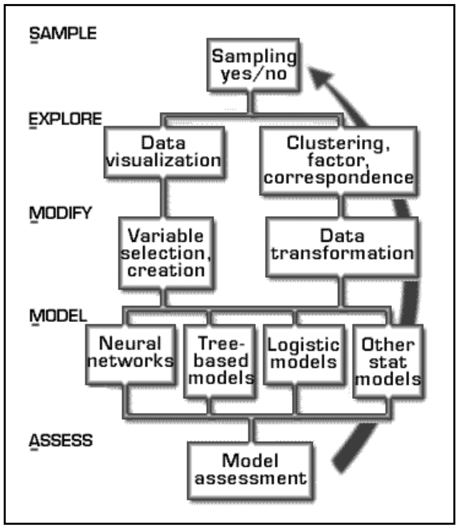

Figure 3: Steps in the SEMMA Methodology

SEMMA is itself a cycle; the internal steps can be performed iteratively as needed. Figure 2 illustrates the tasks of a data mining project and maps those tasks to the 5 stages of the SEMMA methodology. Projects that follow SEMMA can sift through millions of records and reveal patterns that enable businesses to meet data mining objectives such as

- Segmenting customers accurately into groups with similar buying patterns
- Profiling customers most effectively for individual relationship management
- Dramatically increasing response rate from direct mail campaigns
- Identifying the most profitable customers and the underlying reasons they choose you
- Understanding why customers leave for competitors (attrition, churn analysis)
- Uncovering factors affecting purchasing patterns, payments and response rates
- Increasing profits by marketing to those most likely to purchase
- Decreasing costs by filtering out those least likely to purchase.

Spatial mining involves using the spatial values during the SEMMA process. Each step of the SEMMA process can involve the spatial data, but some are more natural. Our example looks at the product registration data that is geocoded with demographic appending during the warehousing process. If this had not already been done to the data, it could be done independently using the same tools before beginning the SEMMA process or even as part of the process using a custom code node in Enterprise Miner.

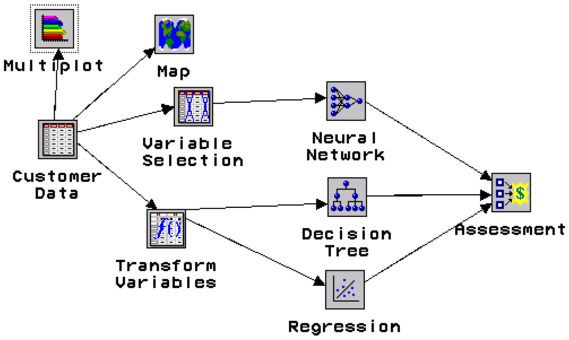

Figure 4: Data Mining Project

SAMPLE

In our case we did not take a sample at all because the total amount of data was not very large. It is possible to do a stratified sample based on geography. This would ensure that your sample has data proportionately representing all geographic areas. To do this you use the sampling tool, set the method to stratify and choose a geography variable, like MSA (metropolitan statistical area), to use for the stratification. For example, since the Dallas/Fort Worth area is about twice as large as the Raleigh/Durham area the sample would have about twice as many observations for Dallas/Fort Worth as for Raleigh/Durham.

EXPLORE

Exploring is an excellent place to use a map. Once the data is geocoded, you can view your customers as points on the map or summarize them by various geographies and look at them that way. In Figure 5 you can see the map showing each customer as a point.

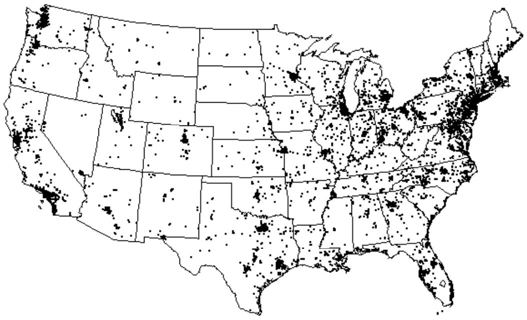

Figure 5: Customer Locations

What you will notice is very common. There are lots of customers where there are lots of people. For example southern California, Houston, and along the coast from Washington, DC to Boston. In this hardcopy version it is impossible to see all of the detail you might want to see. Fortunately the actual tool is interactive allowing you to zoom in and look at areas in more detail. Figure 6 shows a view of the greater Seattle area from the same session after you zoom in.

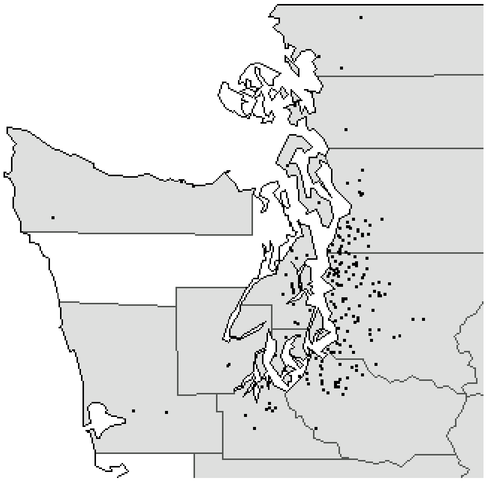

Figure 6: Seattle Area Customer Locations

You can use a summary to display a thematic map. But first we will look at modifying values.

MODIFY

The thematic map in Figure 7 shows the relative strength of each county as a market. This value is calculated in two steps. First the market penetration is calculated as the number of customers divided by the total number of households in each county. The second step is to compare each county's market penetration to the average market penetration and assign it to one of four categories:

1. low – less than 80% of average,
2. medium – from 80% to 120% of average,
3. high – from 120% to 200% of average,
4. very high – over 200% of average.

Again you are looking at a static map while the system is actually interactive and color. This gives you much more flexibility to explore the map, add labels for the cities, query the under-lying data, and other features.

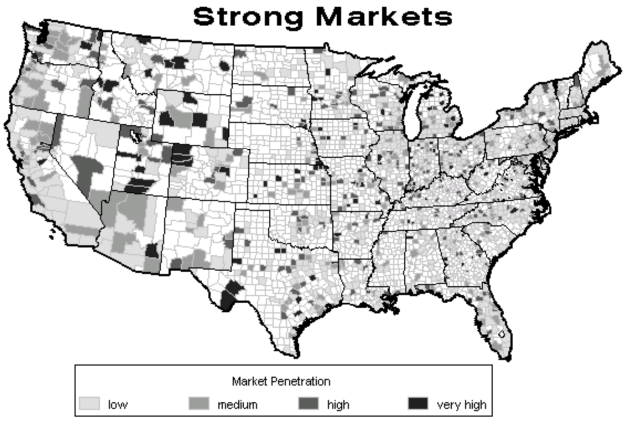

Figure 7: Market Strength Map

For our example, we wanted to look at the sales per person for various counties. Our summary data table had total sales for each county and we added county level demographics. Then we used the transformation tool, labeled *Transform Variables* in Figure 4, to create a variable for sales per person.

MODEL

We used three techniques - regression, neural net, and decision tree - to try and model sales per person in each county. The results of each model can be viewed to see what input variables were important for predicting the output. To compare the effectiveness of each model we used the assessment tool.

ASSESS

The assessment tool, labeled *Assessment* in Figure 4, can be used to compare the lift of different models.

CONCLUSION

Adding the where to the who is all about adding spatial values to your data. Spatial values can be added at anytime, but adding them during the warehouse loading process gives you all of the advantages of warehousing and makes these values available to a wider number of people and applications in your enterprise. After you add spatial values to your data, you can display your data on various maps. But you can also use the additional values for summaries, reporting, analysis, and calculating new variables. The spatial component is a powerful way to add value and gain better insights into your data.

ACKNOWLEDGMENTS

Thanks to various developers working on SAS/Enterprise Miner and SAS/Warehouse Administrator who answered my questions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jack Bulkley
SAS Institute Inc
100 SAS Campus Drive
Cary, NC 27513
Email: sasjtb@wnt.sas.com