Paper 3200-2015

The Use Of SAS® Maps For University Retention And Recruiting

Allison Lempola, South Dakota State University;
Dr. Thomas Brandenburger, South Dakota State University

ABSTRACT

Universities often have student data that is difficult to represent, which includes information about the student's home location. Often, student data is represented in tables, and patterns are easily overlooked. This study aimed to represent recruiting and retention data at a county level using SAS® mapping software for a public, land-grant university. Three years of student data from the student records database were used to visually represent enrollment, retention, and other predictors of student success. SAS® Enterprise Guide® was used along with the GMAP procedure to make user-friendly maps. Displaying data using maps on a county level revealed patterns in enrollment, retention, and other factors of interest that might have otherwise been overlooked, which might be beneficial for recruiting purposes.

INTRODUCTION

University student data bases contain information about each student which can be further summarized into information by state and county. This information can be displayed using SAS® tools to create choropleth maps. This paper aims to show how SAS programmers can use these tools to create professional graphs for institutional research purposes.

A student database is used which contains relevant information for each student including ZIP code, high school grade point average,etc. The SASHELP.ZIPCODE file is used to convert student's home ZIP codes into both state and county FIPS codes. In conjunction with the student data, SAS/GRAPH® map datasets are used which contain U.S. state boundaries along with the county borders within each state.

The data in the following examples contain three years of student information for first-time, full-time Bachelor's cohorts with the vast majority of students having home addresses located within the upper Midwest. Possible institutional research applications are discussed using the relevant student data. Maps include student population, retention rate, mean high school grade point average, and average ACT scores by county.

FORMATTING THE RESPONSE DATASET

Student databases often do not include State and County FIPS codes, but do contain student ZIP codes. The SASHELP.ZIPCODE file contains ZIP code level information for the United States and will convert ZIP codes to state and county FIPS codes, to be used in the GMAP procedure.

If the original data set is named Student_Data with one row per student with relevant information and ZIP codes stored under the variable name zip, the ZIP codes can be converted to state and county FIPS codes with the following SAS® code:

```
data With_FIPS;
   set Student_Data;
   set sashelp.zipcode (keep=zip county state) key=zip/unique;
   if _error_ then do;
   call missing(county, state);
   _error_ = 0;
end;
run;
```

The dataset With_FIPS now is a dataset with one row per student with relevant student variables (ACT score etc), the state FIPS code for each student (state), and the FIPS county code (county). This dataset can be reformatted to summarize specific variables (such as High School grade point average) by county using the SQL procedure. This response dataset will be used to map the student data. The response dataset is made using the following code:

```
proc sql;
   create table By_County as
   select county,
   state,
   count(distinct ID) as Population,
   mean(ACT) as MeanACT,
from Student_Data
group by state, county;
run;
```

The table By_County will contain any relevant variables. For example, MeanACT contains the average comprehensive ACT score in each county and Population counts the population in each county. Any other variables of interest can also be used. For example, mean high school grade point average, mean estimated family contribution, or retention rate by county.

USING THE GPROJECT PROCEDURE

In this paper, student data is to be mapped by county. SAS supplies numerous maps that can be used in a multitude of ways including the all states of the United States, selected groups of states, counties within selected states, or single states. The map data set MAPS.US contains the map data set with a state map of the United States. Similarly, the map data set MAPS.COUNTIES produce county maps of any selected states.

A majority of students in the dataset have home addresses in the upper Midwest region. Therefore, the maps are limited to the counties in North Dakota, Minnesota, South Dakota, Nebraska, Iowa, Wisconsin, and Illinois which included 95.41% of the original first-time-full-time Bachelor's cohorts. The number of students considered in the study is 5,856.

The GPROJECT procedure is used to create a projected map of counties in selected states. Without the GPROJECT procedure, the GMAP procedure (used later in the paper) would result in a distorted map as the GMAP procedure uses longitude/latitude coordinates based on spherical coordinates instead of a flat surface (distortion does not occur for maps with only states and not counties).

The following code will create a map data set named County_Map containing projected coordinates for the counties in North Dakota, Minnesota, South Dakota, Nebraska, Iowa, Wisconsin, and Illinois:

```
proc gproject
   data=maps.counties
   out=County_Map;
   id County;
   where state in (46 27 38 19 31 55 17);
run;
```

The state FIPS code for the chosen states were used to determine which states would be in the map data set. The dataset County_Map will be used throughout the rest of the examples to map various student information of interest by county in the selected states.

STUDENT POPULATION MAP

The following code produces a choropleth map summarizing the number of students from each county:

```
proc format;
   value Count
   1='1'
   1-5='1-5'
   5-15='5-15'
   15-100='15-100'
   100-high='Above 100';
run;
```

```
pattern1 v=ms c=CXFFFFCC;
pattern2 v=ms c=CXC7E9B4;
pattern3 v=ms c=CX41B6C4;
pattern4 v=ms c=CX2C7FB8;
pattern5 v=ms c=CX253494;

proc gmap
   data= By_County
   map= County_Map
   all;
   id county state;
   format Population Count.;
   choro Population/discrete;
run;
```

The FORMAT procedure is used to group counties with similar numbers of students. Since five groups were defined, five patterns (colors) are used to color the map with solid shades. The colors are defined via hexadecimal notation.

The GMAP procedure is used to create the following map. The format option is needed to format the coloring as described within the FORMAT procedure on the variable Population which contains the number of students from each county. The DISCRETE option is used to control the formatting of the groups.

Figure 1 describes student population by county.

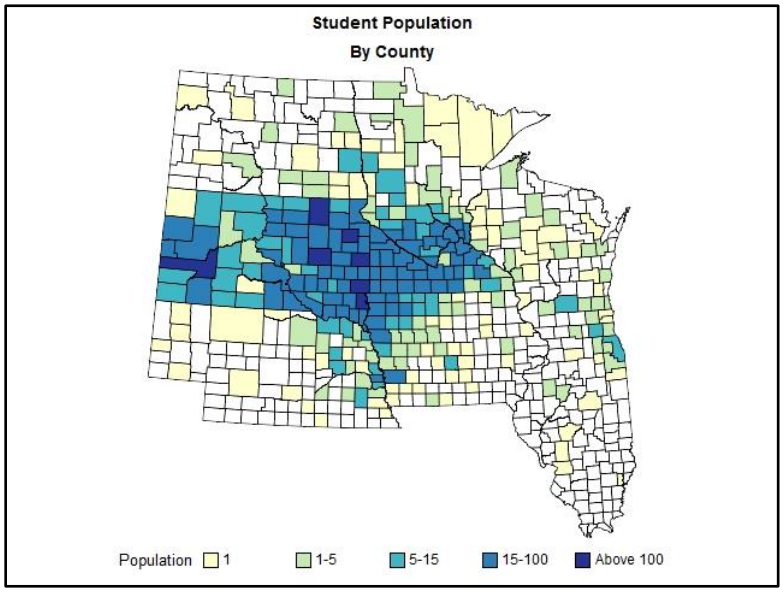

Figure 1

Although institutions often know general home locations of students, it is beneficial to concretely visualize the home locations of students. County by county student populations could be represented in a table; however this would result in an unreasonably large amount of information to sort through. Questions like the following could be answered using a student population map:

- Are there large metropolitan areas we expect to see a higher student population from?
- Are there states from which a larger student population is expected?

- What is the student population from the county of the University compared to the neighboring counties?
- Are there any counties from which a surprisingly large number of students list as their home address?

RETENTION RATE MAP

Retention rate is defined as the percentage of first-time-full-time students who continue to study full-time at the institution their second year. The overall retention rate of the three years of first-time, full-time students studied was 74.67%. The map in Figure 2 following describes the by county retention rate and was created using identical code as above, with exception to changing the variable name in the GMAP procedure and changing the format in the FORMAT procedure.

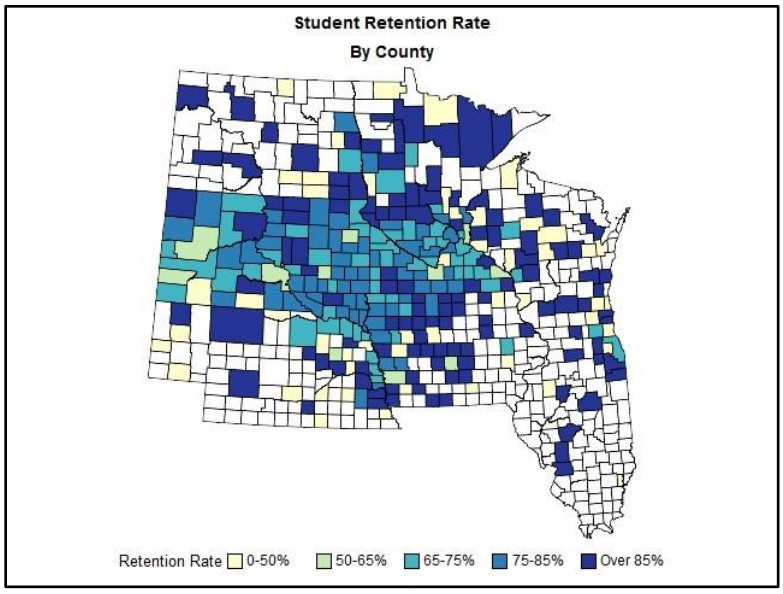

Figure 2

Institutions often know overall retention rate of cohorts but retention rate could fluctuate based on the home location of the student. Note that the map should be used in conjunction with the population map, as some rates will appear very high or low depending on the number of students from that county. Questions such as the ones below could be investigated through a retention rate map:

- What is the retention rate of the county where the university is located? Is it higher or lower than average?
- Does retention rate appear to increase or decrease as distance from campus increases?
- Is the retention rate higher or lower in metropolitan areas?
- Are there any areas with particularly higher than average or lower than average retention rates?

AVERAGE COMPREHENSIVE ACT SCORE MAP

Comprehensive ACT score is often analyzed for incoming freshman. The average comprehensive ACT score for this data set was 23.04. The map in Figure 3 below summarizes the average comprehensive ACT score by county for the selected states. The code for making the map is identical to as above, with exception to changing the variable name in the GMAP procedure and changing the format in the FORMAT procedure.

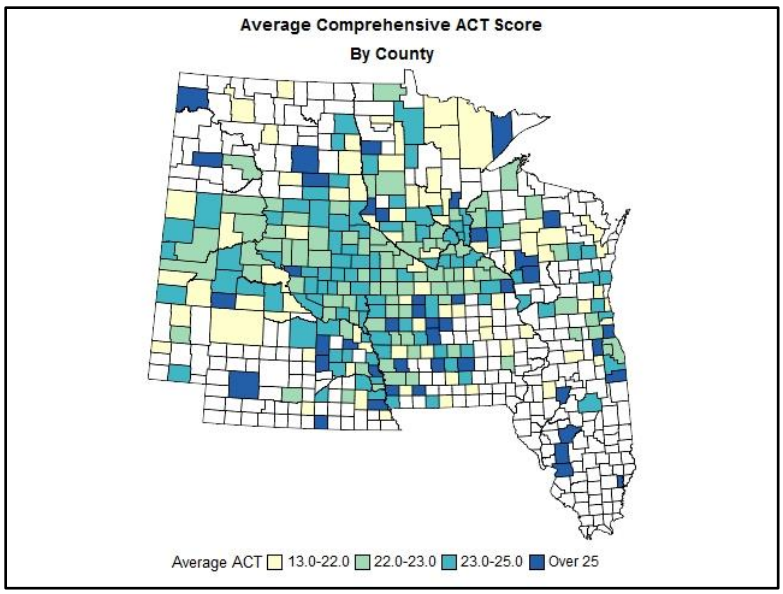

Figure 3

Institutions often know average comprehensive ACT scores for incoming freshman. But these scores could fluctuate geographically based upon the home location of the student. Note that the map should be used in conjunction with the population map, as some averages may be misleading depending on the number of students from the county. Questions such as the ones below could be investigated through the map:

- What is the average ACT score of students whose home county is the same as that of the University? Is it higher or lower than average?
- Do average ACT scores appear to increase or decrease as distance from campus increases?
- Are there any areas with particularly higher than average or lower than average ACT scores?
- What are the average ACT scores of counties with large metropolitan areas?

AVERAGE HIGH SCHOOL GPA SCORE MAP

High school grade point average is common in analyzing a cohort of incoming freshmen. The map in Figure 4 summarizes the average high school grade point average by county for the selected states. The code for making the map is identical to as above, with exception to changing the variable name in the GMAP procedure and changing the format in the FORMAT procedure.

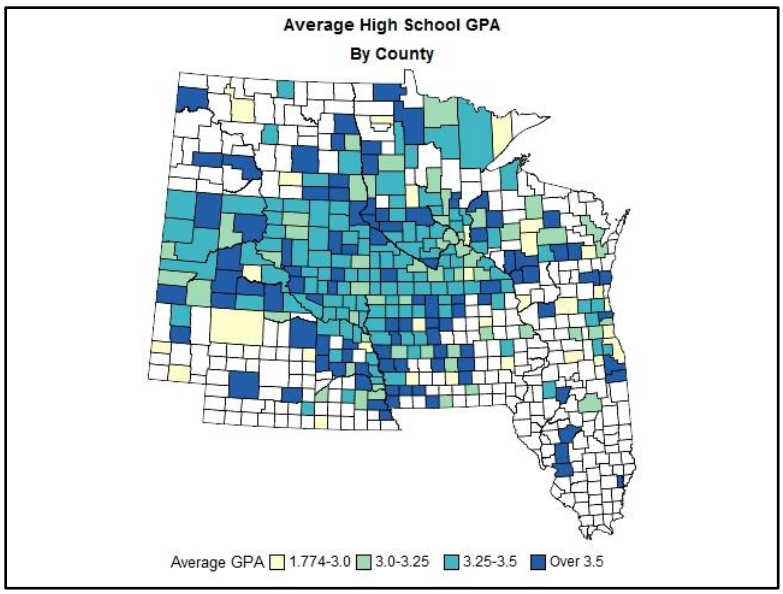

Figure 4

The average high school grade point average for the three years of student data was 3.325. Like ACT scores, high school grade point averages could fluctuate based on the home location of the student. Once again, the map should be used with caution when analyzing averages for counties with small numbers of students. Question like the following could be investigated through the map:

- What is the average high school grade point average in the county where the university is located? Is it higher or lower than average?
- Do average high school grade point averages appear higher or lower as distance from campus increases?
- What is the average high school grade point average from which a large student population comes from?

CONCLUSION

This paper made use of the many different many different features of SAS® software. This included the SASHELP.ZIPCODE file which converted student's home ZIP codes into both state and county FIPS codes which were used to create a response data set with all needed variables. SAS/GRAPH® map datasets were used to create meaningful maps of the counties within selected states.

The maps described could be used in a variety of ways to better understand the student population at a university. This information could further be utilized in university recruitment to target areas from which to recruit from or analyze other student data such as estimated family contribution. Furthermore, this information could be analyzed to study retention with the aim of increasing student success. Similar maps could be used in a variety of fields of study to study customers, students, or other populations.

REFERENCES

Hadden, Louise S and Zdeb,Mike. 2010. "ZIP Code 411: Decoding SASHELP.ZIPCODE and Other SAS® Maps Online Mysteries" *Proceedings of SAS Global Forum 2010*. Cary, NC. SAS Institute Inc Available at http://support.sas.com/resources/papers/proceedings10/219-2010.pdf

Zdeb, Mike. 2004. "The Basics of Map Creation with SAS/GRAPH®." *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*. Cary, NC. SAS Institute Inc. Available at http://www2.sas.com/proceedings/sugi29/251-29.pdf.

RECOMMENDED READING

- *http://colorbrewer2.org/*
- *http://support.sas.com/documentation/cdl/en/graphref/63022/HTML/default/viewer.htm#gmap-proc-statement.htm*

CONTACT INFORMATION

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

Allison Lempola
allylempola@gmail.com