

USING SAS FOR SPATIAL ANALYSIS

Christiana Petrou, University of Louisville, Louisville, KY

ABSTRACT

The aim of this paper is to identify major trends in several cardiac and respiratory conditions, and to analyze correlations such as median income and geographical locations for Jefferson County patients in Kentucky. This study dealt with geographical interpretation and statistical data manipulation. ArcGIS was used to compile, author, analyze, map, and publish geographic information and knowledge in combination with the statistical software package, SAS, using the SAS Bridge to ESRI developed jointly between the SAS Institute, Inc., and ESRI, Inc. This study used approximately 5000 patient records.

Using ARC toolbox, the proximity of patient to the nearest provider was computed using the Near command. Conducting text mining in Enterprise Miner, the patients were clustered according to similarities of the word descriptions of the DRG codes, or Diagnosis Related Group, which refers to a system of describing and classifying the hospitals' patients. Text Miner found five clusters, where each cluster has similar trends in words more so than any other cluster. Kernel density (PROC KDE) was used on the proximity variable by cluster to determine the distance to a healthcare provider by specialty. Results were compared to patient socio-economic status using census data.

INTRODUCTION

The science of learning plays a key role in the field of statistics and data mining. We are interested in learning from the data that we have available. Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and to understand what the data mean. The purpose of this paper is to examine the methodology and application of data mining, particularly text mining, in the context of performing a statistical analysis. Large amounts of data are available in a number of fields such as medicine, biology, finance, and marketing. It is a great challenge to understand these data and to develop efficient and effective statistical tools to extract meaningful information from the data. Much of the available data are in unstructured, text format and traditional statistical methods are inadequate to examine that data. This paper examines a project regarding the medical sector and using statistical tools, namely text mining, clustering, and data visualization. The study also enables further understanding of statistical methods and the application of statistical software.

The aim of this study was to identify major trends in several cardiac and respiratory conditions, and to analyze correlations such as median income and geographical locations for Jefferson County patients in Kentucky. This study used geographical interpretation and statistical data manipulation. ArcGIS was used to compile, author, analyze, map, and publish geographic information and knowledge in combination with the statistical software package, SAS, using the SAS Bridge to ESRI developed jointly between the SAS Institute, Inc., and ESRI, Inc. This study had 5000 patient records. In order to complete this project, the author had to learn to use the statistical software, SAS and the geographic software, ArcGis.

DATA MINING AND TEXT MINING

Data Mining is an analytic process designed to explore large amounts of data - typically business or market related data in search of consistent patterns or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction. The purpose of Text Mining is to process unstructured text information, to extract meaningful numeric indices from the text, and to make the information contained in the text accessible to the various data mining, statistical, and machine learning algorithms. Investigators who use text mining must recognize that a complete understanding of natural language text, a long-standing goal of computer science, is not immediately attainable. Instead, text mining focuses on extracting a small amount of information with high reliability. The information extracted might be the author, title and date of publication of an article, the acronyms defined in a text or the articles mentioned in the bibliography. Information can be extracted to derive summaries of the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, the investigator can analyze words or clusters of words used in documents, or analyze entire documents to determine similarities or relationships between them, or how they are related to other variables of interest in the data-mining project. In the most general terms, text mining will "turn text into numbers", which can then be incorporated in other analyses such as predictive modeling projects, and the application of unsupervised learning methods such as clustering.

THE SAS BRIDGE TO ESRI

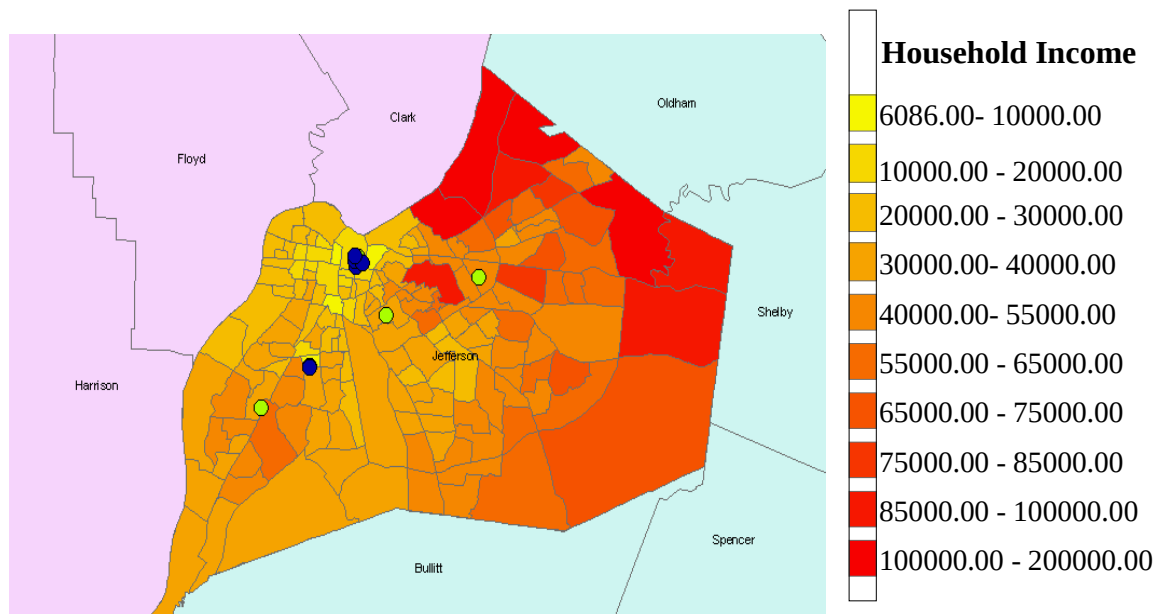
DATA MINING AND GIS TO EXAMINE HEALTH CARE PROVIDERS PROXIMITY TO PATIENTS IN JEFFERSON COUNTY, KENTUCKY.

The aim of this study was to identify major trends in several cardiac and respiratory conditions, and to analyze correlations such as median income and geographical locations for Jefferson County patients. The health conditions were coded through the DRG code list. DRG stands for Diagnosis Related Group, and refers to a system of describing and classifying the hospitals' patients. All possible principal DRGs are divided into 25 mutually exclusive categories referred to as Major Diagnostic Categories (MDCs), which are loosely based on organ systems. This study contained MDC 4 (respiratory diseases) and MDC 5 patients (circulatory/heart diseases and conditions).

Examples of Respiratory Diseases found in this study were bronchial asthma, or an allergic condition resulting from the reaction of the body to one or more allergens. It is one of the most fatal respiratory diseases. Other examples include pleural effusion or "water on the lungs" which is too much fluid collected in the space between the lungs; and pneumonia or an inflammation of the lung, usually caused by infection. Examples of Cardiovascular Diseases found in this study are heart failure when the heart does not pump enough blood and is unable to meet the body's demand for oxygen. The heart then over-compensates by enlarging. Others include Angina Pectoris when the oxygen demand of the heart muscle exceeds the oxygen supply because of a narrowing in the coronary arteries, causing a possible heart attack; and Acute Myocardial Infarction when the heart muscle (myocardium) changes due to the sudden deprivation of circulating blood, usually caused by the narrowing the of the arteries.

The data set did not contain any economic, social or educational information on the patients. A table containing census tracts for each patient was downloaded from the ESRI site. The result was a new table including the respective census tract for each patient. This new table was then merged with a table consisting of information about the median household income. This new variable was collected from the website of the United States Census. With more variables in the dataset, a map of median household incomes according to the respected census tract was constructed. Bringing in other information with respect to the patients such as educational background was very difficult. The education information provided by the Census website was set in two categories of males and females and was divided into every category of educational status. The original dataset contained no demographic data, such as gender or age of the patients. The healthcare providers of cardiac and pulmonary problems presented an interesting variable for the project: the addresses of each of the providers affiliated with the hospital were collected. A table of all of the cardiac and pulmonary healthcare providers was created. That table was then imported in ArcMap and geocoded. A map is given in Figure 1 of the location of the providers compared to the median household income. The providers are displayed according to type, cardiac and pulmonary. Visually, it is obvious that all the providers are located within the low-income level areas (although somewhat surprisingly).

Figure 1. Socio-economic geographic areas and healthcare providers



The proximity of patient to the nearest provider was computed using the Near command from ArcToolbox in ArcGIS Software. Conducting text mining in Enterprise Miner, the patients were clustered according to similarities of the word descriptions of the DRG codes. The result was five clusters, where each cluster has similar trends in words more so than any other cluster. The clusters are shown in Table 1.

Table 1. Text Clusters

Cluster Number	Descriptive Terms	Frequency	Percentage
1	pleurisy, pneumonia, simple, simple pneumonia	411	9%
2	diagnosis, respiratory, system, respiratory infections, support, ventilator, + infection, + inflammation, + neoplasm,	371	8%
3	obstructive, chronic, chronic obstructive pulmonary disease, pulmonary embolism, embolism,	337	7%
4	with, without, other, coronary, + procedure, bypass, cardiovascular, ptca, percutaneous, percutaneous	2671	55%
5	circulatory disorders, catheterization, cardiac catheterization, failure, shock,	1033	21%

Kernel density estimation was then applied to each cluster. This allowed for a smooth approximation of the density function across each observed point. The procedure of computing the kernel density was performed in Enterprise Guide. We have n independent observations $x_1, \dots, x_n$ from the random variable X , which in this case distinguishes the two kinds of providers. Two graphs were constructed, comparing kernel densities to the nearest pulmonary provider, and likewise to the nearest cardiac provider. From these graphs, it is evident that most patients live respectively around 10,000 or 20,000 feet (approximately 2-3 miles) from the nearest cardiac and pulmonary healthcare providers' location. The code for the kernel density estimation is

```
proc kde data=work.cadg gridl=1 gridu=20
method=srot out=outkde;
var packyears;
run;
```

The kernel densities of proximity to nearest provider are shown in Figures 2 and 3. For example, there is a high density of patients from Cluster 3 that live within 10,000 feet from the nearest pulmonary provider, while there is also a high density of patients in cluster 5 that live approximately 15,000 feet from the nearest cardiac provider. Furthermore, Figure 2 has more irregular bumps than Figure 3. That is because the bandwidth for Figure 2 was set at 1 while the bandwidth for Figure 3 was set at 2. The larger the bandwidth, the smoother the approximation of the graphs. Using Text mining in this project gave the opportunity for efficient analysis treatment of the data. The data were grouped into the different clusters, where each cluster contained words of the different health conditions of the patients. This allowed for patients having similar health problems to be treated as one variable, namely one cluster, and every patient belongs to exactly one cluster. Proximity of patient to provider was computed on average according to the 5 clusters as opposed to each patient individually.

Figure 2. Comparison of Kernel Density to the Nearest Cardiac Provider

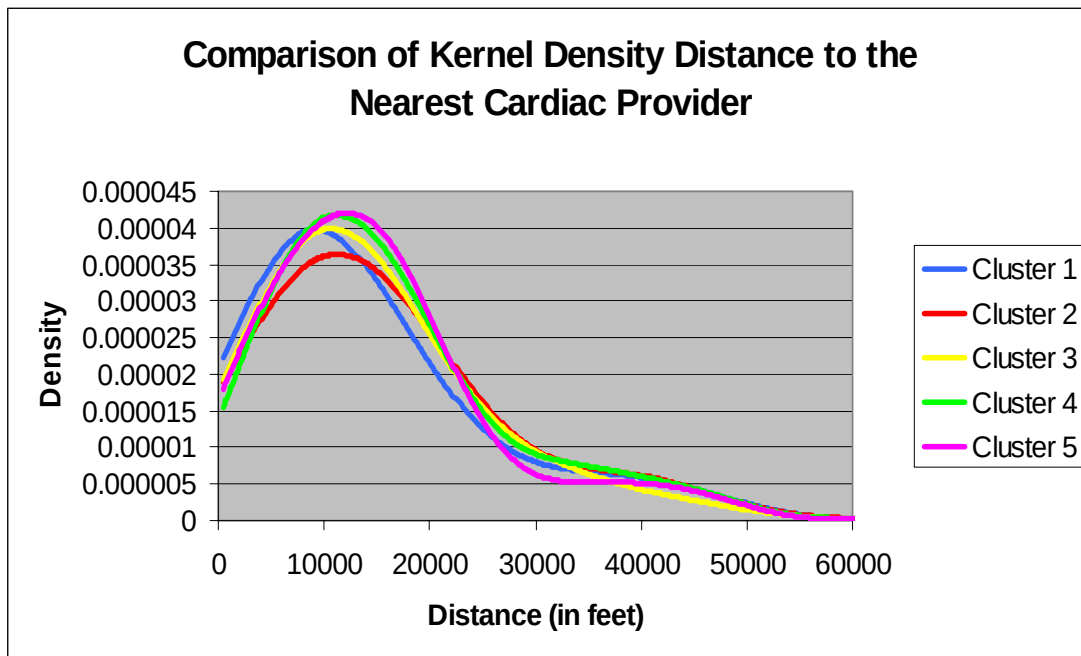
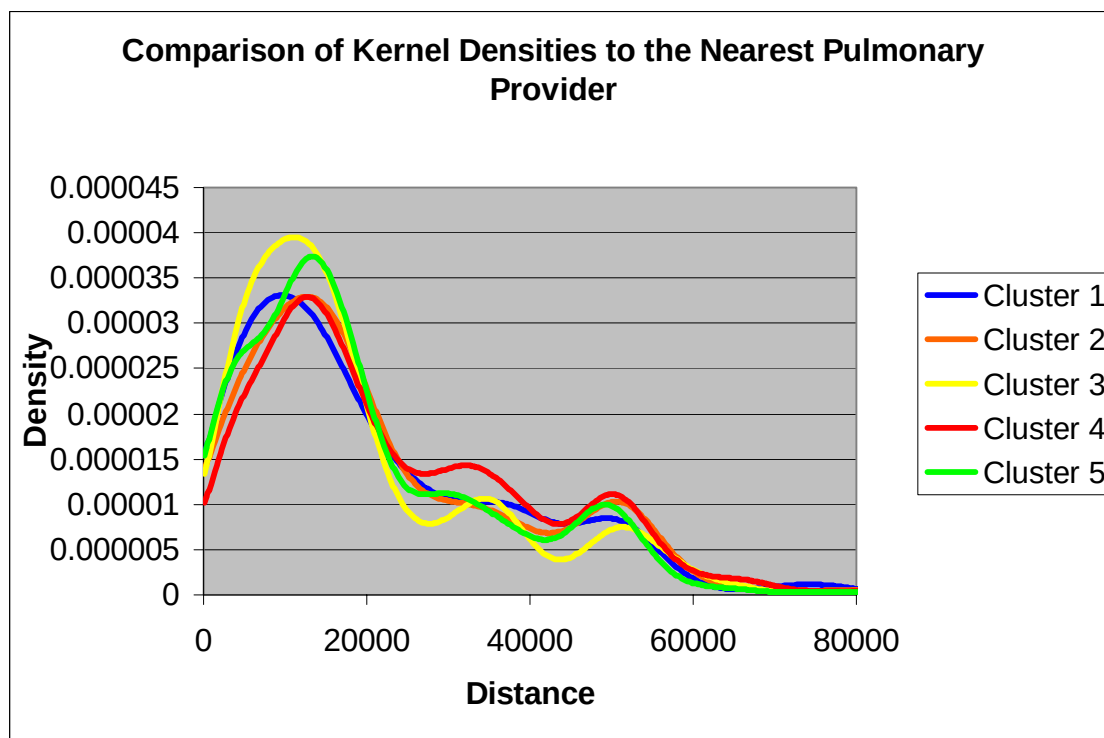


Figure 3. Comparison of Kernel Density to the Nearest Pulmonary Provider



CONCLUSION

Text mining and clustering allowed for an efficient and time-consuming statistical data analysis. Furthermore, the bridge between SAS and ESRI allowed for a statistical formulation of the data as well as a visual interpretation of the patients and health care providers locations in Jefferson County. The purpose of this study was to undertake a spatial statistical analysis perspective in order to examine the distribution of patients and health care providers and the SAS- ESRI bridge allowed for a successful formulation. The combination of statistical methods such as text-mining and kernel density estimation alongside geographic mappings of the findings allows efficient statistical interpretation.

ACKNOWLEDGEMENTS

I would like to thank Dr. Patricia Cerrito for all her guidance and support throughout all of my work and for the preparation of this paper.

CONTACT INFORMATION

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

Author Name: Christiana Petrou
Company: University of Louisville
Address: 2231 Arthur Ford Ct, Apt #1
City state ZIP: Louisville, KY, 40217
Work Phone: 502-852-6240
Email: cspetr01@louisville.edu