ST16

Data Mining of Dental Information

Christiana Petrou, University Of Louisville, Louisville, KY

ABSTRACT

The purpose of this study is to examine standards of care in the Dental School of the University of Louisville. The issue of compliance on behalf of the patients is the central theme in this project. We will examine the relationship of visit intervals, treatment needs, and patient compliance. We will test hypotheses such as whether treatments vary by clinic group, or by entry point. We will examine the relationship of visit intervals, treatment needs and patient compliance. We will test hypotheses such as whether treatment schedules exist in the clinical database that will optimize patient compliance, and whether the compliance level varies by clinic group or by entry point. Another aim is to examine the relationship of patient demographics, including socio-economic status to compare to the level of compliance and to test the hypothesis that patients living in lower socio- economic neighborhoods will have a lower rate of compliance

INTRODUCTION

The purpose of this paper is to demonstrate the use of linear models, clustering, neural networks, decision trees and kernel density estimation to enable us to examine a large, complex data set. The main focus will be the use of generalized linear mixed models in spatial analysis and to investigate information from a large dental clinic using SAS combined with ArcGIS software to make inferences on the cost of dental care. We will examine the relationship of patient visit intervals, treatment needs and patient compliance. To examine the relationship of patient demographics, Tapestry data, provided by ESRI, Inc. that provides detailed socio-economic status will be combined with the dental database. We will investigate a large number of variables simultaneously and remove the ones that are not statistically significant.

The main theme is to examine standards of care in the Dental School of the University of Louisville. Our dataset consists of over 30,000 patient visits. An issue of concern is patients' non-compliance. One of the difficulties in determining patient compliance is to define it and to be able to measure it. Compliance is basically adherence to a drug regimen as in taking medications correctly and on time. For dental patients, it also involves routine checkups. It encompasses the patient's active participation in his or her own healthcare; seeking medical advice, keeping appointments, following recommendations concerning lifestyle, as well as following medical regimens. Here are some examples of noncompliance:

1. Failure to take medications, which includes missed doses and stopping therapy too soon.

2. Taking too much medication. More is not always better.

3. Taking a drug for the wrong reason, especially if the patient takes multiple drugs and is confused about their purpose.

4. Improper timing of drug administration, especially with complex regimens.

5. Not having the prescription filled or refilled.

The first step was to construct a definition for compliance based directly from our data. For a specific treatment, patients undergoing that treatment should be identified. The time interval between visits can be computed from visit dates. Median treatment intervals were estimated from the data (median used to avoid the influence of an outlier). The patients who were within 75% confidence of the median at least 80% of the time will be defined as fully compliant, with decreasing levels of compliance as the percentage decreases (for example at 60 %, 40%, 20% and 0%). For a patient undergoing several treatments, the compliance is computed for each treatment individually. To determine compliance, patient visits are first defined sequentially. Once the visits are sequential, visits can be examined. The difference between the visit date and time t and time t+1 is used to find the length between visits. The median of these differences is computed, as well as the 80th percentile. The 80th percentile value is then compared to the desired time between visits for that patient's particular diagnosis (CPT code). Compliance with treatment requirements is essential to good health. However, many patients do not comply and the reason may be partly due to convenience, money, or bad habits, or could even occur because of a disagreement about a treatment. We wanted to examine the issue of visit intervals, treatment needs and patient compliance. Patients who have an interval in excess of the norm are defined as non-compliant.

We will examine the relationship of visit intervals, treatment needs and patient compliance and test hypotheses such as whether treatment schedules exist in the clinical database that will optimize patient compliance, whether treatments vary by clinic group and whether compliance varies by entry point. The model equation to be examined will be

$$\text{Compliance}= \beta_0+ \beta_1\text{*patient code}+ \beta_2\text{*clinic}+ \beta_3\text{*treatment code}+ \beta_4\text{*treatment scheduling}+ \beta_5\text{*patient demographics}+ \beta_6\text{*entry point}+\varepsilon$$

Multiple pair wise comparisons can be made to determine which dentists have the highest level of patient compliance. Once they are identified, then the treatment schedules of dentists with higher compliance will be compared to those with lower compliance to determine difference. Another aim is to examine the relationship of patient demographics, including socio-economic status to level of compliance to test the hypothesis that patients living in lower socio- economic neighborhoods will have a lower rate of compliance. The power of GIS can be combined with the data mining tools in SAS to examine the relationship of spatial distance to attribute data. The first law of geography states that "everything is related to everything else, but near things are more related than distant things". This law merely states that nearby neighborhoods tend to act similarly and this similarity weakens as the neighborhoods become more distant from each other. Patients living in lower socio-economic neighborhoods will have a lower rate of compliance. Education and income interactions will be used to define the socio-economic index and will be used to investigate the levels of compliance. It is anticipated that older patients will have a higher level of compliance; this is because older people have more need of medical services, and they have more time to wait at the clinic Also, patients living a greater distance from the dental school are expected to be less compliant. In this case, GIS can be used to investigate spatial statistics involved with a geographic measure of proximity. Kernel density estimation will be used as a data visualization technique to investigate the relationship between proximity and compliance. In particular, it will be determined if a cut point exists in terms of proximity so that patients living at greater distances than the cut point will be less complaint; patients living at a lesser distance will be more compliant.

From the original dataset of approximately 142,000 patient records, three tables were obtained. The data were divided into patients that lived in Jefferson County, which accounted for about 88,000 records, patients who lived in the rest of the Kentucky counties, and finally patients who lived in Indiana. The patients belonging to the Jefferson county area were considered first; eventually the methodology used for the Jefferson county area will be projected to the rest of the data. The first objective was to find the distance from each patient in Jefferson County to the Dental School in Louisville. ArcGIS was utilized to achieve this goal. The patients' addresses were geocoded using the ArcGIS version 9. The results of the geocoding procedure are displayed in Figure 2. Jefferson County is one of 120 counties in Kentucky. The county includes the Louisville metro area and it has a total area of 399 mi². The Ohio River forms its northern boundary with the state of Indiana. According to Mapstats in 2004 the population of Jefferson County was a bit over 700,00.

Figure 1. Patient locations in Jefferson County

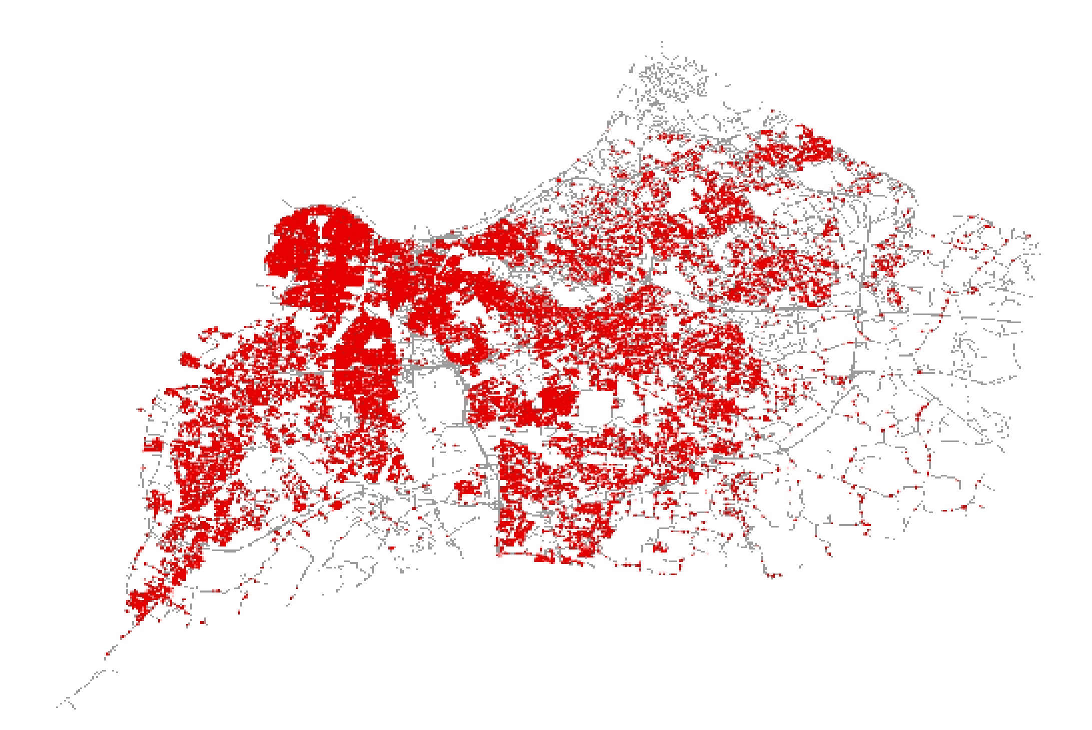
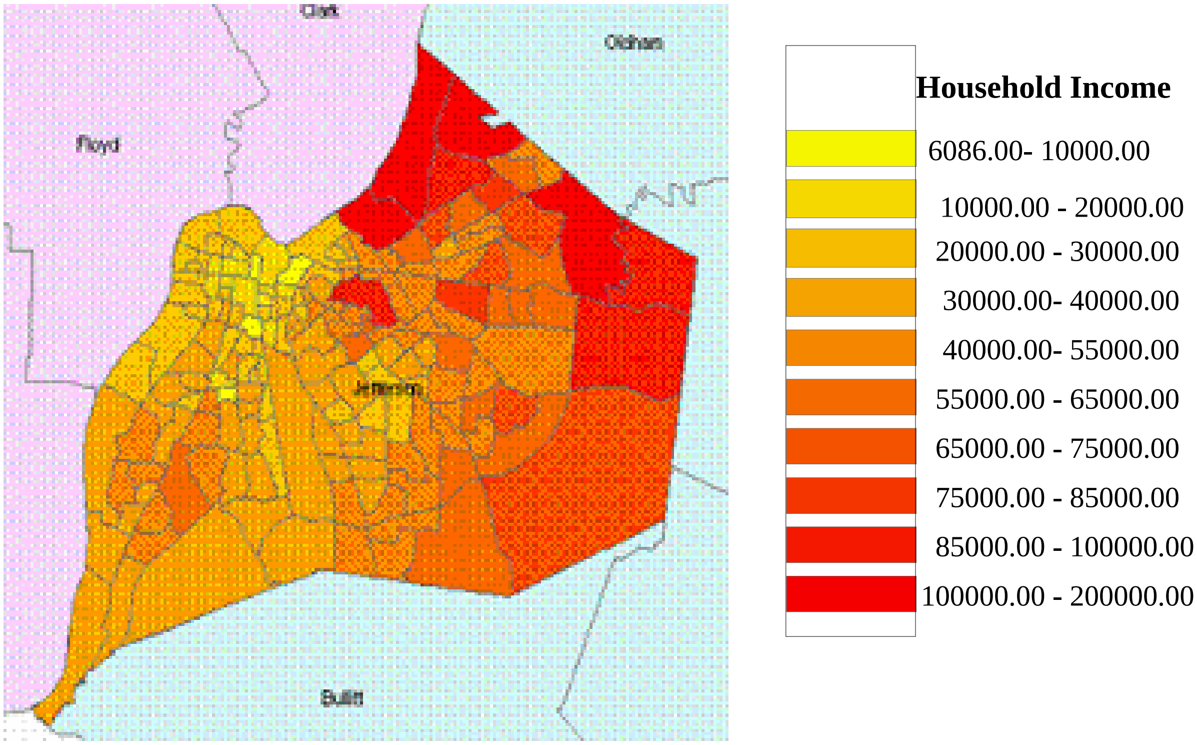


Figure 2. Socioeconomic map of Jefferson County



Viewing the socioeconomic map of Jefferson County, we conclude that a cluster of patients are situated in the low income areas of Jefferson County. Furthermore, the address of the Dental school was obtained and located on the map. Arctoolbox includes the Proximity Near command that was used to compute the length of the shortest distance between the locations of each patient to the Dental School. The table consisting of the distances was then imported into SAS 9.1 to run the kernel density code. Density estimation is the construction of an estimate of the density function from the observed data. A popular data visualization tool is a histogram. Histograms, however, have some disadvantages. They depend on the width of the bins; that is, equal sub-intervals in which the whole data interval is divided, and the end points of the bins, which is where each of the bins start. The problems with histograms are that they are not smooth, and depend on the width of the bins and the end points of the bins. We can eliminate these problems by using kernel density estimators. Kernel estimators smooth out the contribution of each observed data point over a local neighborhood of that data point. The contribution of data point x_i to the estimate at some point x_j depends on how far apart the two points are. The extent of this contribution is dependent upon the shape of the kernel function adopted and the width (bandwidth) accorded to it. If we denote the kernel function as K and its bandwidth by h, the estimated density at any point x is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$

where $\int K(t)\,dt = 1$ to ensure that the estimates f(x) integrates to 1 and where the kernel function K is usually chosen to be a smooth, unimodal function with a peak at 0. This procedure uses the following code:

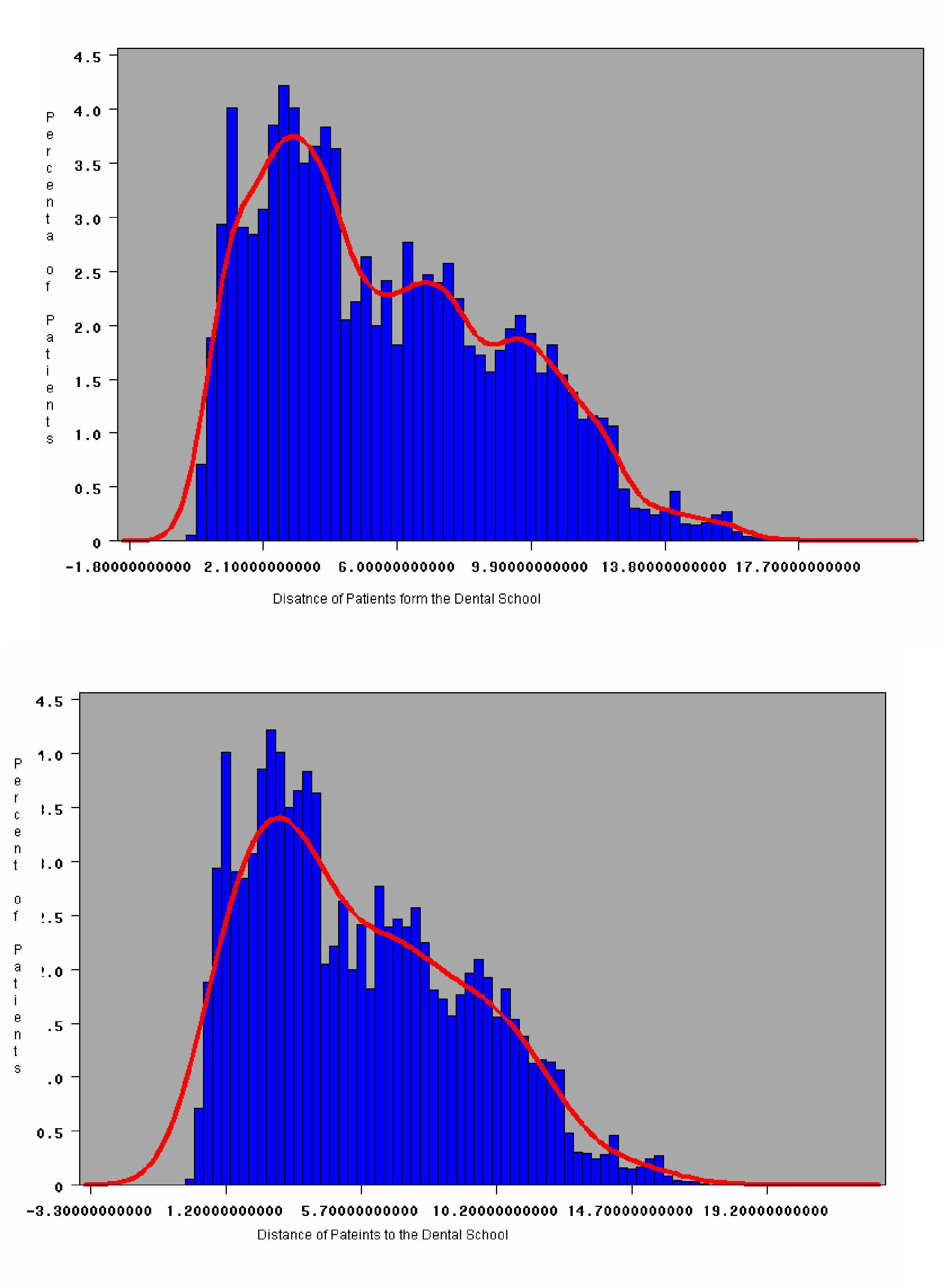
```
    PROC KDE DATA=SASUSER.DISATNCE GRIDL=1 GRIDU=87965 METHOD=SROT
    BWM=1.00 OUT=SASUSER.OUTKDE1;
    VAR DISTANCE;
    RUN;
```

To get the histogram together with the kernel density plot, the following code was used:

```
    PROC UNIVARIATE DATA=SASUSER.DISTANCE
    NOPRINT;
    VAR DISTANCE;
    HISTOGRAM / Kernel (W= 4 L= 1 COLOR=RED C=1 K=Normal)
    CFRAME=CXA8A8A8 CAXES=BLACK WAXIS=1
    CBARLINE=BLACK
    CFILL=BLUE
    PFILL=SOLID;
    RUN;
```

Figure 3 shows that approximately 3.8% of the patients live within a distance of 3.5 miles from the Dental school. Then there is also another peak with approximately 2.5% of the patients living within a distance of 7 miles from the Dental school. In general, we observe that about 85% of the patients are within a distance of 13 miles from the Dental school. Hence, we conclude that the downtown Dental school attracts patients that are situated nearby and furthermore, most of the patients are situated in the lower income areas of Jefferson County.

Figure 3. Kernel Density Estimation



The next task was to examine the relationship between distance from the dental school for each patient and the time between visits to the school. The time between visits was computed by determining the time difference between a patient's first visit and their last visit. If a patient only visited the Dental school once, then their time between visits was set to 0. The procedure was executed in SAS with the following code:

```
DATA DIFFERENCE;
SET SASUSER.CHARGES;
BY PAT;
IF FIRST.PAT AND LAST.PAT THEN DELETE;
RETAIN R_CHTYY R_CHTMM R_CHTDD;
IF FIRST.PAT THEN DO;
R_CHTYY=CHTYY;
R_CHTMM=CHTMM;
R_CHTDD=CHTDD;
END;
IF LAST.PAT THEN DO;
DIFF_CHTYY = CHTYY - R_CHTYY;
DIFF_CHTMM = CHTMM - R_CHTMM;
DIFF_CHTDD = CHTDD - R_CHTDD;
OUTPUT;
END;
DROP R_:;
RUN;
```

In the outcome however there are negative values for the months and the days. To avoid this issue, the following code was utilized to convert years to months and add any left over months from the months column.

```
DATA WORK.DIFFERENCE;
MONTHS= DIFF_CHTYY*12+DIFF_CHTMM;
RUN;
```

The table that consisted of the time between visits at the clinic was merged together with the table that consisted of the distance of each patient to the dental school. This was done in Enterprise Guide 3. Finally we used a Generalized Linear Model to see the dependence of proximity of a patient to the length of attendance at the school. The general equation was Months =1.97138 – 0.00438Distance. Therefore, if a patient lives right next to the Dental school then the average amount of between visits to the school is approximately two months. As the patient lives further and further away from the school, then the amount of time between visits decreases. It is noteworthy that the significance level of the model was 0.927.

CONCLUSION

The next step in this project will be to define compliance. This will be attempted by observing the patients visiting dates and times. There are a number of investigations that should be carried out as the study progresses. At this point, we have a visual perspective of the social status of the patients, at least for the ones in Jefferson County. A cluster of patients live in the low income areas of Jefferson County. Based on the definition of compliance, we will determine whether patients in lower incoming areas are more or less compliant than patients who live in higher incoming areas. Furthermore the data set consists of the CPT codes for the patients. CPT stands for Current Procedure Terminology and these codes are part of a code list selected by HIPAA, used to describe health care services in electronic transactions. Using the CPT codes will enable us to cluster patients in groups using text miner by examining associations of words that are within the CPT codes. This procedure will allow for a reduction of the thousands of patient records to a small finite number of clusters allowing for a simpler manipulation of the data set.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Patricia Cerrito for all her help in this paper.

CONTACT INFORMATION

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

Christiana Petrou
2239 Arthur Ford Ct, Apt 1
Louisville, KY, 40217
Work Phone: (502)852-6240
Email: cspetr01@louisville.edu