Data Mining and Visualization: Real Time Predictions and Pattern Discovery in Hospital Emergency Rooms and Immigration Data

by

Ashley M. Snyder

B.S. Operations Research
United States Air Force Academy, 2008

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN OPERATIONS RESEARCH
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2010

Signature of Author: _____
Sloan School of Management
Interdepartmental Program in Operations Research
May 14th, 2010

Approved by: _____
Dr. Natasha Markuzon
The Charles Stark Draper Laboratory, Inc.
Technical Supervisor

Certified by: _____
Professor Roy Welsch
Professor of Statistics and Management Science and Engineering Systems
Thesis Advisor

Accepted by: _____
Professor Dimitris Bertsimas
Boeing Professor of Operations Research
Co-Director, Operations Research Center

[This Page Intentionally Left Blank]

Data Mining and Visualization: Real Time Predictions and Pattern Discovery in Hospital Emergency Rooms and Immigration Data

by

Ashley M. Snyder

Submitted to the Sloan School of Management on

May 20th, 2010 in partial fulfillment of the requirements for the

Degree of Master of Science in Operations Research

Abstract

Data mining is a versatile and expanding field of study. We show the applications and uses of a variety of techniques in two very different realms: Emergency department (ED) length of stay prediction and visual analytics. For the ED, we investigate three data mining techniques to predict a patient's length of stay based solely on the information available at the patient's arrival. We achieve good predictive power using Decision Tree Analysis. Our results show that by using main characteristics about the patient, such as chief complaint, age, time of day of the arrival, and the condition of the ED, we can predict overall patient length of stay to specific hourly ranges with an accuracy of 80%.

For visual analytics, we demonstrate how to mathematically determine the optimal number of clusters for a geospatial dataset containing both numeric and categorical data and then how to compare each cluster to the entire dataset as well as consider pairwise differences. We then incorporate our analytical methodology in visual display. Our results show that we can quickly and effectively measure differences between clusters and we can accurately find the optimal number of clusters in non-noisy datasets.

Technical Supervisor: Natasha Markuzon
The Charles Stark Draper Laboratory, Inc.

Thesis Advisor: Professor Roy Welsch
Professor of Statistics and Management Science and Engineering Systems

[This Page Intentionally Left Blank]

Acknowledgements

There have been many people who have provided me with insight, help and encouragement throughout the writing of this thesis. I would like to thank the many individuals for their support.

First, I would like to thank Dr. Natasha Markuzon of Draper Laboratory. You spent many hours looking over my work, helping me to come up with ideas, and always pointing me in the right direction. It has been a pleasure working with someone as knowledgeable and professional as you.

I would also like to thank Dr. Roy Welsch of MIT. You taught me the fundamentals of Data Mining and it was in your class that I was first able to apply specific techniques to actual data sets and become aware of the myriad of Data Mining applications.

Furthermore, I would like to thank Dr. Larry Nathanson and Dr. Leon Sanchez from Beth Israel Deaconess Medical Center for some very interesting talks regarding the operations and processes of the Emergency Department as well as providing me with the ED data to work with.

Thank you to Kimberley and Tommy for helping me with the problem sets, trying to answer the copious amounts of MatLab questions I had, but above all for working with me to make the best door of the 2009 Holiday Door Decorating Contest, which remained up for the majority of 2010.

To all my friends and Beantown and MIT Rugby: I would have never made it without you. Thanks for the memories, a place to live, and all the opportunities you gave me from a rugby perspective and in life. A special thanks to the 294, the porch of 9 Derby, Sha, Mel, Emilie, Dowty, Meredith, Meat, Richard, Kitty, KateO, Seary, Brandon, Freddie, and Mark Green. I am going to miss you guys.

Finally, I'd like to thank my family for their invariable support: to my Dad for listening to my stories about the "working world," my Mom for always seeing the silver lining in any situation, and to Wes for always keeping it light, and finally starting to answer his phone. Most of all, thank you *all* for answering my calls that I would of course only make while driving home.

This thesis was prepared at The Charles Stark Draper Laboratory, Inc., under Internal Company Research Project 22928-001 and 23921-007, Data Mining for Draper.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or The U.S. Government.

Ashley M. Snyder, 2nd Lt., USAF May 14, 2010

[This Page Intentionally Left Blank]

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The goal of this research is to explore the application of data mining methods in two different domains in which a lot of sparse noisy data exists. We explore data mining as a prediction tool for the patient lengths of stay at the Beth Israel Deaconess Medical Center (BIDMC) Emergency Department (ED) as well as provide an analytical data mining background for the development of a human guided interface to allow a user to see in near to real time a mathematically optimal number of clusters in an immigration location dataset as well as the similarities and differences between these clusters. Below is an outline of the remainder of the thesis.

Chapter 2 – Techniques for Data Mining Analysis and Prediction

In Chapter 2, we provide an overview of the supervised and unsupervised learning techniques used. The methods, applied to the ED dataset and the immigration location dataset, will be discussed in detail and include k means clustering and regression to uncover the natural

groupings of the data, as well as Logistic Regression, k Nearest Neighbors, and Decision Trees for prediction.

Chapter 3 – BIDMC Patient Length of Stay Predictions

Chapter 3 provides an introduction to the work of optimization in hospitals and Emergency Departments around the county. We show how the raw data was modified and how different types of data were combined to explore both information about a patient's visit as well as statistical information about the condition of the ED when that patient arrived. We then describe how the data is incorporated into supervised learning methods, conduct a full analysis of results, and make conclusions for predicting patient length of stays for the ED's most frequent chief complaints.

Chapter 4 – VAST Dataset: Real Time Clustering and Information Processing

In chapter 4, we focus on the geospatial evolution of migrant boat landings. We discuss application to a variety of optimal clustering methods to k means clustering in order to determine the mathematically optimal number of clusters for the landing locations. We then use this information to drive the background analysis behind the visual analytics user interface to address and determine differences between each of the clusters and the entire dataset as well as pairwise differences between clusters. We also discuss the incorporation of this analysis into a visual analytics tool.

Chapter 5 – Contributions and Future Work

We finally discuss the overall contributions of our current models and results for ED length of stay prediction and automatically uncovering a mathematically optimal number of clusters as well as their differences for the immigration location dataset. We propose ideas for future work including the use of Artificial Neural Networks in the ED setting and increased specificity of patient chief complaints as well as the implementation of multiple clustering algorithms to provide complementary analysis tools.

Chapter 2

Techniques for Data Mining Analysis and Prediction

This chapter introduces both the unsupervised and supervised learning techniques that we use in this thesis: *k* means clustering, Logistic Regression, *k* Nearest Neighbors, and Decision Trees. We concentrate on these methods as they satisfy the main criteria necessary for our research including:

- Develop models that are data-driven and based on stored historical information
- Perform classifications for multiple labeled classes
- Operate close to real-time

2.1 Science of Learning

Data mining is generally defined as the process of taking large amounts of data and analyzing it to find patterns and relationships resulting in useful information. Closely related to data mining is machine learning. Machine learning is a discipline that allows for the development of algorithms in which a computer is able to change its behavior, or learn, based on a specified set of data, which is called the training set. Through a variety of algorithms, the computer is able to model and recognize complex patterns and apply this knowledge to unseen sets of data.

With advances in technology and ever expanding ways of collecting records, data mining has evolved as a very useful way of understanding large quantities of data. It allows users to look beyond basic statistical information such as averages or standard deviations, and instead into deeper aspects such as the relationships between variables and subtle patterns within the data. Data mining contains three main parts: training, testing, and validation. A training set is used to model the data and must be a good representation of the entire dataset to be effective. Testing and validation apply the model to unseen sets of data and evaluate the model's performance to dictate how well the model fits the data.

Both machine learning and data mining are used to help a computer model a variety of real world situations and apply those models to assist a user with information extraction. Machine learning has been applied to many fields including machine perception, language processing, pattern recognition, medical diagnosis, detecting fraud, stock market analysis, classifying DNA sequences, and cell mapping [1]. Machine learning and data mining include four main classes of tasks [2]:

- Classification
- Clustering
- Regression
- Association

These main classes can be broken down into specific categories of algorithms, the most common are unsupervised and supervised learning. In unsupervised learning we look at how

data is organized when there are no class labels on the target variable. It is used largely through clustering algorithms that attempt to characterize inputs and determine associations among the variables and the natural separations in the data [3].

Supervised learning learns from a training set of data, which associates input variables and a target output. There is a labeled output which is what separates it from unsupervised learning. This learned function can then be applied to a new set of inputs with a purpose of predicting the output label. A regression model predicts for a continuous output, while classification predicts a specific output label that can be binary, numerical, or qualitative. We used unsupervised learning for both the ED and immigration location datasets, and supervised learning for the ED dataset because the data were labeled with a patient's overall length of stay.

Exploring first the ED dataset, we focus on clustering, regression and classification. Our goal is to use clustering and regression to find the main variables of interest and then use classification to predict a patient length of stay. For the immigration location dataset that had no labels attached, we perform clustering, focusing on automatically finding the optimal number of clusters and computing near real time comparisons of cluster differences.

2.2 Unsupervised Learning

The goal of unsupervised learning is to understand the underlying structure of a dataset. Because unsupervised learning does not have a class label, it is more difficult to determine a strict measure of success as compared to supervised learning, and the effectiveness of a method cannot be directly verified. Cluster analysis has a variety of goals that aim to group elements of a dataset together based on their attributes, specifically, grouping within one cluster items that are more similar than those of another cluster. One of the most popular unsupervised learning techniques is the iterative process of *k* means clustering [3].

2.2.1 K Means Clustering

K means clustering is one of the most popular iterative descent clustering methods [4]. It is a partitioning clustering technique that is based on the notion of a center point, or centroid, that represents the mean of the points around it. The main dissimilarity measure for this algorithm is

Euclidean distance, but other distance/dissimilarity measures are valid and include the Manhattan distance metric. These metrics are defined as:

$$\textbf{Euclidean Distance: } D(x_i, c_0) = ||x_i - c_0|| \qquad (2.1)$$

$$\textbf{Manhattan Distance: } D(x_i, c_{ij}) = \sum_{j=1}^{p} |x_i^j - c_0^j| \qquad (2.2)$$

where p is the number of attributes, x_i is the individual data point, x_i^j and c_0 are the data points/clusters of a specific attribute that we are interested in.

Because a numerical distance metric is used to compare variable distances within or between clusters, the variables must be quantitative. If categorical or qualitative variables are used, they are transformed into a numerical representation either by taking on specific integer values or taking on values over a specific range. In order to preserve the weighting of the attributes, each variable must be normalized to have a mean of zero and a variance of one. The k-means clustering algorithm can be defined in the following steps:

1. Choose "k" the number of clusters
2. Randomly or manually chose centers (centroids) for each of the k clusters
3. Find the elements of the data set that are closest to each of the centroids using the predetermined distance measure
4. For a given cluster assignment, C, recompute the centroid by calculating the mean of all elements in C and make this the new centroid
5. Repeat steps 2 and 3 until the centroids do not change

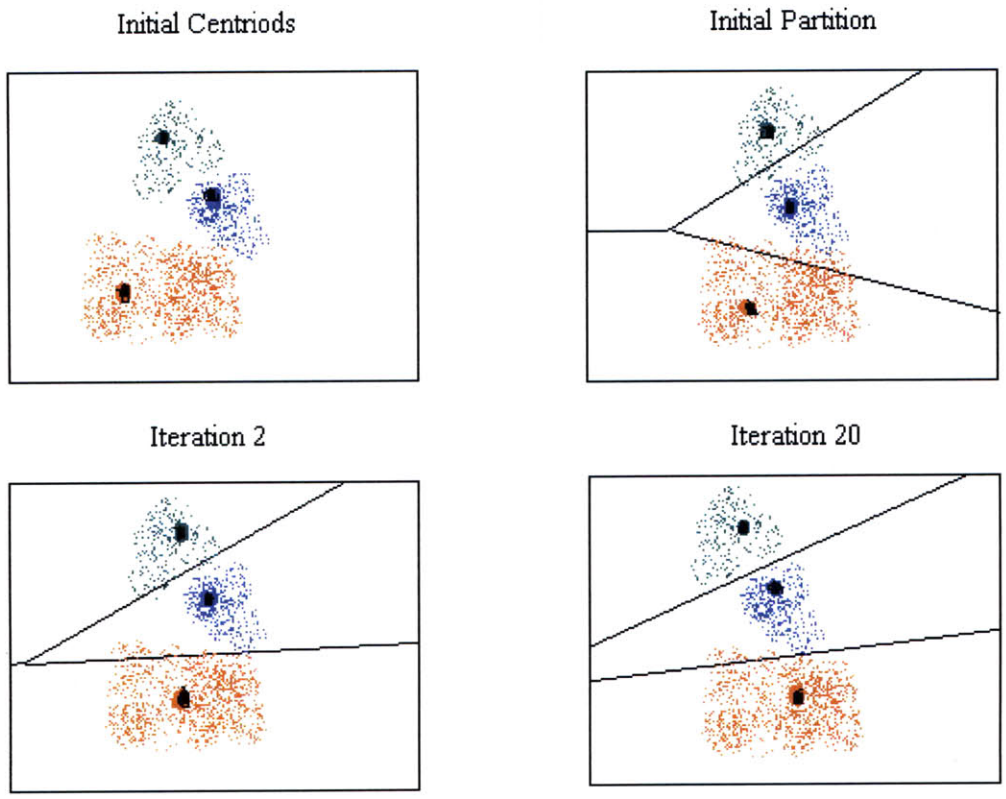


Figure 2.1 Iterations of K Means clustering

When the data are not clearly divided into visible clusters, the starting locations of the initial centroids become very crucial. To account for this, one must either manually choose centroid centers based on the density of the data or try many different random starting points, settling on the centroids that result with the smallest sum of squared distance. This sum of within cluster squared distance is sometimes called the sum of squared error (SSE) and is defined as:

$$\boldsymbol{Error = \sum_{i=1}^{k} \sum_{x \epsilon C_i} \sum_{j=1}^{p} (x_i^j - c_0^j)^2} \qquad \textbf{(2.3)}$$

where p is the number of attributes in the dataset x_j is the *jth* attribute for a data point in a cluster, c_{ij} is an element of cluster C_i referring to the *jth* attribute, k is the predefined number of clusters. This measure defines how close elements of the same cluster are to a given centroid and allows for reassignment if the SSE for a particular data point is smaller for a different cluster.

2.3 Supervised Learning

Supervised learning is different from unsupervised learning in that it deals with labeled output variables. With supervised learning, a variety of inputs are assessed by an algorithm and an output is produced. The input vector consists of all the attributes or "x" values that are available and can be made of up categorical or numerical values. The data is separated into two sets. The first is a training set, which is used to learn and model the data by both its inputs and labeled outputs through a selected algorithm and compares its predicted classification to the actual target label. If the computer is unable to correctly classify the data, the error will be much higher and the ability of the model to predict on the unseen data will likely be poor. If the model is able to accurately classify the elements, the error will be low, the accuracy high, and we can apply the model to the test set. The test set consists of a separate unseen dataset and is used to validate the performance of the training set.

One of the measures of performance in classification is Loss. Loss is a measure of how different the prediction results are from the actual class labels in the training set. It is often referred to as misclassification error or in the case of a training set, the training error. Training error is a very loose estimate of the test error we can expect when we run our model against the unseen data in noncomplex settings, but the test error is usually higher than training error overall and increases greatly as model complexity increases. Figure 2.2 shows this relationship and represents the test error as a red line and the training error as a blue line.

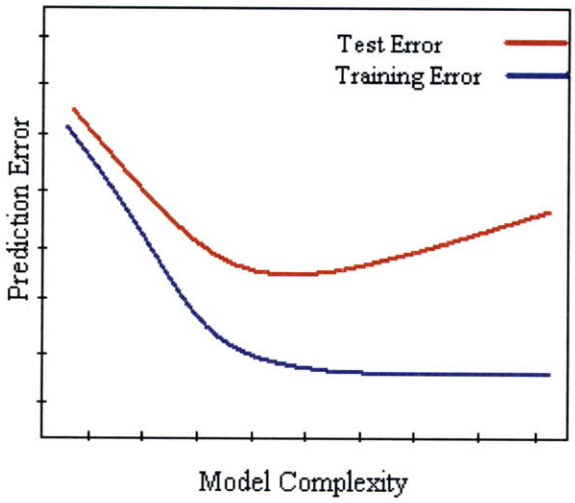


Figure 2.2 Prediction Error vs Model Complexity for Training and Test Error

If we try to minimize training error too much, we begin to see a process called overfitting. Overfitting is when the model matches the training set too closely in order to maximize accuracy, but in reality tends to exaggerate the importance of minor fluctuations and inevitably will have poor predicting power on the test set. This is why choosing a training set that is both large and representative of the entire dataset is very important. Below, a green line represents an overfitted line of the blue and red data points and the black line represents a less accurate, but better fitted line.

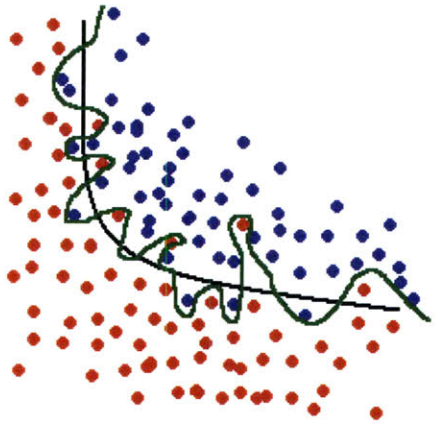


Figure 2.3 Visual Example of Overfitting

2.3.1 Logistic Regression

Logistic Regression is used for predicting the likelihood or probability of an occurrence by fitting the data to a logistic curve. It is sometimes referred to as the Logistic Model or the Logit Model and can be used with numerical or categorical data [5]. Logistic Regression can be used as a classification tool, but also as a tool in data analysis and inference in which the goal is to discover the importance of each input variable and its explanatory power, usually in terms of an odds ratio.

The logistic curve is a function that can range from negative infinity to positive infinity on the x axis, but is constrained between zero and one on the y axis. It is defined by the following function:

$$f(z) = \frac{e^z}{e^z+1} \qquad (2.4)$$

The input variable z is usually defined by the regression coefficients (β) of each of the attributes $(x_1, x_2, \ldots, x_k)$ and the values of the input variables:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \qquad (2.5)$$

The result, *f(z)*, is a probability ranging between zero and one for a particular outcome. To use this for classification, first the regression coefficients are found then a probability for each input vector is calculated for a binary prediction. A threshold probability is determined that minimizes the misclassification error for the training set data points which are classified based on this threshold. Using the regression coefficients and the probabilistic threshold previously determined by the training set, the test set is classified to evaluate how well the model generalizes to unseen data. An example can be seen in Figure 2.4.

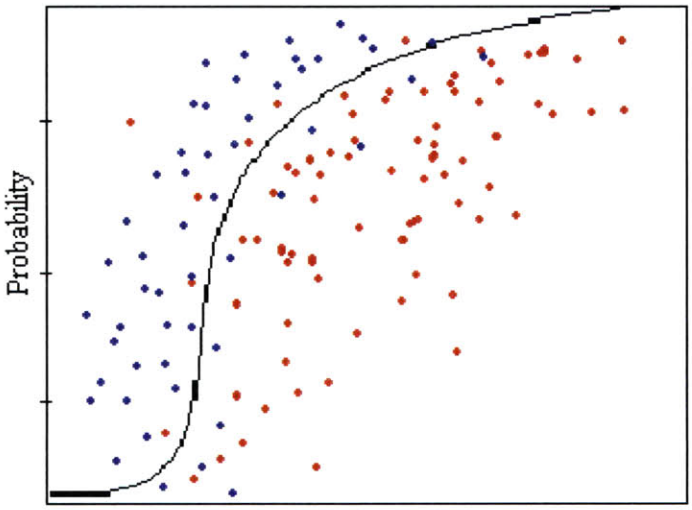


Figure 2.4 Visual Example of Logistic Regression

Logistic Regression is a common and simple way to perform a binary prediction because the input variables do not need to be normalized, have equal variance, or be normally distributed and Logistic Regression is able to handle non linear relationships and effects [6]. Two main disadvantages of this method are that it needs a large enough dataset to produce stable and meaningful results and having more than two target variables greatly increases the complexity of the model.

2.3.2 K Nearest Neighbors

K Nearest Neighbors algorithm (kNN) falls into the category of lazy classifiers because it is based completely on the training set and requires no fitting of a model [7]. Lazy classifiers store all of the training points and do not build a classification model until a new point needs to be classified, rather than building a general model and applying to the entire set. For a specific vector from a test set, we find the k closest training data points (neighbors) and classify based on a majority vote with ties being arbitrarily broken. The input parameter, k is specified by the user and is typically an integer larger than 1. The distance metric can be chosen by the user and is generally Euclidean or Manhattan distance. Usually each attribute is normalized to have a mean of zero and a variance of one to account for the possibility of vastly different ranges of quantitative values or units of measurement [8]. Figure 2.5 shows the main idea of the kNN algorithm. The red "x" represents one class and the black "+" represents another class. The green minus sign is the point of interest that we are classifying and in this case, we are looking at the three closest neighbors (k = 3). Out of the three closest points, the red x is closest two times and the black plus only once, so we would chose to classify the point of interest as a red x. If however, we chose k to equal 7, we would then classify the point of interest as a black plus because there are four black plus signs and only three red x's within 7 nearest neighbors.

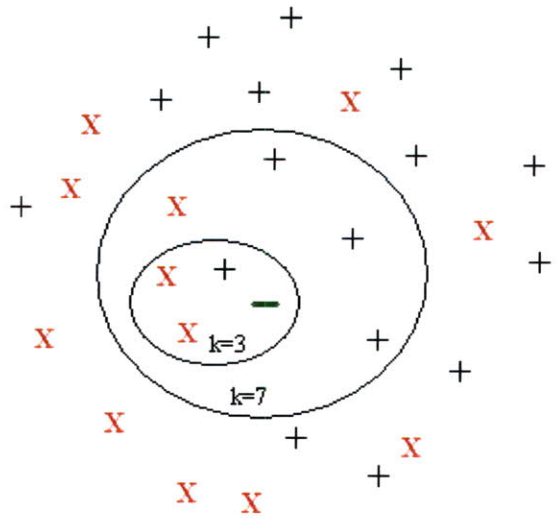


Figure 2.5 Visual Example of K Nearest Neighbors

An important disadvantage of the kNN algorithm is in the pre-selection of the parameter k by the user, as it is very important in the overall classification of a data point. Another problem is in the dominance of a single class. If one of the classes is more prevalent compared to the other classes, the dominating class will tend to be the majority of the neighbors simply because it is more frequent.

Although a simple algorithm, kNN is still very effective. We can alleviate some of the disadvantages described above by performing cross validation and balancing the dataset, which helps to keep one class from dominating. kNN has had much success in many areas of research including handwritten digits, satellite imaging, and EKG patterns and has the most success in areas where each class has many possible prototypes and the decision boundary is highly irregular [8].

2.3.3 Decision Trees

In Decision Tree Classification, there exists a tree with attribute nodes, decision criteria, and an overall classification. The attribute nodes correspond to an input variable with the branches coming off representing each of the possible values for that variable or a specific decision criterion. The decision criteria can be qualitative (Yes/No) or mathematical and be determined by a probabilistic threshold ($p<0.77/p>0.77$) [3]. The final node, or "leaf," represents a classification resulting from decisions at each of the interior nodes. The algorithm starts with the entire data set and then a variable is chosen at each step to be used to partition the data. The goal for determining which variable should be split first is to start with the "best" variable, followed by the next best, and so on. What constitutes "best" depends on the formulae used for measuring the variables, but the main idea is to make splits resulting in the lowest misclassification error [8]. Figure 2.6 is an example of a simple decision tree used to determine a binary classification of Yes or No. Each box represents an attribute or input, four total, each branch contains a means of making the decision, and each circle is the final classification.

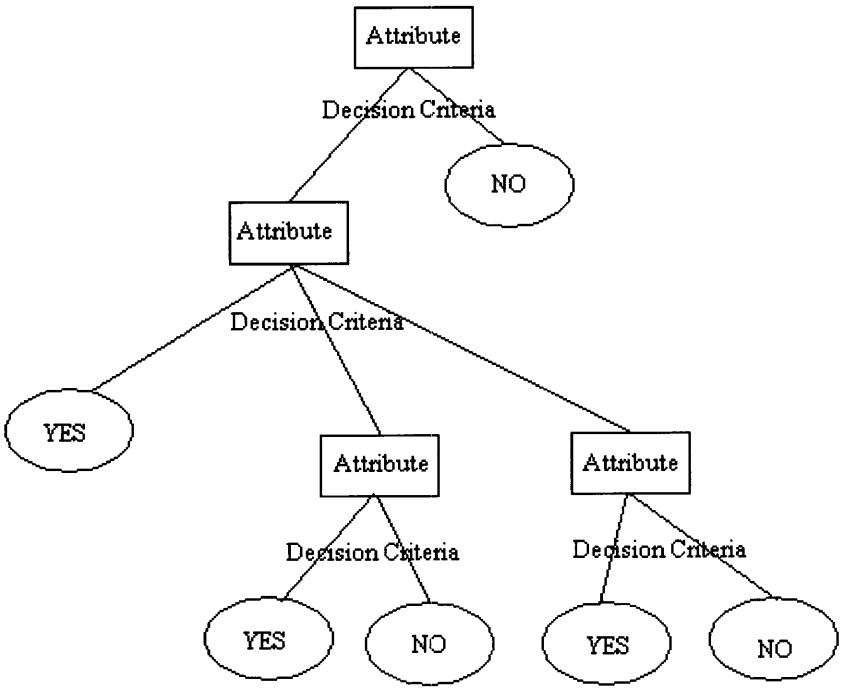


Figure 2.6 Visual Example of Decision Tree Classification

Decision Trees come with many advantages because they are simple to understand and easy to quickly visualize. They also do not require data normalization, can work with blank values, and are functional with both categorical and numerical data entries. Because of its computational simplicity, it is able to handle large amounts of data quickly and it is simple for a user to alter nodes or decision criteria without much underlying computational knowledge. Because the Decision Trees are making locally optimal decisions at each node, they cannot guarantee a globally optimal final answer, which is their main disadvantage [9]. Furthermore, Decision Trees can easy become very complex and "wide" because of overfitting. To overcome this problem, some scaling or pruning is required.

2.4 Multiclass Extensions

Logistic Regression does not easily translate to multiple classes because the algorithm is looking for a logarithmic curve to separate the classes of values. Instead we use Multinomial Logit modeling which separates data with planes and hyperplanes. This method differs from

Logistic Regression because it works in higher dimensions and has limitations when variables are highly correlated as the model cannot easily differentiate the individual impact of these variables on the target class [10].

kNN is designed to incorporate many target variables and is not limited to the binary case. Because it takes an unknown data point and compares it to the k nearest points, we can increase the number of neighbors we are comparing to our unknown data point. This allows for smoother boundary lines made from the classification of the multiple target variables and minimizes the effects of overfitting.

Decision Trees are also designed to deal with multiple target classes with no necessary changes because the tree is built on binary recursive partitioning. The partitioning does not change when there is more than one class and simply continues splitting until there is a homogenous set of labels in which the number of label can be greater than two [11].

2.5 Summary

This chapter discussed several widely use algorithms of data mining. We discussed methods that will be used for both labeled and unlabeled datasets. For unsupervised learning, there was k means clustering, and for supervised learning, we discussed three classification methods: Logistic Regression, k Nearest Neighbors, and Decision Trees. The next two chapters focus on each dataset separately, demonstrating the application of these methods and results gained from each.

Chapter 3

BIDMC Patient Length of Stay Predictions

Whether it is a possible cardiac arrest, a woman unexpectedly going into labor, or a simple prescription refill afterhours, the ED tends to be a very crowded place with patients having a variety of medical needs. In this chapter, we present the problem statement, results of recent changes in the ED dynamics, and our methodology for prediction of patient length of stay, results, and conclusions. We show a data mining approach for real time prediction to forecast a patient's length of stay in the ED within minutes of their arrival based on their age, the capacity of the ED, the patient's initial complaint, and the time of day in which the patient arrives. These results will help to predict ED overcrowding and can serve as the basis for patient planning, with an overall goal of minimizing the time a patient is taking up a bed in the ED while other patients are waiting.

3.1 Background

Emergency Departments evolved during the French Revolution by Dominique Jean Larrey, who is often known as the father of emergency medicine, but it wasn't until after the Second World War that ED and Emergency Medicine made an appearance in the United States [12]. This centralized type of practiced medicine began to replace house visits by doctors and allowed for a single doctor to treat many patients at once in a centralized location. Prior to the 1960s, hospital EDs were staffed mostly by doctors currently working in the hospital including physicians, interns, and surgeons that would simply rotate to the ED as part of their work schedule. Eventually some physicians chose to dedicate all their time to the ED and in 1970 the first Emergency Medicine training program was developed [12].

The goal of any ED is to treat a wide range of patient complaints that require immediate attention and to stabilize acute patients until they can get more specialized and permanent care in the hospital. A US government report stated that there were 199 million ED visits in 2006, which was a 36% increase from a decade earlier. During this time span, the actual number of emergency rooms decreased from 4,019 to 3,833 and the rate of visits per 100 people increased from 34.2 to over 40 [13]. This shocking statistic has made the need for a high efficient, highly streamlined system of operations necessary for any ED.

3.1.1 Technology

The ED at Beth Israel Deaconess Medical Center (BIDMC) has been able to monitor their operations by tracking their patients' and ED statistics through a system called Dashboard [14]. The Dashboard system records information about a patient including the rooms/beds the patient occupies, the doctors that see the patient, the tests requested, the time a bed in the hospital is requested, and the time they are either discharged home or to a room in the hospital. It also takes multiple readings per hour of the condition of the ED including how many total patients are being treated, how many are in the waiting room, the number of doctors and nurses on staff, and number of patients in each of the five acuity levels. The data is accurate and thorough, and has led to a wide variety of investigation of the BIDMC operations including a report by Bentley College [15] and a simulation by Clay Noyes [16].

3.1.2 Literature Reviews

Prediction of LOS in the medical realm is by no means limited to Emergency Departments. In fact, the majority of LOS prediction is performed to predict how long a patient will remain in the hospital after having specific types of operations such as knee replacement [**17**] or cardiac procedures like a bypass surgery [18]. Much research has also been done to predict a patient's LOS in the main hospital [19] or in specific departments such as the Intensive Care Unit (ICU) [20].

Artificial Neural Networks (ANN) were used to predict patient length of stay in the main hospital using similar variables to those used in this paper including patient age, condition of the hospital, and chief complaint [19]. Results showed the ANN could predict the LOS in a hospital within one day of their actual LOS, with accuracy up to 95%. Bayesian classifiers and Decision Trees were also used as methods of prediction for hospital length of stay [17]. This research focused on patient LOS in a hospital in the range of 61 or more days and compared the strength of the two algorithms with different variable selection criteria. Regression analysis was similarly used in a paper to determine the most important variables in predicting psychiatric patient LOS in an acute Psychiatric Hospital, and to determine if these variables remained valid prediction inputs from one year to the next [21].

Other research focused on predicting LOS for aftercare treatments of specific conditions such as Cardiac Arrest, Stroke, and Appendectomy. One such paper uses ANN to predict LOS using information of Cardiac Arrest patients in the ICU [18]. This study showed the applicability of the ANN in predicting the LOS for operations and revenue management in the ICU to account of unused beds and possible overflows. Another study took a step further into ICU length of stay by developing software called EuroSCORE to allow for doctors to predict a post cardiac arrest patient's level of risk based on specific postoperative conditions and inevitably predict that patient's LOS [22]. This software used logistic regression to determine a patient's risk level and the results of this study showed that EuroSCORE was able to predict patient LOS when the risk levels were low, but tended to underestimate the LOS when the patient risk levels were very high and the patient had many complications.

Some research focused on stroke and post stroke rehabilitation patients using linear regression to determine what factors caused LOS in the hospital to be increased [23]. Another set of research accessed LOS for Appendectomy patients from a business prospective using data

mining techniques [24]. This article used Support Vector Machines (SVM) to predict which patients will have a LOS longer than the "typical" LOS that is reimbursable by the healthcare system. Their results show that these SVM are useful in early prediction of which patients would have longer LOS enabling doctors and administrators to take the appropriate measures.

Finally, there is research surrounding the prediction of patient LOS for ICU patients using Linear Regression with variables that represent only the acuity of the patient and complications they have had since being admitted into the ICU [20]. The end results showed there are ways to use this data mining method to have more efficient use of the ICU beds.

3.1.2.1 Artificial Neural Networks to Predict Patient LOS in the ED

Medical Doctors and scholars from Colby College and Vanderbilt performed a study on predicting a patient's LOS using artificial neural networks from an ED that serves over 42,000 patients annually [25]. The results showed that the neural networks were successful in predicting the training set patient LOS within a range of 2 hours, but when the model was applied to the validation set, the prediction ability fell drastically to a range of 7.5 hours. When the data was broken into subsets of specific chief complaints that included abdominal pain, chest pain, and multiple wounds, both the training and validation were successful in predicting for a range of about 3.5 hours of the patients actual length of stay.

The variables used in the model included number of patients in the waiting room, average wait time for WR patients, emergency department capacity level, average patient acuity, number of patients with beds requested in the hospital but still occupying a bed in the ED, number of patients with health risk indicators (latex allergy, blood-borne disease, or respiratory isolation), number of patients waiting to be discharged, and the ED diversion status. This thesis uses a similar set of variables.

Our results are different than those of the paper largely due to the nature of the ED at BIDMC. This ED serves higher acuity patients and is a primary place for transfers of complicated cases and injuries from other Boston area emergency departments. It also is one of the primary centers for patients suffering from a cardiac arrest due to its 24 hour Angioplasty Program which allows for the completion of a Percuntaneous Transluminal Coronary Angioplasty (PTCA), a common procedure done during a bypass surgery, within a 90 minute window from notification to open artery [14].

3.1.2.2 Information Technology at BIDMC

Another paper that centers on patient LOS is a report done in 2006 by the McCallum Graduate School at Bentley College [15]. This paper is not specifically centered on predicting a patient's LOS, but acknowledges that there are processes that increase the LOS and thus areas that can be streamlined. LOS is defined as the time from which the patient is time stamped as arriving into the ED, to when they are discharged, either to the main hospital (approximately 30%) or out of the ED (approximately 70%). This study used and analyzed 12 months worth of historical ED data taken from the Dashboard system. Two main processes were found to bottleneck ED operations: MRIs and the admission process. They found that the time a patient waits, and occupies an ED bed while waiting for an MRI to be taken, read, and returned takes more than 2 hours and with proper planning could be greatly reduced. Similarly, they found that because nearly a third of all patients coming into the ED will be admitted into the main hospital, the hospital admission process could be better streamlined and result in an overall lower patient LOS.

The results were to be implemented as an Information Technology solution that could be added into the Dashboard system to help with short term, intermediate term, and long term tactical decision making regarding when and how many MRIs and beds to request ahead of time. The thesis by Clay Noyes used these results and suggestions to further streamline ED processes by creating a simulation of the ED operations and then altering variables to find areas of improvement [16].

3.1.3 Observations

In order to better understand the events and operations taking place in the ED, we went on site to BIDMC to ask nurses and doctors questions regarding the environment of the ED. Each patient has a triage level called the Emergency Severity Index (ESI) and the levels range from 1-5, with 1 being the most severe and 5 being the least. A patient of ESI 1 is of top priority and has likely been through some sort of intense trauma. And ESI 5 patient might be coming by afterhours to get sutures removed, and can wait in the ED for a long period of time without a change in their condition. Generally ESI 1 and ESI 2 patients have priority based on their complaint and ESI 3, 4, and 5 are served in a first come first served basis. In this particular ED the general break down of patients by ESI level can be seen in Figure 3.1:

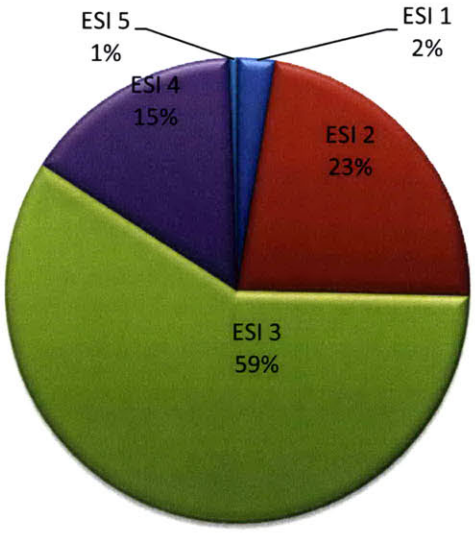


Figure 3.1 ESI Breakdown by Percent

The process of registering and treating a patient is: registration, triage, waiting room/bed, tests, results, discharge. When a patient comes into the ED by their own means (e.g. not via ambulance), he would see a front desk triage nurse and report his chief complaints, have his vital signs taken, and take a seat in the waiting room until a bed becomes available. Once a bed has opened up, he will be taken into the room, seen by a doctor and go to another area for specialized testing such as x-ray, CAT scans, MRIs, or lab tests. The patient will then return to the ED, not necessarily to the same bed as before, and wait for the results to be read and for the doctor to receive these readings. The doctor will assess the next move for the patient which might be a transfer to the main hospital or discharge him to his home. If a hospital bed is needed, a request is made from the doctor to the hospital and the patient remains in the ED until the hospital has a room for the patient. During each of these processes, timestamps are taken in the Dashboard system and valuable information for each patient is recorded. If a patient comes in via ambulance, he is likely to be of a higher triage level, and he will be registered into the system with their vitals and information being transferred straight from the paramedics. These patients will generally be given a bed quickly and then they will go through the same processes as the other patients.

At times, this process is not very optimal and is accomplished in a step by step fashion, with no way of predicting if a patient will be occupying a bed for a long period of time or if he

will be quickly admitted into the hospital. Furthermore, at times the ED is over its maximum capacity with respect to the number of patients being treated and hall space is used or rooms are split to handle a greater capacity.

3.2 Assessment of the 2008 Room Split

In his thesis, Clay Noyes showed through simulation of the BIDMC ED that beds were a major limiting resource, rather than the number of staff working or various testing equipment like MRI machines or CAT scans [16]. He found that if one extra bed were added the result would be an average decrease in the time between the patient being registered and actually being assigned a bed by 12.1%. BIDMC elected to implement this change on September 1, 2008. To assess the effects of this decision, we analyze data three months before the bed split and three months after, a range from July to December for the years of 2007 and 2008.

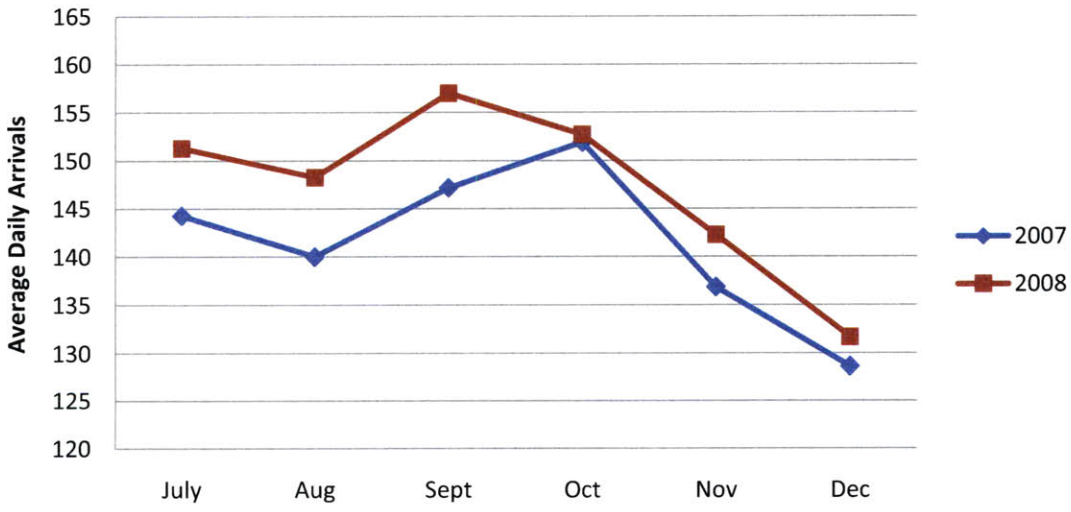


Figure 3.2 2007 and 2008 Average Daily Arrivals

In Figure 3.2, 2007 average daily arrivals are represented as the blue line and 2008 average daily arrivals as the red. Overall there were nearly 29,000 patient arrivals in 2008 and 26,000 in 2007, showing a 10 percent increase between the two years. There are more average daily arrivals for each of the six months in 2008, with October being the only month in which

both 2007 and 2008 are nearly the same. We expect both 2007 and 2008 patient arrivals to have similar curves with no observable differences before or after the September split because the split should does not affect the number of patients arriving to the ED.

When we instead look at the average number of patients in the waiting room (WR), we expect that with an addition of a bed in September 2008, there should be fewer patients. Figure 3.3 confirms this as the average number of patients in the WR in 2008 is consistently lower than those in 2007 after the September split.

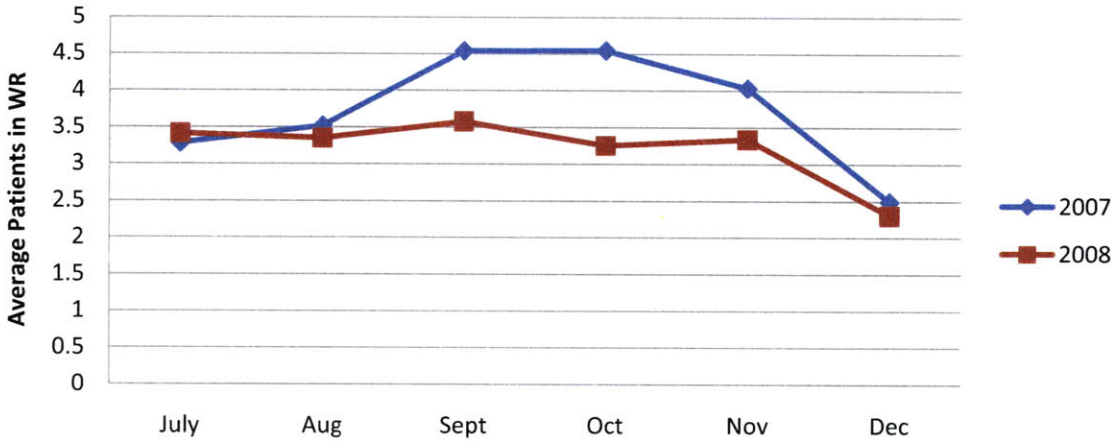


Figure 3.3 Average Number of Patients in Waiting Room

In July and August, the average number of patients in the WR between 2007 and 2008 is fairly similar at about 3.3 patients. In 2007, that number loosely follows the trend of patient arrivals and increases in September and October and then falls in November and December. In 2008, we see a different trend. After the September room split, the average number of patients in the WR only slightly increases in September and decreases again from October to December. As the actual number of patients arriving in 2008 was higher than in 2007, our observation confirms this simulation conclusion that the additional bed has positive results for patients coming into the ED and beds were in fact a limiting resource for BIDMC.

We validated the results of the bed split using patient length of stay (LOS), focusing on ESI 2 and ESI 3 patients, which make up over 80% of all patients seen by the ED. Because ESI 2 and ESI 3 patients are the most common type of patient being treated, they are the patients in which the effects of the change would be most evident; the other ESI's length of stay changes

can be viewed in Appendix B. In Figure 3.4 and 3.5 the LOS is lower for 2008 than 2007, but the greatest differences are observed after September. Overall, the decision to conduct a bed split had a lowering affect on the patient length of stays and the ED waiting room times.

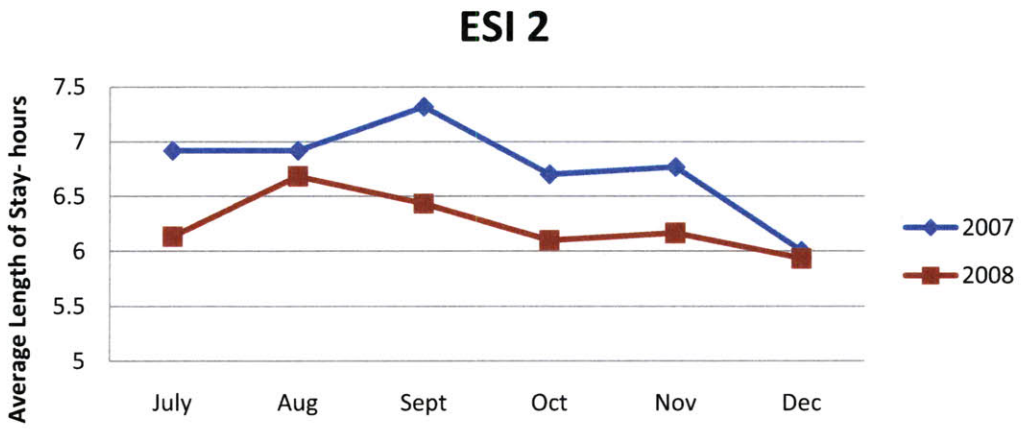


Figure 3.4 ESI 2 Average Patient Length of Stay for 2007 and 2008

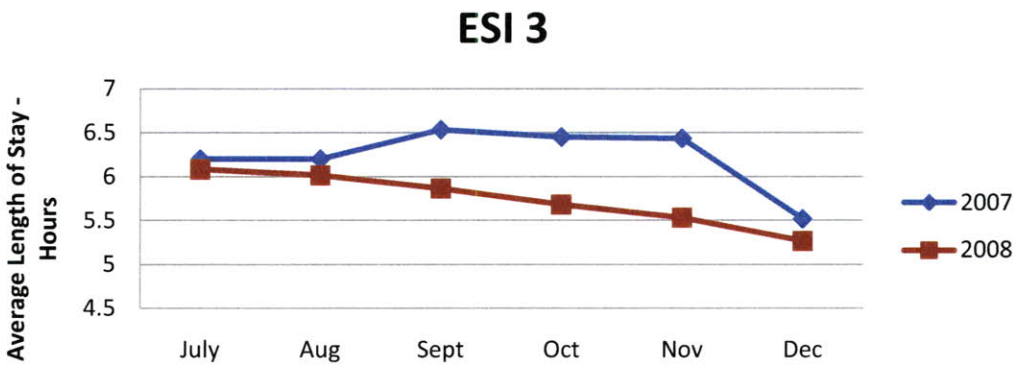


Figure 3.5 ESI 3 Average Patient Length of Stay for 2007 and 2008

3.3 Predicting Patient LOS in Real Time

For this chapter, the problem statement is to use data mining techniques to predict a patient's LOS upon his arrival to the ED based on chief complaint, age, time of day of his arrival, and the waiting room statistics. This chapter discusses the data available, including the preprocessing and mapping that was conducted, the variables chosen for the prediction,

classification as a means of prediction using Logistic Regression, kNN, and Decision Trees, the parameters for validation, the summary of results and conclusions, and applications and extensions of the results.

3.4 Description of Available Data

In our research, we focused on the Dashboard data available from the time period of July 2008 to December 2008. This data includes ED Statistics and Patient Visit information. The ED Statistics consist of count statistics for the waiting room and the ED taken in ten minute intervals. These statistics are taken every day, resulting in 144 readings each day. No individual patient data was included in the ED Statistics (Table 3.1).

Field Name	Description
WR	Number of patients in the waiting room
InDept	Number of patients being treated in the entire department
CDU	Number of patients physically in the Clinical Decision Unit
ADM	Number of patients in the department in "Admitted" status –
REQ	Number of patients that we have requested a bed for, but have not had one assigned (must be cared for by ED staff)
TODAY	Number of patients registered since 12 midnight
RNCOUNT	Number of nurses assigned to patients at that moment in time. *(note that may spuriously spike around change of shift)*
OBS	Number of patients on Observation status
NEW	Number of patients registered in the last 60 minutes
AVGWR(min)	Average waiting time for patients in the waiting room in minutes
MAXWR(min)	Maximum waiting time in the waiting room in minutes
ICU	Number of patients that have had an ICU bed requested.
Tele	Number of patients that have had a Telemetry bed requested
Floor	Number of patients in the dept on ADM or REQ status who are not ICU or Tele status.
ESI-1	Number of patients in the department having acuity level 1
ESI-2	Number of pts in the department assigned ESI level 2
ESI-3	Number of pts in the department assigned ESI level 3
ESI-4	Number of pts in the dept assigned ESI level 4
ESI-5	Rarely used – very minor issues

Table 3.1 ED Statistics Available Data

The Patient Visit data includes all the specific data on each patient that comes through the ED and all the information related to their visit. It is not updated at intervals of time, but through a series of timestamps recorded as each of the events occurs (Table 3.2).

Field Name	Description
MRN	Medical Record Number, modified to be unique to the patient
Age	Patients Age
Chief Complaint	Reason for coming to ED, note often has typos
Room	History of room occupation in ED and what time patient was moved
Visit Milestones	Time into system, time moved from WR to room, time moved to another room, time bed requested, time patient discharged
Physician ID	Unique ID for primary care physician
ED Physician ID	Unique ID for ED attending physician
ED Resident ID	Unique ID for ED resident physicians and time of change
ED Nurse ID	Unique ID for ED nurses and time of change
Referral	Whether or not there was an electronic primary care referral placed
Lab	Times that labs were ordered, resulted and the values
Radiology	Types of radiology studied ordered and time ordered, performed, resulted and type
Diagnosis	Primary and Secondary diagnoses for the visit
Disposition	Home vs Admitted vs Transferred vs Died, note if admitted time a bed was requested, time bed was assigned, area admitted to)

Table 3.2 Patient Visit Available Data

3.4.1 Preprocessing

In order to make sense of the massive amounts of data for six month period the data was available, a series of preprocessing steps are taken. We start with some summary statistics of the average arrivals by month and hour as well as average length of stay by the patients to better understand the dynamic of ED at BIDMC.

Trends of the average hourly arrivals show very similar curves for each day of the week. An average of hourly arrivals over the time period is shown in Figure 3.6. Number of arrivals is at a minimum during the early hours of each morning, then increases greatly during the mid-day period, and then decreases again as evening approaches. Because of this data and the results of regression analysis of the arrival process into the ED, we are able to break the day into three sections: A period from midnight to 0800, from 0810 to 1600, and from 1610 to 2400.

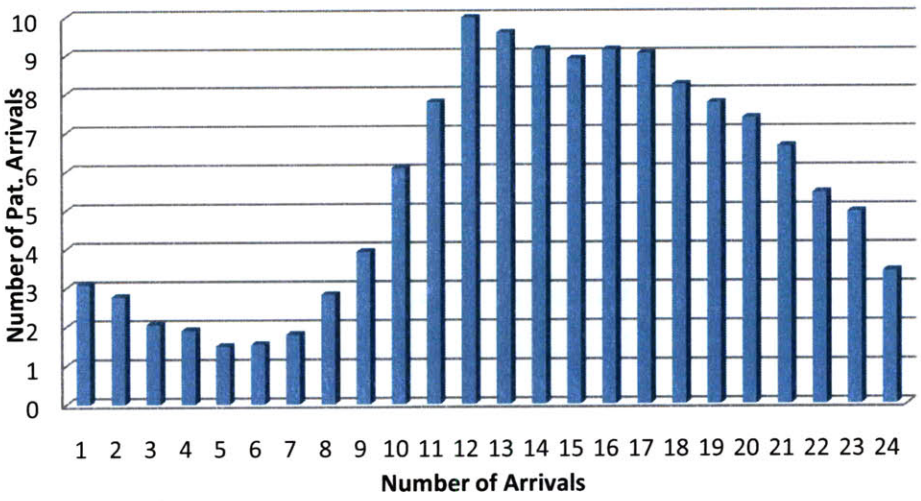


Figure 3.6 Average Number of Hourly Arrivals in a Day for July-Dec 2008

We also observed that the average length of stay was consistently between 5 and 6 hours for each day of the week, and there was not a lot of variance in LOS across the week (Figure 3.7).

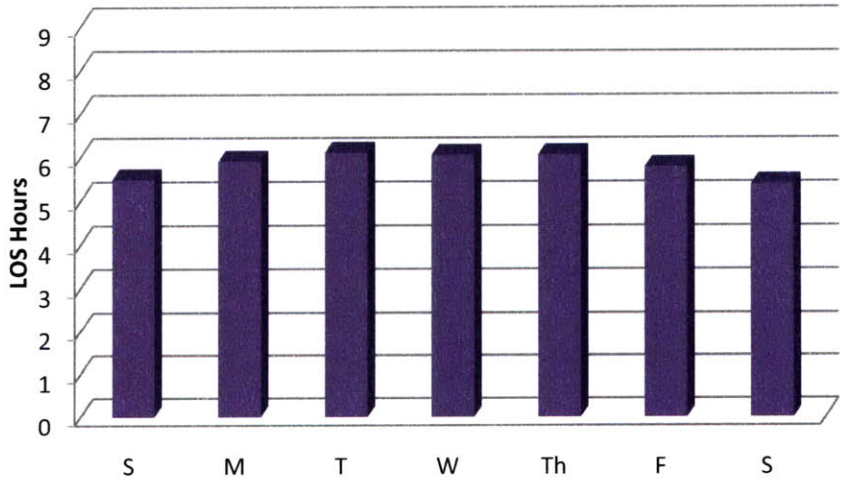


Figure 3.7 LOS vs Day of Week for July-Dec 2008

3.4.2 Mapping

In order to predict a patient's length of stay based on both patient information and elements in the hospital, we need to combine the data into a dataset of both patient information

and the ED statistics. To do this we merge the condition of the ED at the patient's arrival time with the standardized patient information. When the data was first received from the Dashboard system, there were some inconsistencies with chief complaints in the six month period. If someone came in with a wrist injury, it could be entered into the system as WRIST INJ, WRIST INJURY, or even accidently mistyped as WIRST INJ. This led to a total of 7,969 unique chief complaints, many of which represented the same underlying patient grievances. To overcome this problem, we spoke with Dr. Larry Nathanson to get a list of mapped accurate and consistent chief complaints. We then wrote a program that would search through each of the 8,000 unique chief complaints and replace them with the accurate complaints resulting in a much more concise list of slightly over 300 complaints.

3.5 Variable Selection

In order to predict a patient's LOS, we need to remove all of the information that we will not normally know about a patient within moments of his arrival including if he were admitted to the hospital or not, the final diagnosis, and the state of the emergency anytime after the initial arrival. This left us with patient age, time of arrival, chief complaint, acuity, and a variety of ED statistics including number of patients in the WR, department, number of nurses, and number of patients in each of the five acuity levels. To determine the best variables to use, we performed clustering and regression.

3.5.1 Clustering

Using *k*-means clustering we are able to find the variables that would be important in classifying patient LOS. Having too many variables when clustering may hide or get in the way of the true structure of the data, while having too few variables may result in not enough input data to correctly classify the outputs. We only considered the variables that would be accessible when a patient initially arrives at the ED. We then began clustering with different combinations of these variables looking for large amounts of changeability in the clusters with respect to each of the variables and included in the clustering classification of a Long or Short LOS. When there was great variability, we explored these inputs more closely and determined if they were good explanatory variables. For example, if a Long LOS cluster and a Short LOS cluster had high

variability in the Capacity of the ED and the Age of the patients variables, we would consider those variables more closely to determine if they were significant. The variables we elected to focus on are the capacity of the ED, the time of day a patient arrives at the emergency room, the age of the patient, and the chief complaint. In Table 3.3 we show each of the variables we considered and have highlighted in red the variables chosen for prediction.

Variable	Significant
No. Patients in WR	No
Capacity of ED	Yes
REQ	No
TODAY	No
RNCount	No
TOD	Yes
OBS	No
ICU	No
ESI	No
Age	Yes
Chief Complaint	Yes
NEW	No

Table 3.3 Variables Chosen to Predict Patient LOS

Looking at the chosen variables, we can see how they are essential in predicting patient LOS. Specifically, if a patient that is 23 years old comes into the ED with a chief complaint of Chest Pain they will likely be seen and monitored, rarely ever admitted into the hospital and will likely have a longer length of stay. But if a 57 year old patient comes into the ED with the same chief complaint, they are more likely to be transferred quickly to the main hospital because of the risk of cardiac arrest. This trend is seen between patients of different ages for many chief complaints such as a fall, broken bones, lacerations, and abdominal pain.

Similarly, clustering shows that the length of stay of patients is largely related to the condition of the ED. If the ED has a lot of patients in it, patients with less acute chief complaints such as cast removal or laceration, have much longer length of stays than when there are not as many patients in the ED. The time of day a patient comes into the ED was also a very important field in this dataset. As is evident in Figure 3.6, if a patient arrives at 0300 when there are few people being treated, the length of stay is generally shorter than if the patient arrives at 1500

when there is a higher volume of people. Lastly, the chief complaint is a very important factor. Each chief complaint has a large variation of LOS from the next chief complaint. For example, a patient coming in with a stroke or cardiac arrest usually has a shorter LOS than other patients because they are generally admitted into the main hospital very quickly. Other patients, with drug overdose or a mental disorder for example, will be in the ED for a very long time because their treatment includes being continuously monitored and tested.

Acuity, or the ESI level, is not a variable of interest in this paper. For BIDMC, over 80% of the patients are triaged into ESI 2 or 3, and less than 3% are ESI 1 or ESI 5. When we broke the patients down by chief complaint we found that nearly all the patients with a certain chief complaint also generally had the same ESI number and when we clustered using Acuity, there was not the variability we saw with other input variables. For example patients coming into the ED for a Fracture or Laceration, were ESI 3 regardless of their age, the time of day they arrived or any other factor. Similarly, if there were multiple ESI levels in a single chief complaint, they were generally uniformly spread out through the patient length of stays and one particular ESI level did not dominate the short or long overall LOS.

3.5.2 Regression

After determining from clustering the importance of the time of day, the ages of the patients, and the different conditions of the ED, we needed to determine the appropriate way to group these categories. In the case of the time of day, we consider many different alternatives such as splitting the day into 0000 to 1200 and 1210 to 2400, splitting it into three equal parts, possibly three unequal parts, and so on. Using regression, we found the best split is three equal times of day that consisted of the time frames of 0000-0800, 0810-1600, and 1600-2400, which produces an R^2 value of 0.84. The results of the regression that proved to be the best are seen in Table 3.4. We can see that the time of day for the 8 hour blocks is very statistically significant with very high t-Statistics. The day of the week, whether it is a weekday or a weekend is also important, but not statistically significant compared to the other variables. We also look into a variety of other variables and breakdowns for time of day including: 1 hour breakdowns, 6 hour breakdowns, 12 hour breakdowns, looking at each day of the week separately, and considering holidays.

Variable	Estimate	t-Statistic
(Intercept)	28.99	8.49
TOD 0000-0800	-24.66	-61.61
TOD 0800-1600	17.55	44.06
TOD 1600-2400	7.11	17.4
Weekend	-1.93	-6.14
Weekday	1.93	6.14

Table 3.4 Regression Results for Time of Day

We find the best breakdown for the patients age is into three groups consisting of Young (less than or equal to 35 years of age), Middle Aged (between 36 and 65 years of age), and Old (greater than 65 years of age). Finally, we determine the appropriate breakdown of the capacity of the ED. There are 46 permanent rooms in the ED as well as extra beds and occasionally hallway space is used in the cases when there is a large surplus of patients [**16**]. This increased capacity allows for over 60 patients to be treated at a given time. Even though there are always a minimum of 46 beds available, initial capacity is usually reached when there are approximately 42 patients. Using this information from Clay Noyes's thesis on ED capacity and working through many comparisons, we split the capacity of the ED into four categories of Low Fill (LF), Initial Capacity (IC), Capacity (C), and Over Capacity (OC).

Low Fill represents the ED when there are less than 38 patients being treated, open beds, and plenty of doctors and nurses available. Initial Capacity represents the range of time when there are enough beds available in the ED and the WR is generally empty and occurs when there are between 38 and 44 patients in the ED. This is when the beds are nearly all filled up and a queue has developed in the waiting room. Hallway space has not been used yet, but the system is at an initial threshold. Capacity occurs when there are 45-52 patients in the ED. Each of the 46 permanent beds in the ED has been filled, hallway space has started being utilized and there are patients actively waiting in the WR. Over Capacity refers to the times when there are greater than 52 patients in the ED and operations are at an absolutely maximum. Specifically, all beds are taken, all hallway space is being utilized, doctors and nurses are being occupied, and there is a surge of patients in the WR. Figure 3.8 represents the frequency in which the ED is at each of these thresholds. We can see that the majority of the time, the ED is in the Low Fill range as seen with the blue bars, but there are many times when the system is being stressed and overloaded, as seen in the Over Capacity range represented by green bars.

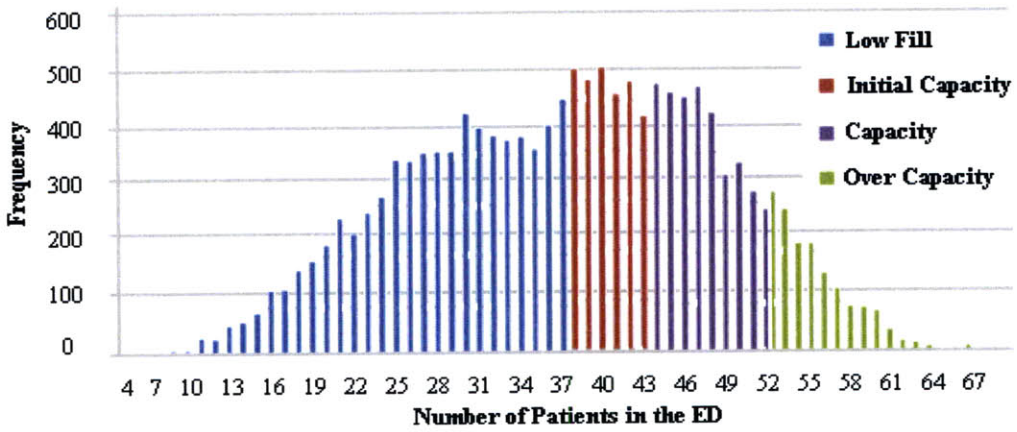


Figure 3.8 Frequency and Capacity for BIDMC ED

3.6 Classification

In order to have enough data for a training and test set for the chief complaints that would be used in the prediction, we elected to focus on chief complaints that had over 100 occurrences in the 6 month period. These 34 chief complaints accounted for over 80% of all the reasons patients came into the ED during the 6 month period in which we studied. We used 75% of the data for a training set and 25% for a test set. The reported results are based on the outcome of the test set. Table 3.5 lists the 34 chief complaints that we used in the study.

Fall	Overdose	Nausea
Trauma	Stroke	Change of Mental Status
Motor Vehicle Accident	Abdominal Pain	Cough
Alcohol Intoxication	Chest Pain	Swelling
Laceration	Head Pain	Diarrhea
Infection	Flank Pain	Suicidal Ideation
Wound	Back Pain	Seizure
Dermatitis	Dyspnea	Syncope
Cellulitis	Evaluation	Sore Throat
Assault	Bleeding	Paresthesia
Fracture	Fever	
Abscess	Dizziness	

Table 3.5 List of Chief Complaints Used in Prediction

3.6.1 Method 1: Binary LOS Prediction

The first method we implement when classifying patient length of stays is to break patient LOS into a binary variable of "Long" or "Short," and use chief complaint, time of day, condition of the ED, and age as our main variables for each patient. We then go through a process of determining what the best cut off for the Long and Short LOS for each of the chief complaints by considering a short LOS to be a stay of less than 2, 3, 4, 5, 6, or 7 hours. In order to best represent the uniqueness of the chief complaints, we go through this process for each and find the best cutoff, which can be seen in Appendix C.

We use both *k* Nearest Neighbors and Logistic Regression for this method. For *k* Nearest Neighbors, we balance the dataset to ensure one class does not dominate the classification and we consider many different numbers of nearest neighbors. For Logistic Regression we also elect to balance the dataset to avoid unnecessary misclassification. We then run the training set through the algorithm and output probabilities between 0 and 1 for each data point, and a threshold that maximizes the classification accuracy. The algorithm and threshold are then applied to the unseen test set for prediction and the final accuracy of the method is assessed.

Both *k*NN and Logistic Regression require an adjustment of parameters during training phases. For kNN, we alter the integer number of neighbors we considered when predicting to have maximum accuracy. This number ranged from 3 neighbors to 19 neighbors using only odd *k* integer values to avoid any possible ties for determining the majority rule. The threshold parameter for Logistic Regression was a probability and ranged from 0 to 1 seen in Table 3.6.

Algorithm	Parameters	Possible Values
Logistic Regression	*p* threshold	[0.27,0.82]
K Nearest Neighbors	*k* number of neighbors	*k*=3,5,7,...,17,19

Table 3.6 Validation Set Parameter Values

After adjusting parameters and trying the variety of different combinations available to determine the best constraints for each method, we come to the conclusion that for kNN, a *k* value of 5 results in a low training error and the lowest overall test error. For Logistic Regression, the probability threshold ranged from 0.27 to 0.82, with an optimal parameter of 0.63.

3.6.1.1 Measure of Accuracy

In both the case of Logistic Regression and kNN classification we discuss the two parts of accuracy, sensitivity and specificity. In the case of this study, if a patient actually has a Short LOS and is classified as a Short length of stay, they are considered a True Short (TS). If they actually have a Short LOS, but are classified as a Long LOS, they are a False Long (FL). Similarly, if a patient has a Long length of stay and are classified by the model as a Long LOS, then they are a True Long (TL). And if they are a Long LOS and are classified as a Short LOS, they are a False Short (FS). Essentially the correct classifications are True Short and True Long (TS/TL) and the misclassifications are False Short and False Long (FS/FL).

Sensitivity is defined as the proportion of all patients that are correctly classified as Short LOS over all the patients with a Short LOS, which includes FL patients, because even though they are misclassified, they are actually Short LOS patients. Mathematically this is defined as:

$$\boldsymbol{Sensitivity} = \frac{TS}{TS+FL} \qquad (3.1)$$

Specificity is similarly defined as the proportion of all patients that are correctly classified as Long divided by those that were classified by the model as Long, again including those patients in the category of FS.

$$\boldsymbol{Specificity} = \frac{TL}{TL+FS} \qquad (3.2)$$

Together these measurements insure that the accuracy value is not misleading. For example, if we are looking at a test set that has 100 patients in which 95 of them have a Short length of stay and only 5 have a Long LOS, we could classify them all as a Short length of stay and have an accuracy of 95%. However we have completely misclassified every single Long LOS. In this case, our sensitivity would be 100% because we did correctly classify all the patients with an actual Short LOS, but our specificity would be 0% because we incorrectly classified all of our Long patients. Looking at both these numbers in conjunction with the overall accuracy gives us a better representation of our model, especially in the case of Logistic Regression and kNN methods.

3.6.2 Method 2: Hourly Range LOS Prediction

The second method used to predict patient LOS was to classify with an hourly range using a Decision Tree. With this method we use the input variables to predict a range of LOS for a given chief complaint based around the average LOS for each combination of variables we are predicting. This length of stay could be a 2 hour range around the average LOS of 3 hours (i.e. 3 hours +/-1 hour) depending on the condition of the ED, time of day the patient arrives, and the patient's age. We provide the LOS ranges for each group predicted to be with 80% accuracy. For example, we can show that with the chief complaint Fracture, a middle aged patient arriving during the first time of day when the ED is at Low Fill, predictability is within 1 hour of the average overall LOS with 80% accuracy. Specifically, we are confident that he will be in the ED 4.5 hours +/- 0.5 hours. Conversely, for another chief complaint, such as Assault, we can only predict a middle aged patient arriving during the first time of day when the ED is at Low Fill within 7 hours of the average overall LOS with an accuracy of 80% (an average LOS of 3.5 hours +/- 3.5 hours). As will be discussed later, there are a few chief complaints that are associated with high uncertainty with respect to time total time spent in the ED.

We first run the training data set through the algorithm and then apply the LOS ranges to the test set to obtain the final hourly ranges we can use to compare results with. For the Decision Trees, we base our LOS ranges on the average LOS for each group of patients. From there a range of 1 hour around the mean LOS is determined and the accuracy assessed. Then we move to ranges of 2 hours, 3 hours, 4 hours, 5 hours, 7 hours, and over 7 hours, in each case assessing the accuracy of each of the LOS ranges.

3.7 Results

After applying the algorithms to the two methods described above, we find that breaking patients down by their chief complaint, age, the time of day of their arrival, and the condition of the ED we are able provide valuable LOS prediction results. In this section we are going to show prediction power and results using no distinguishing variables, which is the baseline scenario, then show results when each patient is broken down by chief complaint, and ultimately show the improvements in predicting LOS when we consider each patient by not only chief complaint, but also age, the time of day they arrive, and the capacity of the ED upon arrival.

At first, taking into account only the patient LOS from all the chief complaints of patients coming into the ED and plotting a histogram, we find a LOS range of about 3 to 10 hours, or a 7 hour range, that encompasses 80% of the all of the patient LOS. This estimate, as we see in Figure 3.9, does not really account for the long LOS (11 or more hours) represented by the right side tail such as patients coming in with Alcohol Intoxications or those with Suicidal Ideation. Furthermore, the left side of figure represents primarily complaints such as Sore Throat and Wound patients, in which the average LOS is generally less than 3 hours.

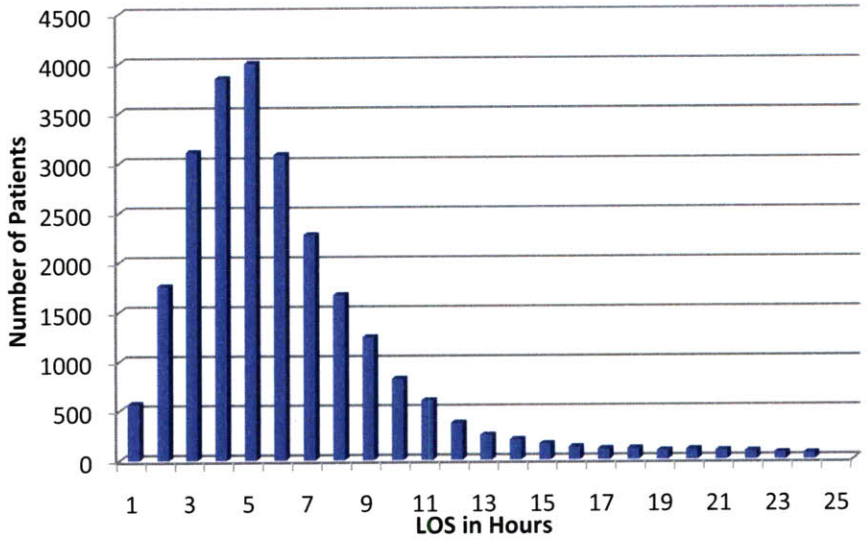


Figure 3.9 LOS by Hour of Patients into BIDMC ED

To account for these problems, we break patient arrivals down further and look into the individual patient's chief complaint and get a more precise range of each based on their individual distributions. When these distributions follow a clear Gaussian form, with small standard deviations, the prediction LOS range is generally fairly small while still encompassing a large portion of the patients. However, when the distribution is uniform, decreasing, or even a highly spread out Gaussian, the prediction range grows very quickly and becomes much less useful.

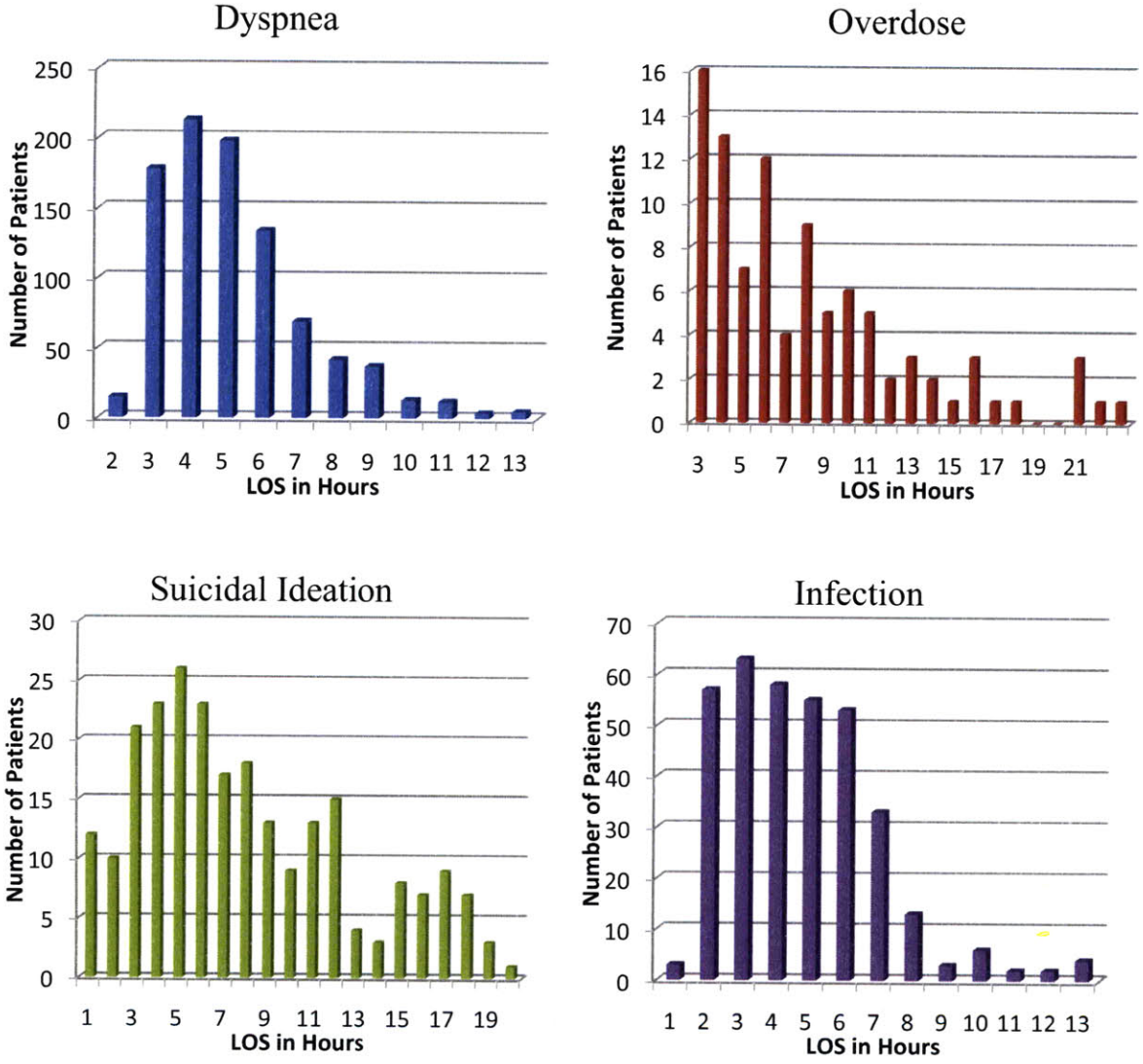


Figure 3.10 Distribution of LOS

Figure 3.10 shows a variety of LOS distributions for specific chief complaints. Dyspnea follows an approximate Gaussian distribution that we are able to predict using the histogram averages and standard deviations to an hourly range of 4 hours (average LOS of 5.15 +/- 2 hours) with 80 percent accuracy. Overdose shows a decreasing distribution for LOS which results in an hourly prediction range of 8 hours (average LOS of 8.04 +/- 4 hours) in order to obtain 80% accuracy. Similarly with Suicidal Ideation and Infection we see two more different types of distributions. Suicidal Ideation has a roughly bimodal distribution with centers around a LOS of

5 hours and again at a LOS of 17 hours. Infection shows a uniform distribution for the bulk of the data, with a LOS between 2 and 6 hours.

Table 3.7 shows the hourly ranges around the average LOS when separating patients only by their chief complaint and predicting to an accuracy of 80%. Laceration is predicted within the smallest range of only +/- 1 hour, while Chest Pain, Diarrhea, and Suicidal Ideation have the biggest hourly ranges.

Chief Complaint	Avg LOS	+/- Range
Change of Mental Status	6.14	1.5
Cough	4.38	1.5
Dermatitis	3.27	1.5
Bleeding	4.89	2.0
Cellulitis	5.86	2.0
Dizziness	5.95	2.0
Dyspnea	5.15	2.0
Fever	5.75	2.0
Abdominal Pain	6.75	2.5
Fall	5.34	2.5
Abscess	5.84	3.0
Back Pain	5.43	3.0
Evaluation	5.41	3.0
Alcohol Intoxication	7.95	4.0
Assault	6.15	4.0
Chest Pain	8.09	5.0
Diarrhea	6.25	5.0

Chief Complaint	Avg LOS	+/- Range
Laceration	3.14	1.0
Fracture	4.88	1.5
Motor Veh. Accident	3.84	1.5
Sore Throat	3.28	1.5
Stroke	4.71	1.5
Syncope	4.54	1.5
Trauma	3.34	1.5
Infection	4.69	2.0
Nausea	5.94	2.0
Paresthesia	5.37	2.0
Seizure	5.86	2.0
Swelling	5.05	2.0
Wound	3.38	2.0
Flank Pain	5.95	3.0
Head Pain	6.15	3.0
Overdose	8.04	4.0
Suicidal Ideation	10.97	6.0

Table 3.7 Prediction Results Using Only Chief Complaint

When we break the data down even further by looking at the capacity of the ED at their arrival, we find there is room for even better prediction ability. In Table 3.8, we have listed the minimum LOS, the maximum LOS, and the average LOS for all patients for each of the capacities of the ED. We are able to see that the average LOS for all patients arriving with various chief complaints changes as the capacity of the ED changes. In most cases, as the number of patients in the ED increases, the average LOS increases. The more acute complaints such as Chest Pain or Stroke do not see this increase as strongly because they are a higher priority and cannot wait to be treated regardless of the number of capacity of the ED. Stroke

patients, for example, usually stay in the ED just under 5 hours at all levels of capacity. The less severe complaints see the largest changes in average LOS as the condition of the ED changes. Abscess patients, for example stay an average of 4.5 hours when the ED is at Low Fill, but that number steadily increases to nearly 8 hours when the ED at Over Capacity. This shows that by separating the patients by the capacity of the ED, we are better able to predict the ranges of their LOS.

	Min LOS	Max LOS	LowFill	InitCap	Cap	OverCap
Laceration	1	7	2.76	3.30	3.61	3.57
Dermatitis	1	8	2.93	3.51	3.68	3.57
Sore Throat	1	9	3.04	3.58	3.68	3.40
Wound	1	9	2.89	3.81	4.10	3.87
MVA	1	9	3.56	3.97	4.46	4.00
Cough	2	10	4.22	4.78	4.34	4.20
Fracture	1	9	4.88	4.81	4.88	5.00
Stroke	2	9	4.82	4.82	4.52	4.57
Seizure	2	13	6.03	5.75	5.45	6.22
Dizziness	2	11	5.81	5.77	6.38	6.63
Cellulitis	1	21	4.79	5.97	7.38	6.60
Abscess	1	16	4.52	6.10	6.59	7.71
Nausea	2	10	5.75	6.14	6.07	6.27
Assault	1	18	5.77	6.45	7.50	7.25
Flank Pain	1	13	5.50	6.57	6.55	6.04
Alcohol Intoxication	2	17	7.47	7.98	9.03	9.78
Overdose	2	23	7.95	8.40	8.72	7.50
Chest Pain	1	23	7.61	8.48	8.92	8.10

Table 3.8 Average LOS by ED Capacity

The following sections show the results of predicting patient LOS with each of the two methods using a combination of the patient information available to doctors when patients initially arrive.

3.7.1 Method 1: Binary LOS Results

kNN is the first algorithm used for prediction of patient LOS with the classifications of Long and Short. As mentioned above, we elect to balance the dataset. The highest we are able to predict binary patient LOS is with an accuracy of 67.28% for Dermatitis, with over 60% for

both sensitivity and specificity. With this method, there are some instances that the accuracy is above 60 percent, but in some of these cases the sensitivity is very high and the specificity is very low, or vice versa. This occurs when the model had little success at predicting one of the classes and simply predicts all the instances as one class. Balancing the data set helps this problem, but as we can see, it was still an issue for predicting patient LOS using kNN.

Looking at the misclassification matrices in Table 3.9, we have two examples of chief complaints that performed the best, Dermatitis and Abscess, Trauma that represents average performance, as well as Swelling that did not predict as well. With these chief complaints that performed the best, we can see that the majority of true short LOS are classified as Short and the majority of true long LOS are classified as Long. There were many chief complaints that had average prediction ability, such as Trauma, in which the majority of the Long and Short LOS are classified correctly, but the overall accuracy is still less than 60%. However, looking at Swelling, we see that the model classifies many of the cases as Long, making specificity significantly higher than the sensitivity. The results for the remaining chief complaints can be found in Appendix D.

Dermatitis

Accuracy 67.28%

Misclassification Matrix

ClassShort	ClassLong	
56	25	ActShort
28	53	ActLong

Sensitivity	69.14%
Specificity	65.43%

Abscess

Accuracy 65.62%

Misclassification Matrix

ClassShort	ClassLong	
33	17	ActShort
20	33	ActLong

Sensitivity	66.00%
Specificity	62.26%

Trauma

Accuracy 57.59%

Misclassification Matrix

ClassShort	ClassLong	
175	148	ActShort
126	197	ActLong

Sensitivity	54.18%
Specificity	60.99%

Swelling

Accuracy 47.55%

Misclassification Matrix

ClassShort	ClassLong	
35	63	ActShort
40	58	ActLong

Sensitivity	35.71%
Specificity	59.18%

Table 3.9 kNN Misclassification Matrix for Dermatitis and Swelling

The chief complaints that have the most success using this method, as well as acceptable sensitivity and specificity values are Dermatitis, Abscess, Alcohol Intoxication, Fever, and Flank Pain. Other chief complaints that have moderate results were Trauma, Cough, Infection, and Chest Pain. Many of the remaining chief complaints are only predictable to a range of about 50% accuracy.

Logistic Regression has a better predicting ability than the kNN algorithm. First, overall accuracy is increased among nearly all of the 34 chief complaints, but more importantly, there is a better balance of sensitivity and specificity. We see that with this method we are less inclined to simply group all of the patients into one class to increase accuracy, increasing only specificity OR sensitivity, but instead finding a balance between the two. Appendix E shows the overall results from Logistic Regression.

Looking at the misclassification matrices in Table 3.10, we again show the two best performers, an average performer, and the worst performer. Abscess and Dermatitis were again the best and able to predict to an accuracy of 74.54% and 68.52% respectively, and had a good balance of specificity and sensitivity. Nausea represents the average chief complaint in which the majority of Long and Short LOS are correctly classified, but again the overall accuracy is still not high enough to be reasonably valuable. Unfortunately, even with an overall increase in accuracy, specificity and sensitivity, we still have some of the same problems we were seeing with kNN as evident in the Seizure misclassification matrix. In this case, the majority of the patients LOS are classified as Short causing the sensitivity to be significantly higher than the specificity and resulting in most of the true Long LOS to be incorrectly classified.

Abscess

Accuracy 74.54%

Misclassification Matrix

ClassShort	ClassLong	
38	15	ActShort
12	41	ActLong

Sensitivity	71.70%
Specificity	77.40%

Dermatitis

Accuracy 68.52%

Misclassification Matrix

ClassShort	ClassLong	
54	27	ActShort
24	57	ActLong

Sensitivity	66.70%
Specificity	70.40%

Nausea

Accuracy 65.87%

Misclassification Matrix

ClassShort	ClassLong	
64	40	ActShort
31	73	ActLong

Sensitivity	61.50%
Specificity	70.20%

Seizure

Accuracy 55.17%

Misclassification Matrix

ClassShort	ClassLong	
57	30	ActShort
48	39	ActLong

Sensitivity	65.52%
Specificity	44.83%

Table 3.10 Logistic Regression Misclassification Matrix for Wound and Seizure

Overall, we find when using a Long or Short binary LOS as the target variable that we are not able to get very convincing results, even when considering all of the main explanatory variables available. With the chief complaints that have the highest prediction correctness and respectable specificity and sensitivity, we are still only predicting with approximately 70% accuracy.

3.7.2 Method 2: Hourly Range Prediction Using Decision Trees

As opposed to the previous method with binary prediction, we concentrate with this method on predicting LOS within a certain hourly range that results in predicting each group of patients with 80 percent accuracy. Table 3.11 is an example for Cough, broken down into each of the 36 combinations of age, time of day, and ED condition for the 251 patients with this chief complaint.

The first column has a 1, 2, or 3 representing the time of day the patient arrived, followed by the letter(s) Y, MA, or O, representing the patients age bracket: Young, Middle Aged, or Old. Finally there are the letter(s) LF, IC, C, or OC that signifies the condition of the ED at the time of the patient's arrival and stand for Low Fill, Initial Capacity, Capacity, or Over Capacity. The next column represents the actual range of LOS for each of the combinations. For example, 1,Y,LF (First TOD, Young patients, at Low Fill), have an average hourly LOS of 3 hours 80% of the time, while 1,O,LF have an average LOS of 4.5 hours with 80% accuracy. The next column represents the range of deviation from the average LOS, in which 1,Y,LF has a range of 1 hour on each side of the average LOS and 1,O,LF has a range of half an hour on either side of the mean. The final column is the number of patients that fall into each category, which allows for us to calculate the average range and average accuracy with appropriate weight distributions. The rows that have been marked with a dash did not have a sufficient number of patients to determine a range. Specifically during the first time of day there simply are not patients coming in with the chief complaint of Cough while the ED is at IC, C, or OC.

Avg Range 2.27 hours
Avg Accuracy 80.87%

Cough	AvgLOS	+/- Range	Tot.Patients
1,Y,LF	3	1	6
1,Y,IC	—	—	—
1,Y,C	—	—	—
1,Y,OC	—	—	—
1,MA,LF	3	1	15
1,MA,IC	—	—	—
1,MA,C	—	—	—
1,MA,OC	—	—	—
1,O,LF	4.5	0.5	5
1,O,IC	—	—	—
1,O,C	—	—	—
1,O,OC	—	—	—
2,Y,LF	4.5	1.5	25
2,Y,IC	3	1	11
2,Y,C	3	1	4
2,Y,OC	—	—	—
2,MA,LF	4	1	41
2,MA,IC	5	1	15

Cough	Act.Range	+/- Range	Tot.Patients
2,MA,C	4.5	0.5	10
2,MA,OC	—	—	—
2,O,LF	4.5	1.5	32
2,O,IC	6	2	9
2,O,C	—	—	—
2,O,OC	—	—	—
3,Y,LF	3	1	7
3,Y,IC	3.5	0.5	4
3,Y,C	3	1	12
3,Y,OC	4.5	0.5	3
3,MA,LF	4	2	7
3,MA,IC	4.5	1.5	11
3,MA,C	3.5	0.5	11
3,MA,OC	—	—	—
3,O,LF	5	1	4
3,O,IC	4	1	7
3,O,C	6	1	8
3,O,OC	6	1	4

Table 3.11 LOS Ranges and Accuracy for Cough Test Set

For Cough, we are able to predict the patients' LOS within the ranges above with an accuracy of 80%, resulting in an average range of 2.27 hours around the mean for all patients (i.e. +/- 1.14 hours on either side of the mean). Similar calculations are done for each of the remaining 33 chief complaints and the hourly range about the mean is recorded in Table 3.12. The overall hourly range of LOS for all of the chief complaints predicted is 4.38 hours. Individual LOS ranges for the remaining chief complaints can be seen in Appendix F.

Chief Complaint	Avg LOS	+/- Range
Abdominal Pain	6.75	2.47
Abscess	5.84	1.71
Alcohol Intoxication	7.95	2.97
Assault	6.15	2.74
Back Pain	5.43	1.92
Bleeding	4.89	1.62
Cellulitis	5.86	1.47
Change of Ment.Status	6.14	1.34
Chest Pain	8.09	4.57
Cough	4.38	1.13
Dermatitis	3.27	0.94
Diarrhea	6.25	2.09
Dizziness	5.95	2.02
Dyspnea	5.15	1.72
Evaluation	5.41	2.69
Fall	5.34	2.01
Fever	5.75	1.97

Chief Complaint	Avg LOS	+/- Range
Flank Pain	5.95	1.77
Fracture	4.88	0.94
Head Pain	6.15	2.59
Infection	4.69	1.59
Laceration	3.14	0.83
Motor Veh. Accid.	3.84	1.59
Nausea	5.94	1.45
Overdose	8.04	3.67
Paresthesia	5.37	1.10
Seizure	5.86	1.85
Sore Throat	3.28	1.35
Stroke	4.71	1.26
Suicide	10.97	4.43
Swelling	5.05	1.54
Syncope	4.54	1.08
Trauma	3.34	1.20
Wound	3.38	1.53

Table 3.12 Summary of Results for Decision Tree

We can see that by using hourly ranges instead of the Short/Long methodology we are able to predict specific ranges for each chief complaint to a much higher accuracy and in a way that delivers specific information regarding LOS for each chief complaint for any of the combination of time of day, ED condition, and age. The chief complaints that have very short LOS ranges include Dermatitis, Laceration, and Syncope, which are all predictable to a range of about +/- 1 hour around their mean LOS. The longer hourly LOS ranges include Suicidal Ideation, Overdose, and Chest Pain, which are all predicted to ranges of +/- 5 hours of their average. These particular chief complaints, however, generally have longer patient LOS, ranging

anywhere from 1 hour to 23 hours, so to predict them within 5 hours of their average LOS is still informative and helpful to ED staff.

3.8 Conclusions

Overall, in this research we found that the Decision Tree Classification provides the best and most descriptive results in predicting a patient's LOS in the ED at BIDMC. It is able to determine within a specific hourly range the average LOS of patients coming into the ED, rather than a binary Long/Short classification, which can be somewhat ambiguous and is not as accurate. The Decision Tree also allows us to see the actual hours a patient will likely be in the ED depending on their chief complaint, age, the time of day, and the condition of the ED, while the other method was not able to do this as successfully. Comparing the results of the Decision Tree to the information originally available to the ED staff, we see the vast improvements these data mining techniques have allowed. Namely, we could decrease the uncertainty of patient LOS from 3.50 hours on average using only overall patient LOS to 2.19 hours on average using a set of information available when a patient first arrives into the ED. Below is a more detailed outlook of these conclusions. Our methodology included several steps: mapping and making chief complaints consistent, using these chief complaints for predicting LOS, and then using multiple variables and individual chief complaint to further narrow the range. In all, we were able to narrow down the range of LOS around the average from 3.50 hours, to 2.60 hours, and finally to 2.19 hours, resulting in an overall 37% decrease.

The Decision Tree is most effective when dealing with chief complaints that do not follow a simple Gaussian distribution (such as Assault or Suicidal Ideation) and those in which the focal variables have the greatest importance on LOS. All but two chief complaints improve their predictability range when compared to using chief complaint alone. Dizziness and Motor Vehicle Accident are the only two that the predicted ranges are worst than using only the chief complaint with no additional information, and each of these saw an increase in prediction range of only 1% and 6% respectively. On average, there is a 20% decrease in hourly LOS range for the total set of data with a maximum decrease for patients with Paresthesia, which shows a 45% decrease. Abscess, Dermatitis, Back Pain, Assault, Paresthesia, Flank Pain, and Fracture each have a percent decrease of over 30% and are highlighted in green in Table 3.13. The overall

improvement using the Decision Tree compared to using only chief complaint to predict the range of patient LOS is a percent decrease of 15.8. The individual percent differences and comparison of ranges between using only chief complaint (CC Range) and the Decision Tree (DT Range) can be seen in Table 3.13.

Chief Complaint	CC Range	DT Range	% Diff
Abscess	3.00	1.71	-43.2%
Dermatitis	1.50	0.94	-37.7%
Back Pain	3.00	1.92	-36.0%
Assault	4.00	2.74	-31.6%
Cellulitis	2.00	1.47	-26.5%
Alcohol Intox	4.00	2.97	-25.9%
Cough	1.50	1.13	-24.7%
Fall	2.50	2.01	-19.8%
Bleeding	2.00	1.62	-19.3%
Diarrhea	2.50	2.09	-16.6%
Dyspnea	2.00	1.72	-14.0%
Change Men Stat	1.50	1.34	-11.0%
Evaluation	3.00	2.69	-10.3%
Chest Pain	5.00	4.57	-8.6%
Fever	2.00	1.97	-1.5%
Abdominal Pain	2.50	2.47	-1.1%
Dizziness	2.00	2.02	0.8%

Chief Complaint	CC Range	DT Range	% Diff
Paresthesia	2.00	1.10	-45.0%
Flank Pain	3.00	1.77	-41.0%
Fracture	1.50	0.94	-37.5%
Syncope	1.50	1.08	-28.0%
Nausea	2.00	1.45	-27.5%
Suicide	6.00	4.43	-26.3%
Wound	2.00	1.53	-23.5%
Swelling	2.00	1.54	-23.0%
Infection	2.00	1.59	-20.5%
Trauma	1.50	1.20	-20.0%
Laceration	1.00	0.83	-17.0%
Stroke	1.50	1.26	-16.3%
Head Pain	3.00	2.59	-13.8%
Sore Throat	1.50	1.35	-10.0%
Overdose	4.00	3.67	-8.3%
Seizure	2.00	1.85	-7.5%
Motor Veh Acc	1.50	1.59	5.7%

Table 3.13 Percent Differences and Range between Histogram and Decision Tree

In this research we found that psychiatric patients are generally the most difficult to predict compared to the other conditions, especially those patients coming in with Alcohol Intoxication, Assault, Overdose, Evaluation, and Suicidal Ideation. We also found some interesting trends related to types of patients and time of day certain patient chief complaints arrive at the ED. First 87.7% of all stroke patients are admitted to the hospital. This means when the ED gets a call of a patient en route suffering from a stroke they can request a bed in the hospital immediately rather than waiting to assess the patient because the likelihood of the patient needing to be admitted is very high.

We predicted 34 of the 304 unique chief complaints coming into the ED in the 6 month period between June and Dec 2008, which accounted for over 80 percent of all patient visits

during this same time period. We found predictive power is highly dependent on the registered chief complaint of a patient and the condition of the ED upon their arrival, as well as their age and the time of day they arrive. Using Decision Trees as a method of prediction, we are able to predict 64.46% of the patients coming in with one of the 34 selected chief complaints within a 4 hour range (+/- 2.0 hours around their average LOS), and over 80% within a 5 hour range, which can be seen in the two graphs in Figure 3.11.

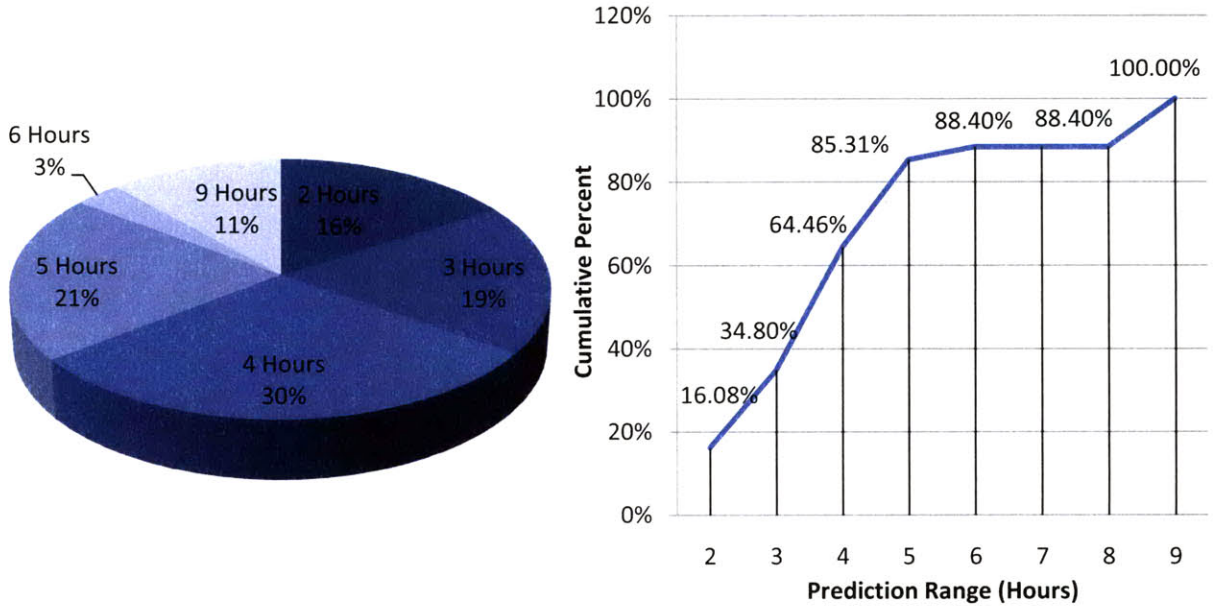


Figure 3.11 Cumulative View of Patients in Predicted Ranges

We showed that Decision Tree Classification has helped the prediction process by decreasing the prediction range originally available by chief complaint alone by 20% for patients coming into BIDMC ED. We assessed two methods of prediction, and found that predicting an hourly range outperformed a Long/Short assessment of length of stay. A related study has shown similar ED LOS predictions using Artificial Neural Networks [25]. The investigators used variables of age, acuity level, coded International Classification of Diseases, 9th Revision (ICD-9) chief complaints, language, presence of laboratory exam, presence of radiology exam, average wait time in the WR, ED capacity, average acuity in the ED, average number of patients requesting a bed in the hospital, ED diversion status, and number of patients with health risk indicators (latex allergy, blood borne disease, or respiratory isolation). The study found that they

were unable to predict all of their chief complaints accurately due to overfitting, but were able to predict within a range of approximately 3.5 hours for their three main ED chief complaints.

The ANN results rely on more inputs than our results do as we elected to only consider inputs from a patient when they initially arrive in the ED. Specifically, we did not include information about the laboratory or radiology work they might eventually need. Also, this study had access to input parameters not available through the Dashboard system including the number of patients with health risk indicators, the diversion status of the ED, and the average number of patients in the ED that have a bed requested in the hospital. Our proposed prediction model allows real time patient LOS prediction within minutes of a patient's arrival to the ED, before they have been assessed by a doctor and before any tests have been requested. With only age, condition of the ED, time of day, and chief complaint, we are able to predict LOS for over 30 main chief complaints within an average hourly range of 4.38, which is only slightly higher than that of the previous research, but with fewer input variables, more chief complaints, and sooner into the patient's overall visit.

3.9 Applications for BIDMC

In a sense, we hope thesis this will lead to even more evolution in Emergency Medicine. Because BIDMC uses their Dashboard system to record all of the ED statistics and the movement of patients through the system, we expect the results of this thesis to allow for real time prediction of expected time a patient will stay in the ED that will give ED staff the opportunity to call in more doctors and nurses if that is where the limitation lies or use additional space such as hallways for their patients, with an overall goal of giving administrators an analytical look into their ED operations. Similarly, if they are not operating at capacity and there are no long staying patients registered, they can send staff home or perhaps take in other hospitals' transfers. It also allows for doctors to see the types of complaints that are most frequently coming in during certain times of the day, the patients that are generally admitted so the doctors can request tests and beds in the hospital earlier, and lastly in the long run allow doctors to hopefully see improvements in patient length of stays reflecting more efficient and streamlined Emergency Department operations.

[This Page Intentionally Left Blank]

Chapter 4

VAST Dataset: Real Time Clustering and Information Processing

Throughout this thesis we have discussed the idea of clustering and how it can be helpful in determining the natural groupings of a dataset with many variables. In the previous chapter we used it as preprocessing before we applied supervised learning techniques for prediction. We have discussed *k* means clustering as the most common iterative method and how k represents the number of centroids used for clustering. But we have yet to discuss what integer value k should take on to correctly represent the clusters. In certain datasets, when there are definite, observable groupings, the value of k is clear and would simply be the number of groupings. When the number of clusters is not clearly determined, one can look at the total sum of squares error for a variety of k values and visually inspect for a "kink" to locate the optimal number of clusters [4]. However, sometimes this kink is not very apparent or one person may think a kink exists while another does not. Similarly, after we cluster a dataset we often want to quickly and efficiently see the similarities and differences between clusters to make sense of the variables and their interactions.

In this chapter, we present the problem statement, background on the current uses of data mining in visual analytics and cluster analysis, of methods to determine optimal clusters and differences between clusters, results, conclusions, and applications. Using the VAST dataset of immigration landing geospatial locations and temporal information, we apply a variety of methods to find the optimal number of clusters automatically. This information is then applied to a visual analytics program to allow a user to quickly see the main variables and elements of each cluster, as well as see differences and similarities between clusters. These results will make sorting through huge amounts of data less reliant on only the obvious patterns a human eye can pick up and provide more information than simple count statistics like averages and standard deviations.

4.1 Problem Statement

There are cases in which we need real time updates in a system, such as visual analytics, to provide analysts with processed information as quickly as possible. For this dataset on geospatial migrant boat landing patterns, the problem statement is to use data mining techniques to develop and evaluate methods that will provide the mathematically optimal number of clusters and answer questions regarding these clusters in near to real time. This chapter discusses the work previously done in this field, data available, the methods used for optimal clustering and determining differences between clusters, the summary of results, and finally application of these results into a visual analytics model that serves to provide visualization techniques for exploring and interacting with data.

4.2 Background of VAST

The VAST dataset comes from the Institute of Electrical and Electronic Engineers (IEEE) Symposium on Visual Analytics Science and Technology. The symposium was founded in 2006 and is the first international symposium dedicated to advances in Visual Analytics [26]. Visual analytics is the science of analytical reasoning supported by human interacting visual interfaces in which the user controls the analytics tools and techniques to manipulate massive amounts of data into valuable information, while providing timely, defensible, and communicable

assessments [27]. Each year the IEEE releases a main challenge that is made up of two to three forged data sets and a variety of mini challenges. The goal is for the competitors to work through the problems and provide a visual analytic tool that allows for ease of presenting and solving the mini challenges and the overall challenge. The dataset used in this thesis is from the 2008 VAST Challenge which was based on a fictitious migration from an island named Isla Del Sueno to various locations in Florida and Mexico.

4.3 Description of Available Data

We focus on the geo-temporal analysis of migrant boat entries, specifically to characterize the choice of landing sites and their evolution over time, characterize the geographical patterns of interdiction over a three year period, and comment on the successful landing rate over the time period using statistics and data mining techniques. The analysis and conclusions in this chapter are then applied to a visual analytics tool that allows for real time information updates. This data includes longitudes and latitudes for the location the migrant boat was encountered or the location of its landing, when the interdiction or landing occurred, information regarding each particular boat, and in some instances the longitude and latitude of the launch point on Isla Del Sueno. An overview of the data can be seen in Table 4.1.

Field Name	Description
Encounter Long	Longitude of the boat interdicted by a Coast Guard vessel
Encounter Lat	Latitude of the boat interdicted by a Coast Guard vessel
Landing Long	Longitude of the boat at its final landing point
Landing Lat	Latitude of the boat at its final landing point
RecordType	Whether the boat was interdicted or successfully reached land
Passengers	Number of passengers on the boat
USCG_Vessel	Coast Guard vessel that interdicted the migrant boat
Encounter Month	The month the boat landed or was interdicted
Encounter Day	The day the boat landed or was interdicted
Encounter Year	The year the boat landed or was interdicted
Record Notes *	Names of passengers on the migrant boat
Num Deaths	Number of deaths on the migrant boat
Launch Long*	Longitude of the launch point on Isla Del Sueno
Launch Lat*	Latitude of the launch point on Isla Del Sueno
Vessel Type	Type of boat the migrants are using to travel (Raft, Rustic, Go Fast)
	* Data not always available in this field

Table 4.1 VAST Available Data

4.3.1 Preprocessing

Before focusing solely on clustering, a series of preprocessing steps were taken to better understand the data. We started by looking into some of the summary statistics such as determining how many departures there were from the island each year, and statistics on the number of landings versus the number of interdictions.

Figure 4.1 shows the number of number of departures per year. There is a distinguishable increase from 2005 to 2007 and nearly double the departures in 2006 than in 2005. This increase in departures is due to the increasing success rate of landing the boats were having across this time frame (Figure 4.2); if the immigrants were able to make it to land more and more effectively, more people would be willing to attempt the journey.

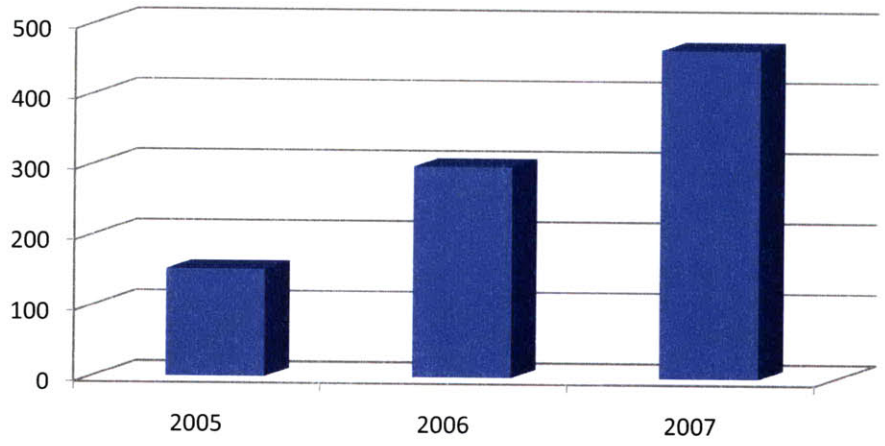


Figure 4.1 Frequency of Departures by Migrant Boats per Year

Looking next into patterns with actual landings and interdictions, we see in Figure 4.2 that although the number of interdictions actually increases over the three year period, the number of landings increases at a higher rate. The interdictions are labeled with a blue line and the landings with a red line.

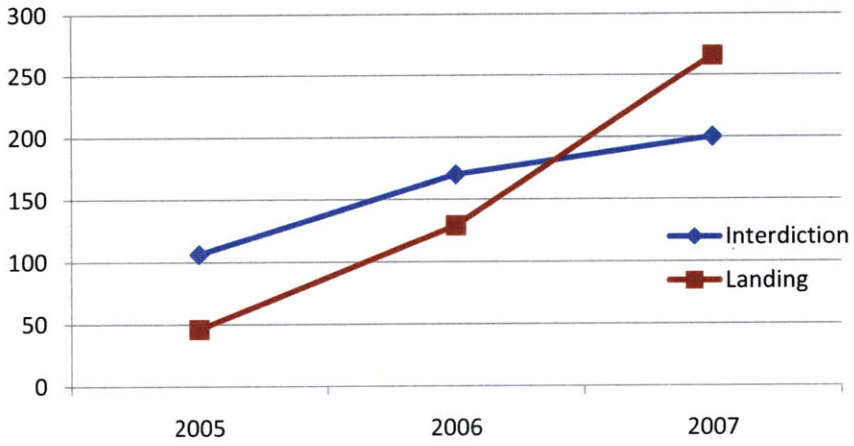


Figure 4.2 Number of Interdictions and Landings by Year

The number of interdictions during this time frame likely increased due to the fact that there were physically more boats attempting to make the journey. When we consider the percentages of interdictions versus landings, we actually see how the migrant boat success rate changed throughout the three year period. In 2005, less than one third of the migrant boats were successfully making it to land, in 2006 that number increased to over 40%, and by 2007 nearly 60% of the boats were evading United States Coast Guard (USCG) vessels and successfully landing as seen in Figure 4.3. This particular figure answers two of the main questions of the VAST challenge regarding the successful landing rate over the time period and the evolution of the USCG interdictions.

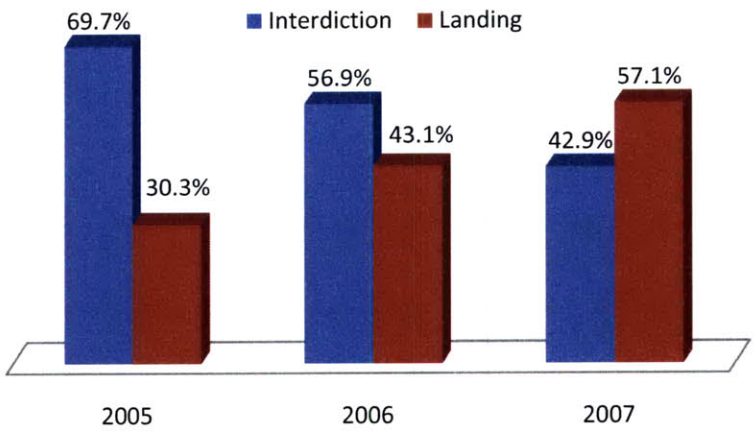


Figure 4.3 Percentage of Interdiction and Landings by Year

Looking next at the types of vessels being used, we see that there are only three distinct options: Rustic, Go Fast, and Raft. A Rustic is a middle sized boat that averages about 14 to 15 passengers and in the case of this data, it is used about 64.4% of the time. A Raft is the second most frequently used vessel, 20.7% of the time, and holds an average of 5 people. Finally is a Go Fast, which is similar to a speed boat. This vessel is used 14.9% of the time and holds between 24 and 25 people. None of the boats was any more or less likely to be interdicted than the others, as each has about a 50% successful landing rate. Additionally the ratio of vessel type remains the same from year to year with about 65% being Rustic, 20% Raft, and 15% Go Fast, that is, there is not a change in vessel selection as time went on or as the locations of landing evolved. We can see this is in Figure 4.4. The Rustic Vessel is labeled in Blue, the Go Fast in Red, and the Raft in Green. The first two columns represent the number of landings and interdiction by each type of vessel, and the remaining columns are type of vessel by year.

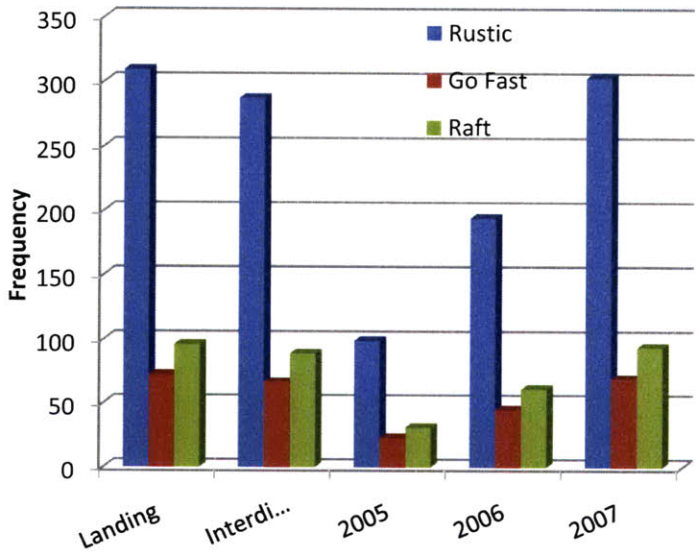


Figure 4.4 Type of Vessel by Landing, Interdiction, and Year

Lastly we look into the evolution of the landing and interdiction locations over the three year period. In 2005, the landings are highly grouped around the Florida Keys, in 2006 there are still landings in the Florida Keys, but landings then moved up the Florida coast on both sides as well as in to Cancun, Mexico. Finally in 2007 there are the majority landings in Cancun and a few scattered along the coastal region of Florida, still including the Florida Keys. Figure 4.8

below shows a scatter plot of the landing location longitudes and latitudes by year with 2005 represented by the markers in Blue, 2006 in Red, and 2007 in Green. The interdiction locations follow a similar trend as the landing locations; there are many interdictions in the Florida Keys in 2005, but as the landing locations begin to expand into other areas, the interdictions begin to follow into these areas as well. In 2006 and especially in 2007, as the migrant boats begin to land in Cancun, the interdiction areas do not follow, which largely has to do with the jurisdiction of the USCG.

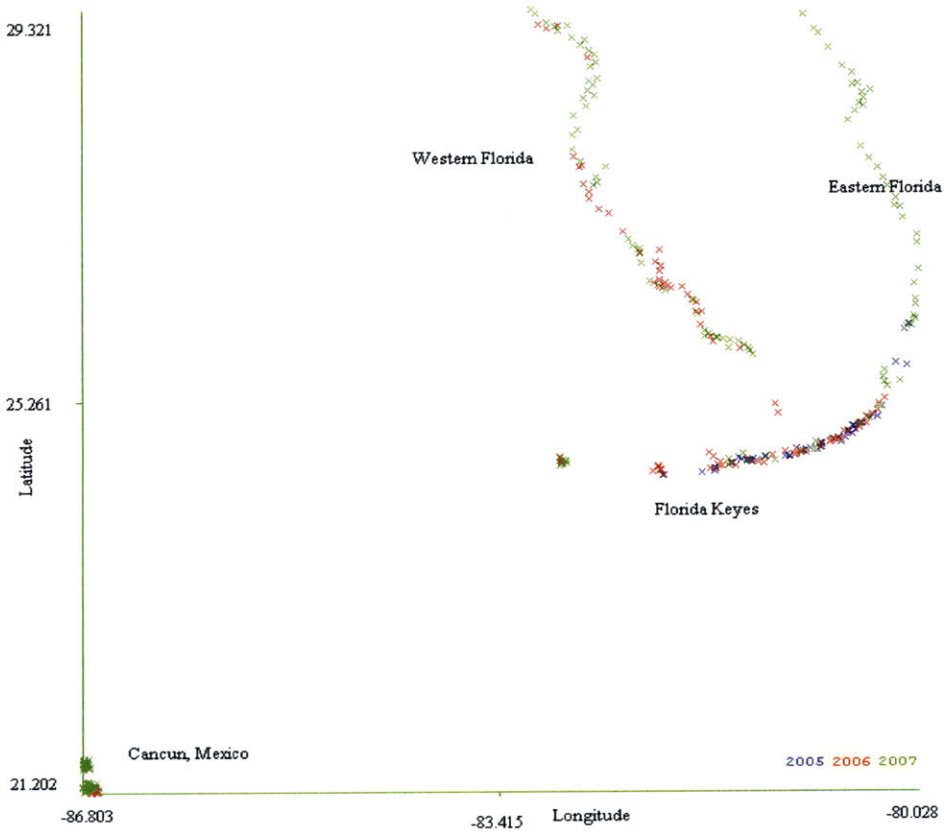


Figure 4.5 Landing Locations by Year

4.4 Literature Review

Currently, in the field of visual analytics, an analyst will observe raw data in the form of historical reports, statistics, or plots, and will report interesting trends and facts. Problems arise with this method when there is too much data for a single person to handle and it becomes hard

to see subtle patterns and interactions accurately. Similarly, the analyst may be biased to a particular set of variables or attributes of a dataset, find patterns where they do not actually exist, or miss patterns they are not accustomed to seeing. Other than graphical representation and basic tools, there are not a lot of visual aids in the field of visual analytics, which is why the VAST symposium works to create competitions and encourage advancement in the field. In the future, we hope to see an integrated interaction between an analyst and various data mining techniques using both visual and analytical technologies to find the underlying patterns and correlations not easily observable by the analyst alone.

Our goal is to use data mining techniques to take the reliance off the analyst and to find valuable information with human guided support. We want to look beyond the simple questions that graphs and averages can answer, and instead look into clustering and find which pieces of data are unusual and then consider the pairwise differences of each cluster compared to every other cluster. We also want to be able to automatically detect the mathematically optimal number of clusters without depending on the user. These things have all been done in the past separately, but we combine them all into an interactive analysis tool.

4.4.1 Visual Analytics

Palantir Technologies is the front runner each year in the VAST symposium and is the same company that developed PayPal [28]. They are working to bring a new approach to building contextually grounded visual analytics environments and produce easy to use interfaces and workflows to reduce human workload and automate anything that would be obvious to a user. Their work for the VAST Symposium Boat Mini Challenge showed an image of Florida and had each of the landings plotted with information about their date and the type of boat as a label. This resulted in a cluttered image with no visual differentiation among data points, over lapping icons, and no other information about a landing other than a simple geospatial plot (Figure 4.6). However, they also had a temporal representation of the data that effectively showed the evolution of landing zones over time and accurately displayed the progression over the 3 year period.

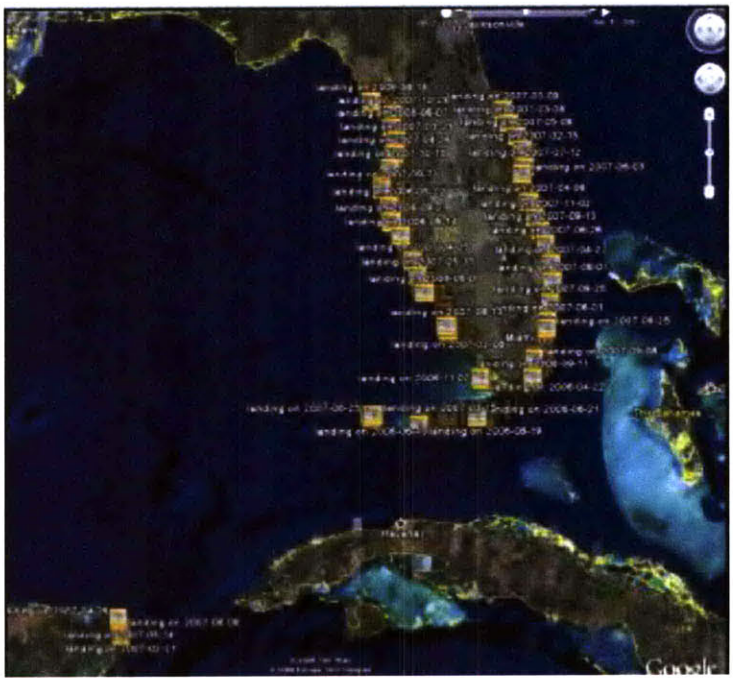


Figure 4.6 Palantir Representation of Landing Zones for VAST Dataset

When doing their analysis, Palantir chose to separate landing zones arbitrarily into Southern Florida, Eastern Florida, Western Florida, and Mexico. With data mining we are able to remove the arbitrarily chosen regions and instead cluster the data to find actual partitions for the landing locations. We also look beyond just physical landing location over time and instead consider the number of deaths, number of passengers, and type of vessel used to compare different regional clusters and make more detailed information available to the user.

There are many other important contributors to the VAST symposium and visual analytics as a whole. Two of these contributors are Oculus Info Inc and the North-East Visualization and Analytics Center (NEVAC). Both focused on using multiple diagrammatic perspectives to support relationship analysis and understanding of behaviors in time, but still lacked solid underlying analytics for their programs [29]. Again, the regional breakdowns of the landing locations were determined without any investigative basis and there was still a lot of supplementary information in these landing locations that was not considered.

4.4.2 Determining The Optimal Number of Clusters

There has been a wide variety of research done in the data mining field to determine the optimal number of clusters using unsupervised learning. These methods generally focused on comparisons of within cluster dispersion, between cluster dispersion, considering nearest neighbors, or assessing the density of data points within an area. One such paper used the Gap Statistic to estimate the number of clusters in a data set [30]. This technique used the output produced by any clustering algorithm and compared the changes of inter-cluster distances to that of an expected reference distribution. The results showed that this method worked well for well-separated clusters, but when the data were not well separated the results were not as strong.

Another study focused on the structural characteristics of clusters in the partitioning process [31]. Here, a validity index was determined by looking at the inter-cluster minimum distance (ICMD) and the mean intra-cluster distance (MICD). ICMD is a pairwise comparison between each of the k clusters, which would drop drastically when the dataset was over partitioned. The MICD is the average distance of the data points to the centroid. When the optimal number of clusters was reached this number also abruptly decreased. The results of this experiment showed the method had success in determining the optimal number of clusters with datasets with some noise as well as object extraction in real images.

K means clustering, although popular, is not always an applicable clustering method, especially when data does not fall into convex regions. When this is the case, density based clustering can be used which allows for the identification of arbitrary, not necessarily convex regions of data points that are densely populated [32]. This paper determined the optimal number of clusters using a nearest neighbor approach and used densely populated regions of data points as possible cluster centers. Each cluster was expanded until the density fell below a parameterized threshold and the resulting cluster widths and centers were determined. This method had success using density based clustering on noisy datasets and images.

A different clustering algorithm used Renyi's entropy as a similarity matrix rather than focusing on the traditional approaches of the partitional algorithms like k means or hierarchical clustering algorithms [33]. The main idea of this research was to assign a data pattern to a cluster, which compared to all other clusters, increased its within cluster entropy the least, using differential entropy clustering. The result created a hierarchy of clusters with a between cluster

entropy value by which the optimal number of clusters was determined. This method had success on both artificial and actual datasets including those with were non-convex.

4.4.3 Differences between Clusters

The Jaccard coefficient is a statistic used to compare the similarity and differences between two sets of data. It is used in a variety of different fields including data mining to compare cluster partitions as well as biology to compare aspects like species diversity [34]. In the case of data mining, the Jaccard coefficient measures the overlap of two clusters based on their attributes, but the presence of the attribute must be binary, and the resulting value will be a percentage of similarity ranging between 0 and 100. The equation takes the total number of times both clusters have the similar attribute, divided by the sum of the total number of attributes in which one cluster has an attribute while the other does not, or vice versa, and the total number of times in which both clusters have the same attribute. This results in the percentage of time both clusters have the same attributes. Because this method deals only with a binary attribute is present or not, it is not very applicable in the comparison of the clusters for this dataset since each cluster will generally have the same attributes, just the number of instances for each attribute will change.

An extension of the Jaccard coefficient is Cosine Similarity. This value measures the similarity between two vectors of any dimension by finding the angle between them [3]. It is measured by dividing the dot product of the attributes of each cluster by the magnitude of the vectors making up those clusters. This method can be used for the non binary case, which increases its applicability. As long as the vectors are of consistent dimension, they can be compared and computed where smaller differences between clusters are then represented by smaller angles and larger differences are measured by greater angles. We can easily implement and use Cosine Similarity for this dataset because the data can be broken into vectors representing each of the clusters.

There are other methods of determining similarities and differences between partitions of information including a simple calculation of the percent difference and correlation [7] or pure distance measures in the case of coordinates.

4.5 Variables

The variables we use to determine the optimal number of clusters are the encounter longitude, encounter latitude, and the encounter year. These variables represent the geospatial inputs of the dataset and allow for us to characterize and show the evolution of migrant boat landing points over time. When we consider cluster differences, we add information to distinguish the clusters including the number of passengers on the migrant boat, the number of deaths on each boat, and the type of migrant boat used. The VAST dataset consists of 917 separate entries, 441 of those being successful landings, which are the focus of determining the optimal number of clusters and the differences between clusters.

4.6 Determining Optimal Number of clusters

Using k-means as the clustering algorithm for this data set, we determine the optimal number of clusters using three specific methods: The Gap Statistic, the Validity Index, and the Silhouette Value. Each method focuses on a variation of finding between cluster distances and inter-cluster distances for each of the points in an established cluster.

4.6.1 Gap Statistic

The Gap Statistics compares the changes in within-cluster dissimilarity, defined as W_k, as a function of the number of clusters k. The curve $\log(W_k)$ is then compared to the curve obtained from data uniformly distributed over a rectangle containing the data. We then estimate the optimal number of cluster, k^* to be the place in which the gap between the two curves is the largest.

To determine the within-cluster dissimilarity we use a distance metric which can be Euclidean distance or the Manhattan Distance seen in Chapter 2. We start by clustering our data into k distinct clusters: $C_1, C_2, C_3, \ldots C_k$, where C_r denotes the elements in cluster r and n_r represents the number of elements in cluster r and $d_{ii'}$ represents the distance between observations i and i'.

To calculate the pairwise distance for all the point in cluster r, let

$$D_r = \sum_{C_r} d_{ii'} \qquad (4.1)$$

and let W_k be the pooled within cluster distance around each of the k cluster centroids:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \qquad (4.2)$$

We then generate B reference datasets using a uniform distribution obtained in the range of actual dataset and cluster each reference set giving W^*_{kb} where $b=1,2,3,\ldots,$ B and $k=1,2,3,\ldots,K$. We then compute the estimated Gap Statistic:

$$Gap(k) = \frac{1}{B}\sum_b \log(W^*_{kb}) - \log(W_k) \qquad (4.3)$$

Because there are B reference sets created, we compute the standard deviations of the estimate gap statistics and call it sd_k. We then define $s_k = sd_k\sqrt{(1+\frac{1}{B})}$ to represent the variation present when we are sampling from multiple reference sets and choose the optimal number of clusters via:

$$k^* = smallest\ k\ such\ that\ Gap(k) \geq Gap(k+1) - s_{k+1} \qquad (4.4)$$

An example of the output graph of the Gap Statistic showing Gap versus k values can be seen in Figure 4.7. On the left side, the k^* is clearly 2, because it satisfies equation 4.4 and has a clear maximum at $k=2$. On the right hand side, the k^* in this example is also at $k=2$ because it satisfies equation 4.4, however the gap statistic begins to rise somewhat drastically again at $k=6$ clusters, implying that there are two well separated clusters and more less separated ones. In cases such as these, it is important to review the entire gap curve rather than simply finding a local maximum [30].

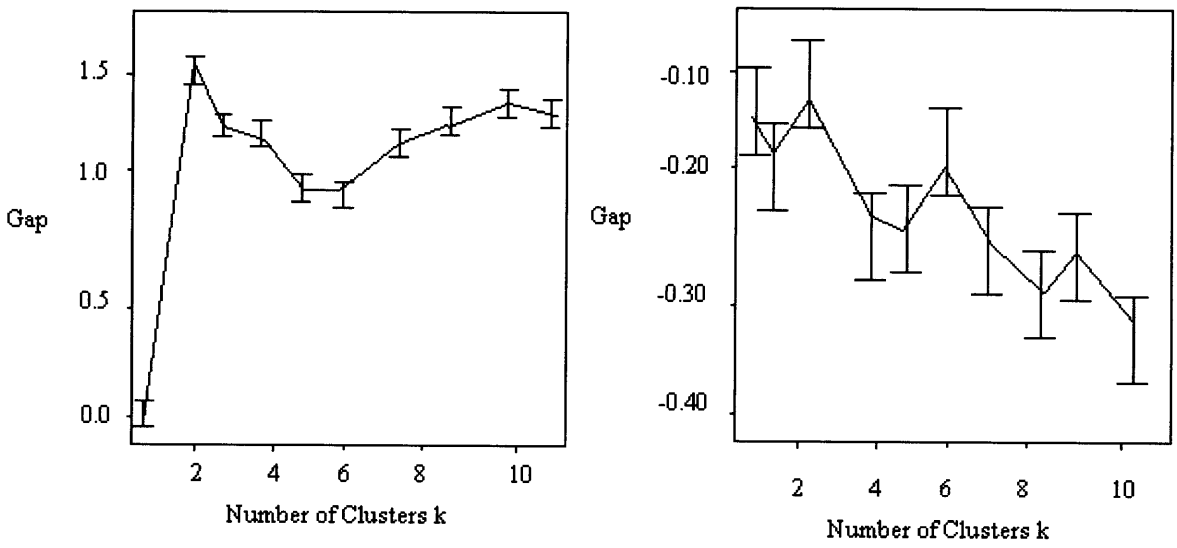


Figure 4.7 Example of Gap Statistic

4.6.2 Validity Index

The Validity Index differs from the Gap Statistic in that it is also looking at the between cluster distances and not only the inter-cluster differences, considering the structural differences around the optimal number of clustering in the partitioning process. It combines these values to produce an index number that is at a minimum when the optimal number of clusters is reached.

This paper considers the distance between different clusters to be the mean intra-cluster distance (MICD) and for a cluster r, is defined as:

$$\boldsymbol{MD_r = \sum_r ||v_r - C_r||/n_r} \qquad \textbf{(4.5)}$$

where C_r is the elements in the cluster r, v_r is the centroid of cluster r, and n_r is the number of elements in cluster r. When the dataset is under clustered, at least one cluster will have a large MICD, but as we move to k^* and past it, the large MICD will drastically decrease.

The inter-cluster minimum distance (ICMD) represents the minimum cluster distance within a single cluster and is defined as:

$$\boldsymbol{d_{min} = min_{i \neq j} ||v_i - v_j||} \qquad \textbf{(4.6)}$$

where v_i and v_j are centroids for each of the cluster centers. This value is large when the dataset is under partitioned and as it becomes over partitioned, the ICMD becomes very small because at least one of the clusters is subdivided too much.

To easily see the drastic changes in both the MICD and the ICMD we define two variables that represent a function of the MICD and the ICMD, v_u and v_o respectively.

$$v_u = \frac{1}{k}\sum_{r=1}^{k} MD_r, \quad for\ 2 \le r \le r_{max} \qquad (4.7)$$

$$v_o = \frac{r}{d_{min}}, \quad for\ 2 \le r \le r_{max} \qquad (4.8)$$

The v_u shows the mean of MICD over each cluster r and measures the compactness of every cluster. When the data are optimally or over partitioned, every cluster becomes compact and v_u becomes very small, therefore this value is large for $r < k^*$ and small when $r \geq k^*$. Conversely, v_o consists of the minimum distance between clusters as the denominator of the function, so when the data are under partitioned the d_{min} value is large, and v_o yields a small value. But when the data become over partitioned, the d_{min} value becomes very small because a cluster has been over divided, and the v_o value spikes. This function, then has very large values for $r < k^*$ and very small values for $r \geq k^*$, the opposite of v_u.

To compute the actual validity index, we normalize v_u and v_o to adjust for the differences in their numerical scales. Let the normalized values of v_u and v_o be v_uN and v_oN, and to lie between 0 and 1:

$$v_{ur}N = \frac{v_{ur} - v_{umin}}{v_{umax} - v_{umin}} \qquad (4.9)$$

$$v_{or}N = \frac{v_{or} - v_{omin}}{v_{omax} - v_{omin}} \qquad (4.10)$$

where v_{ur} is the v_u value for cluster r, $v_{umin}$is the minimum v_u value for all clusters, and $v_{umax}$is the maximum v_u value for all clusters. Similarly, v_{or} is the v_o value for cluster r, $v_{omin}$is the minimum v_o value for all clusters, and $v_{omax}$is the maximum v_o value for all clusters.

The final value is simply the sum of these two values and is defined as:

$$v_{SVr} = v_{ur}N + v_{or}N \qquad (4.11)$$

The goal is then to find the optimal cluster, k^*, with smallest value of v_{SVr} for $r=2$ *to* r_{max}. An example of a data set and the resulting index function can be seen below in Figure 4.8. The v_uN value decreases and the v_oN value increases sharply at 6 clusters, which from the input data picture on the left, we can see is the optimal number of clusters.

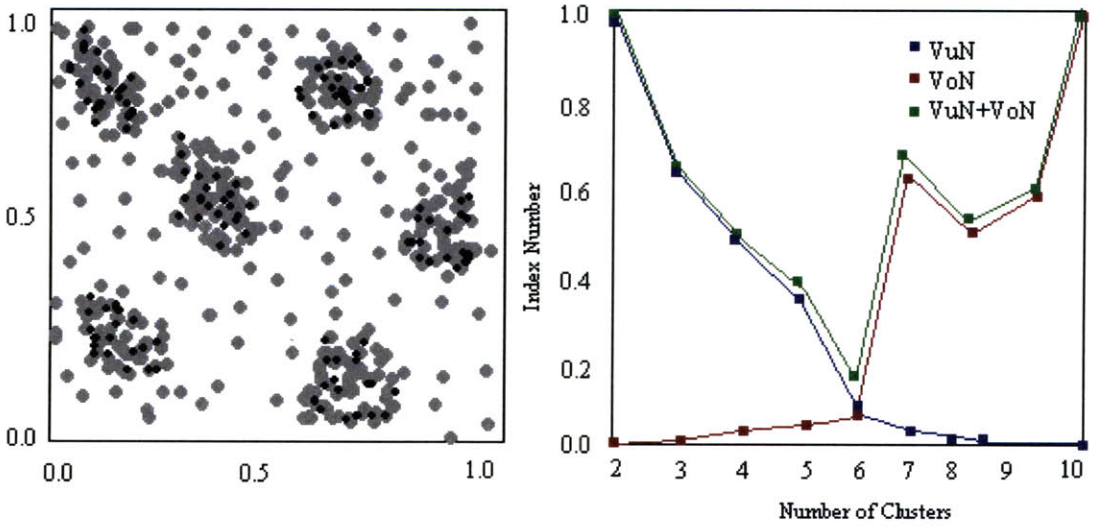


Figure 4.8 Example of Validity Index

4.6.3 Silhouette Value

The last method for mathematically determining the optimal number of clusters is using a Silhouette Value. The Silhouette Value measures how close the elements in one cluster are to points in the closest neighboring cluster. Each data point receives a Silhouette Value that ranges from -1 to +1. If the Silhouette Value is close to +1, the element has been assigned to the appropriate cluster. If the Silhouette Value is close to 0, it means it could be assigned to its current cluster or the next closest one, lying approximately equally between the two clusters. Finally, if the Silhouette Value is -1, the element is likely part of the wrong cluster. The Silhouette Value is defined as:

$$S(i) = \frac{b(i)-a(i)}{\max\{b(i),a(i)\}} \qquad (4.12)$$

where *a(i)* is the average distance of element *i* to all other elements its assigned cluster, *b(i)* is the average distance of element *i* to all other objects in the next closest cluster.

Each element in the dataset is given a silhouette value and each cluster is assigned a silhouette value which represents the average of all of the silhouette values of the element in that cluster. Finally, a silhouette value can be assigned to the dataset as a whole and is again the average of the silhouette values for each point in the dataset. When the overall Silhouette Value is found for a variety of numbers of clusters, the optimal number of clusters will be when the Silhouette Value is the highest representing the most well assigned elements.

We can show these Silhouette Values graphically using a Silhouette Plot. This plot organizes each silhouette value for a cluster from largest to smallest and plots them. In this plot, we are looking for the values of each cluster to be nonnegative and as large as possible. In Figure 4.9, the left side shows a Silhouette Plot of three clusters that is not optimal because there are negative values in the first and third cluster and first cluster having many values that are below 0.50. The right side shows cluster assignments in which there are no negative values and the average silhouette values for each cluster are maximized around 0.80.

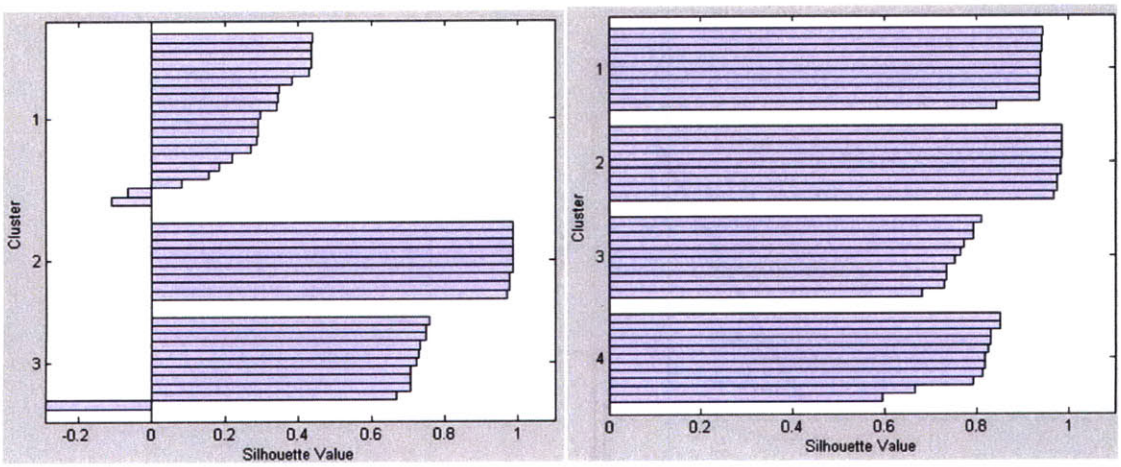


Figure 4.9 Example of Silhouette Value Plots

4.7 Determining Differences between Clusters

The second part of this chapter deals with finding cluster differences after the optimal number of clusters has been determined. We focus on the differences between the individual

attributes such as number of deaths and the types of boat used for landing, as well as provide a single number representing the ultimate difference between a given cluster and the total dataset. This allows us to see which clusters had attributes that were outliers and which seemed to follow the patterns of the data as a whole. We then do pairwise comparisons to assess which clusters were most similar to each other and which are the most different. We use the Cosine Similarity Method for assessing the differences:

$$\cos(\theta) = \frac{a \cdot b}{||a||||b||} \qquad (4.13)$$

$$a \cdot b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \quad (4.14)$$

where a is one vector of interest and b is the other vector of interest, n is the number of elements in each vector, and θ represents the angel between the two vectors that will be between 0 and 180 degrees. Vectors with a smaller angle between are more similar and those with a larger angle are less similar. Equation 4.13 show the dot product of vector a and vector b divided by the magnitude of each vector which equals the cosine of the angle between. Figure 4.10 shows a visual representation of this angle between vectors.

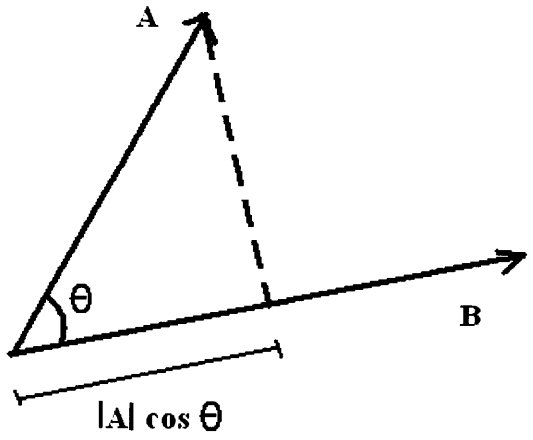


Figure 4.10 Angle Between Vectors

Feature vector representation for each cluster is comprised of the geographical location of that cluster centroid and associated with the cluster categorical information that includes the total departures, average number of passengers on each boat, average number of deaths on the boat, and types of boats used for the landings. Each of these vectors is characterized by the difference

when compared to the average of the entire set of landing data points. Similarly with the pairwise comparisons, we compare each cluster to every other cluster and create a dissimilarity matrix of computed angles to show the differences.

4.8 Summary of Results

In order to apply each of these methods for optimal clustering on the VAST dataset and accurately assess the results, we first applied each method to a benchmark dataset to see how the algorithms will perform under non-noisy and controlled conditions. The benchmark dataset consists of 40 data points: 10 from the Eastern side of Florida, 10 from the Western side of Florida, 10 from the Florida Keys, and 10 in Cancun, Mexico. This dataset, with cluster centers labeled below in Figure 4.11 consists of four very separate clusters and each algorithm should be able to accurately determine four to be the optimal number of clusters. Also, because of the nature of k means clustering, we run many iterations of each method to account for different random seeds for the initial cluster centers for both the benchmark dataset and the actual VAST Landing dataset: 500 repetitions for the benchmark set, looking at possible k^* values from 2 to 6 clusters and 1,000 repetitions for the VAST Landing set with k^* values ranging from 2 to 10 clusters.

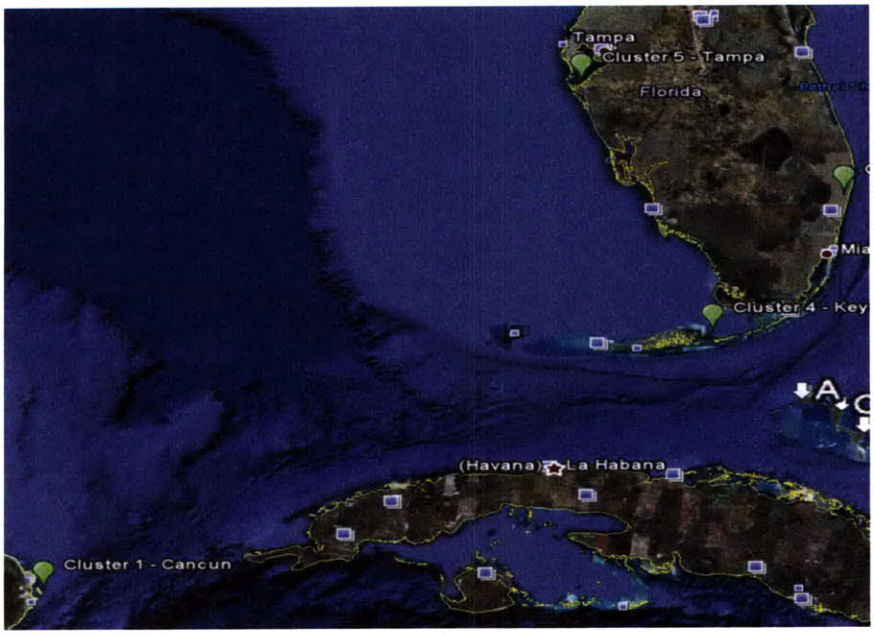


Figure 4.11 Benchmark Dataset Cluster Areas

4.8.1 Method 1: Gap Statistic

Looking first at the Gap Statistic, we use 10 uniform reference sets *(B)* for each iteration of this algorithm to compute the expected Gap Statistic and the standard deviation, or sd_kvalue. On the majority of the 500 trials with the benchmark dataset, the Gap Statistic finds optimal number of clusters to be 4. These results are what we would expect to see because the data is so well separated and the algorithm should not have trouble accurately determining the mathematically optimal number of clusters. The importance of having multiple repetitions or accurately choosing centroid centers is clearly seen in this example, because even with four well defined clusters, the algorithm still chooses 2, 3, and 5 as the optimal number of clusters in some cases (Figure 4.12)

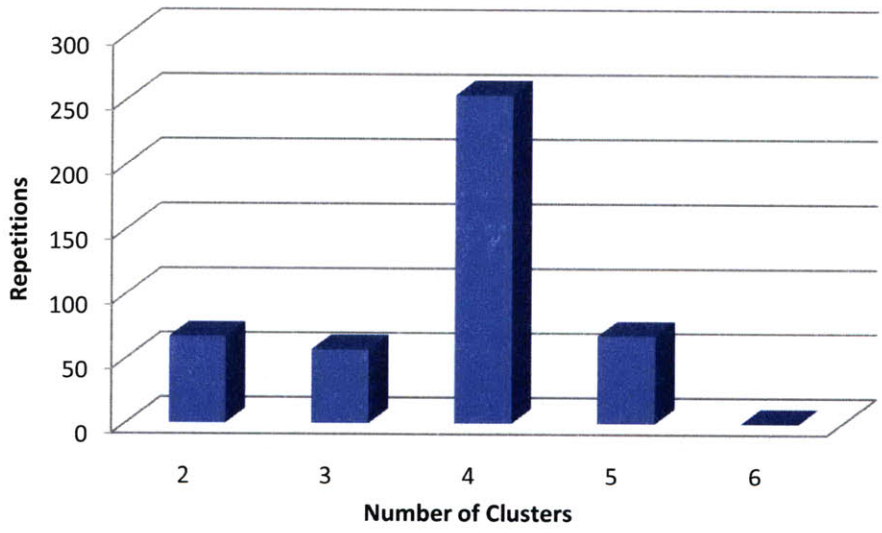


Figure 4.12 500 Repetitions of Gap Statistic on Benchmark Dataset

Applying the Gap Statistic to the full VAST Landing dataset, we find that out of the 1,000 repetitions, 991 times the algorithm computes the optimal number of clusters to be three. The other 9 repetitions find two to be the optimal number of clusters. When two clusters are determined to be optimal, the cluster centers are located in Cancun, Mexico and Central Florida. The Mexico cluster has all the landings centered in Cancun, while Central Florida has everything in Florida. When the optimal number of clusters is determined to be three by the algorithm, the cluster centers are: Cancun, the Florida Keys, and Northern Florida. Figure 4.13 shows a Gap Statistic Plot of the VAST Landing Data. The first time the smallest k^* is achieved satisfying

Equation 4.4, is at 3 clusters meaning a local maximum in the Gap Statistic is achieved. Other local maximums occur at 5 clusters and again at 7 clusters. This implies that based on this algorithm, the optimal number of clusters is three, but there are other less defined clusters within the data.

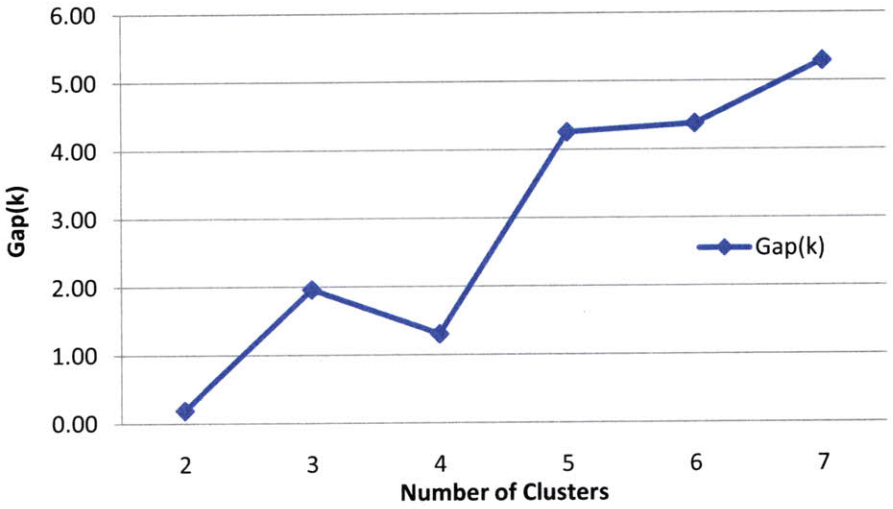


Figure 4.13 Example Gap Statistic Plot of the VAST Landing Data

4.8.2 Method 2: Validity Index

Next we explore the Validity Index as a method for mathematically computing the optimal number of cluster. Using the benchmark dataset first, we find four to be the optimal number of clusters for each of the 500 repetitions. Looking at Figure 4.14 below we see the red line representing the normalized ICMD *(vuN)*, the red line representing normalized MCID *(voN)*, and the green line representing the sum of the two values *(vSV)*. The optimal number of clusters for this method is represented when the *vSV* value is minimized, which is clearly when there are four clusters.

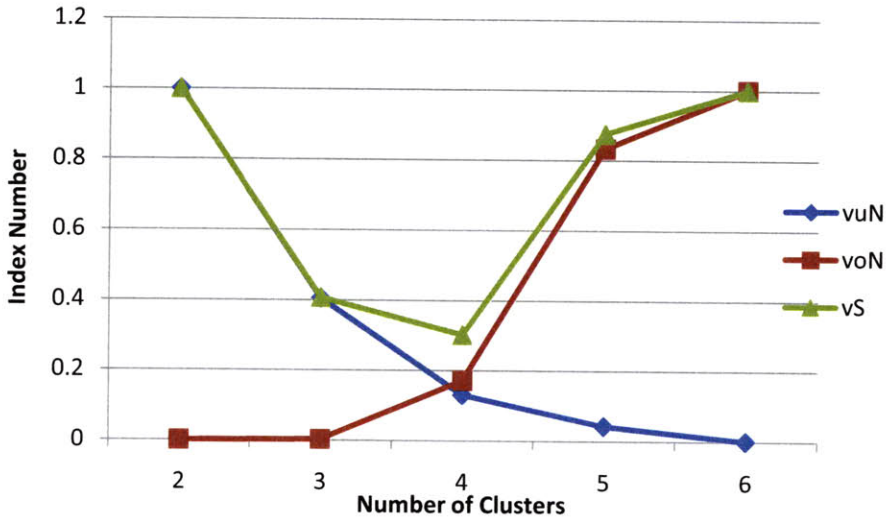


Figure 4.14 Validity Index Graph for Benchmark Data

When this method is applied to the full VAST Landing Dataset, the results are not as clear cut as they were with the benchmark dataset. The algorithm determines the optimal number of clusters to be 9 clusters the majority of the time, and the range of values determined to be the optimal number of clusters ranges from 5 clusters to 10 clusters, as seen in Figure 4.15. When we look closer at which data points are in each of the respective 9 clusters, we find that the algorithm simply breaks the state of Florida into many regions along the coastline, while keeping Cancun separate as its own cluster center.

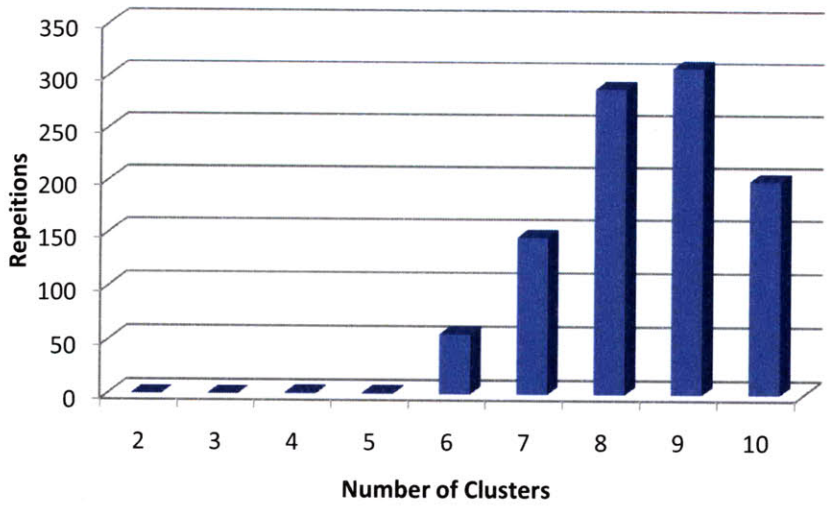


Figure 4.15 Validity Index for VAST Landing Dataset, 1000 Repetitions

4.8.3 Method 3: Silhouette Value

The next method we explore with both the benchmark dataset and the VAST Landing dataset is the silhouette value. For the 500 repetitions on the benchmark dataset, the Silhouette Value shows four as the optimal number of clusters nearly 75% of the time and shows five as the optimal number of clusters 17% of the time. Figure 4.16 shows a Silhouette Plot for the benchmark dataset for 3, 4, and 5 clusters. It is clear that in this case, four is the optimal number of clusters.

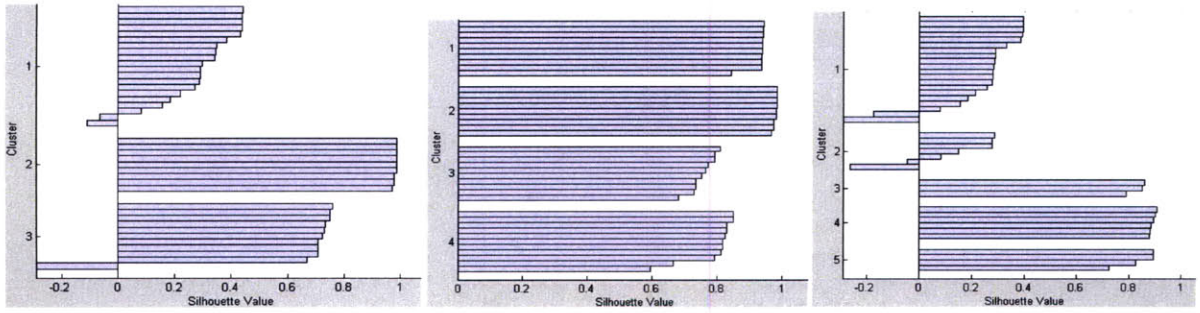


Figure 4.16 Silhouette Plots for Benchmark Dataset, when k=3, k=4, k=5 clusters

When applied to the full dataset, the method essentially separates the data into Cancun data points and Florida data points and returns the optimal number of clusters to be two, 100% of the time. When there are more than two clusters, the Silhouette Values are around the 0.65 to 0.70 range (highlighted in red), but because we are looking for the maximum Silhouette Value to determine the best number of clusters, none of these other cluster numbers are ever high enough to be considered optimal (Table 4.2).

Number of Clusters								
2	3	4	5	6	7	8	9	10
0.8388	0.5634	0.6883	0.3815	0.7379	0.6735	0.6762	0.5998	0.5537
0.8388	0.795	0.455	0.6667	0.6649	0.5836	0.6892	0.6674	0.5727
0.8388	0.7292	0.6883	0.4249	0.6323	0.5576	0.6397	0.4608	0.6417
0.8388	0.7292	0.6883	0.6132	0.7275	0.5949	0.6657	0.5569	0.745

Table 4.2 Silhouette Values for VAST Landing Dataset

4.8.4 Cluster Differences

To show the differences between clusters for the VAST Landing dataset, we elect to focus on five specific clusters. The cluster centers break the data into regions that consist of Cancun, Mexico, Florida Keys, Eastern Florida, Northwest Florida, and Southwest Florida. These regions and their centroids can be seen in Figure 4.17. We then compute all of the valuable information for each of these clusters including number of landings per year, average number of passengers per boat per year, average deaths per boat per year, and the type of boat used for the landing, which can be seen in Appendix G using the Cosine Similarity Method described above. Each of the values from each of the five clusters are normalized to be between zero and one and then compared to the entire dataset.

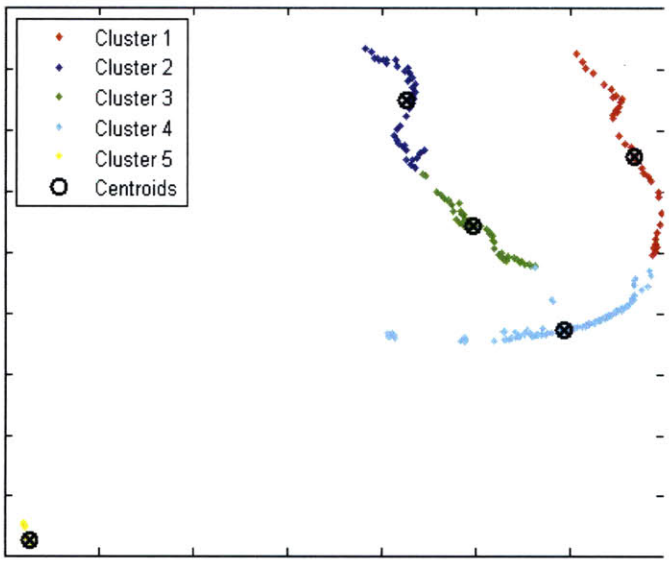


Figure 4.17 Cluster Centroids and Data Points for VAST Landing Data

The results show that when looking at overall differences, Cluster 4 (Florida Keys) is the most similar to the dataset as a whole, while Cluster 1 (Eastern Florida) and Cluster 5 (Cancun) are the most different (Table 4.3). These results follow our intuition about the temporal characteristics of the dataset because for the entire 3 year period boats are consistently landing in the Florida Keys, similar to the overall distribution, but in 2007 boats venture most frequently to Cancun and up the western Florida coastline. The total difference (bolded in "Total Diff" column in Table 4.3), represents the aggregate of the differences from the individual variables to

give a single number in which each cluster can be compared to determine the ones that are most similar and most different to the dataset as a whole.

We can also look at each of the variables separately and find which clusters are most similar to the dataset for a specific attribute. For example, Cluster 5 (Cancun) followed the dataset most closely for type of vessel used for the landings, while Cluster 2 (Northwest Florida) was the most different in that field (Table 4.3). With variables such as average number of passengers per year and average number of deaths per year, there is not a lot of variability between the clusters. In general, each cluster is similar to that of the average deaths and passengers for the overall set of data. The total differences and the break down between each variable can be seen in Table 4.3.

	Total Diff	Location	Year	No. Pass	No. Death	Vessel
Cluster 1 - Eastern Florida	**102.588**	2.938	26.311	37.680	28.041	7.618
Cluster 2 - Northwest Florida	**113.910**	2.998	11.536	32.235	54.044	13.097
Cluster 3 - Southwest Florida	**122.468**	1.854	22.066	32.420	59.560	6.568
Cluster 4 - Florida Keys	**77.348**	0.959	43.552	3.051	26.360	3.427
Cluster 5 - Cancun, Mexico	**103.579**	2.241	13.761	32.298	53.055	2.224
Most Different	**Cluster 3**	Cluster 2	Cluster 4	Cluster 1	Cluster 3	Cluster 2
Most Similar	**Cluster 4**	Cluster 4	Cluster 2	Cluster 4	Cluster 4	Cluster 5

Table 4.3 Cluster Differences Compared to Entire Dataset

Next we perform a pairwise comparison between clusters. A cluster compared to itself would have an angle difference of 0 because clearly there are no differences. The results of the comparison can be seen in Table 4.4 with yellow values representing the largest differences and blue values representing smallest differences. Cluster 1, for example is least similar to Cluster 3, but is quite similar to the other three clusters as a whole, while Cluster 3 is least similar to Cluster 4 and most similar to Cluster 5. Appendix H has the complete breakdown by individual variables for each of the pairwise comparisons.

		E Fla.	NW Fla.	SW Fla.	Keys	Cancun
	Cluster	1	2	3	4	5
E Fla.	1	0	149.0988	197.4104	148.8074	151.2978
NW Fla.	2	149.0988	0	119.1893	185.6811	26.91569
SW Fla.	3	197.4104	119.1893	0	136.8644	110.2238
Keys	4	148.8074	185.6811	136.8644	0	176.4463
Cancun	5	151.2978	26.91569	110.2238	176.4463	0

Table 4.4 Dissimilarity Matrix of Pairwise Comparisons of Clusters

4.9 Conclusions

4.9.1 Conclusions for the Optimal Number of Clusters

Overall, when considering the Gap Statistic, the Validity Index, and Silhouette Values as methods for determining the optimal number of clusters, we found that all methods work well for the benchmark dataset of clearly defined clusters with no outliers seen in Figure 4.11. The Validity Index proved to have the best results for the benchmark dataset with 100% accuracy of determining the optimal number of clusters to be four. The next best method was the Silhouette Value with 75% accurate finding the optimal number of clusters, and finally the Gap Statistic with 55%. In Table 4.5, the breakdown of each of the 500 representations can be seen.

	Number of Clusters				
	2	3	*4*	5	6
Gap Statistic	79	68	277	76	0
Validity Index	0	0	500	0	0
Silhouette Value	0	1	372	87	40

Table 4.5 Benchmark Dataset: Performance of Methods

When we instead compared the results of the VAST Landing dataset, which does not have clearly defined clusters and is more noisy, we do not see a lot of consistency among the methods. The Silhouette Value produced two clusters as optimal for all of the repetitions. The Validity Index showed the optimal number of clusters to be between 5 and 10, with 9 being the majority rule. And finally the Gap Statistic computed the best number of clusters to be three 99% of the time. These results can be seen in Table 4.6.

	Number of Clusters								
	2	3	4	5	6	7	8	9	10
Gap Statistic	9	991	0	0	0	0	0	0	0
Validity Index	0	0	1	1	56	146	288	308	200
Silhouette Value	1000	0	0	0	0	0	0	0	0

Table 4.6 VAST Landing Dataset: Performance of Methods

Despite the inability for these methods to perform consistently using the actual full dataset, the results on the benchmark dataset were promising and showed that these methods can actually produce a mathematically optimal number of clusters for clearly separable datasets.

4.9.2 Conclusions for Calculating Cluster Differences

We showed that we can accurately and simply compare clusters by looking at the angles between the vectors through the Cosine Similarity method for the elements that make up a specific cluster. We showed that we can compare these to the total dataset to see how similar a cluster is to the averages of the set, as well as how similar a specific cluster is to another cluster in the dataset. These results are helpful in the application of visual analytics and allow for a user to quickly and effectively see similarities and differences between smaller sets of data and determine if they want to explore specific pieces further.

4.10 Visual Analytics Applications for VAST Dataset

The results of this chapter were implemented into a visual analytics model to show each of the five clusters chosen as well as charts representing their differences on an individual attribute level as well as with an overall comparison. The visual tool allows for the user to see differences in a display and then determine which clusters they want to look into more closely. The first screen, a user is able to see is Figure 4.18, which shows all of the data points and their respective centroids. The centroids are labeled with red and black boxes and the circle inside the box represents how dissimilar a cluster is to the data as a whole.

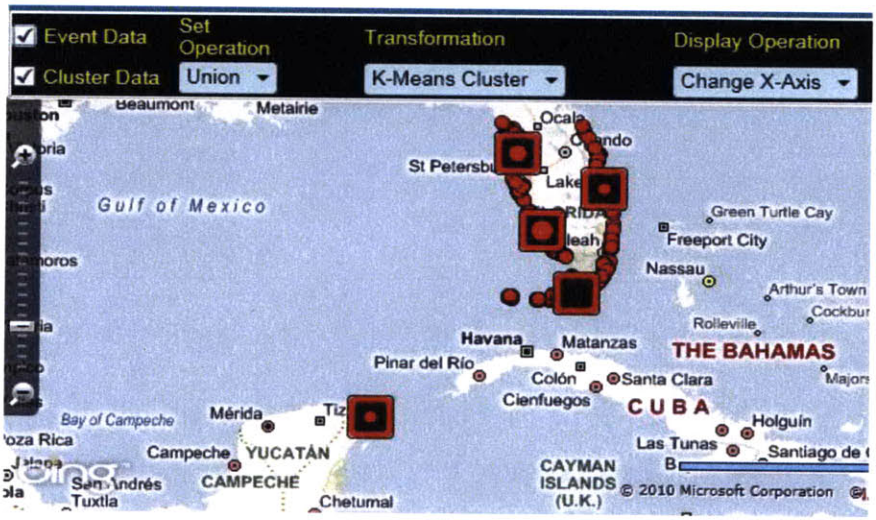


Figure 4.18 Visual Aid for viewing Cluster Centers

The cluster in the Florida Keys has no red circle which implies it is the most similar to the entire data set, while the cluster in Southwest Florida has the largest circle showing it is the most different when taking into account the cumulative differences for each of the variables in the cluster.

From there a user has the option of clicking on two separate clusters and exploring only the elements that make up those chosen clusters. Figure 4.19 shows the Florida Keys cluster and the Northwest Florida cluster, as well as graph of Year vs. Passenger. From the graph on the right hand side we can see the Keys cluster had passengers in all three years because the light blue dots span this time frame. Similarly we can see that the other cluster was predominately in 2007 and that year, generally had more passengers than the Keys cluster.

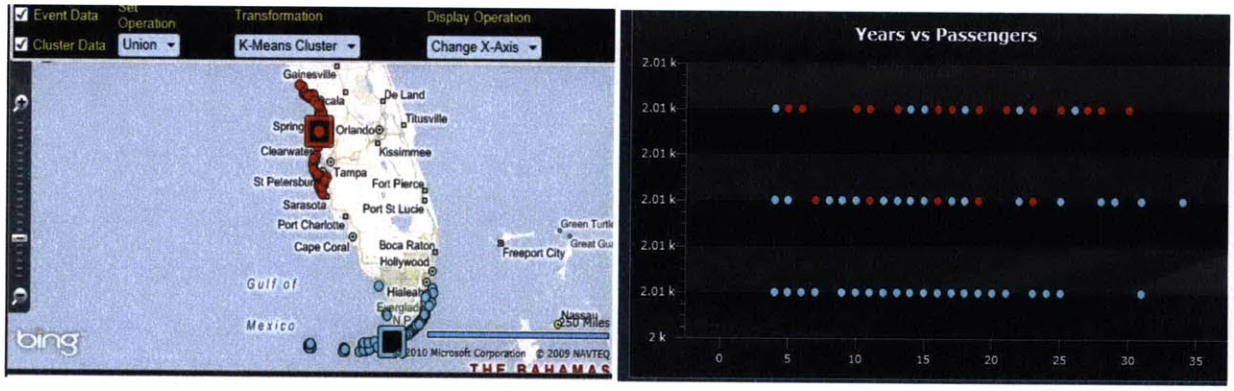


Figure 4.19 Visual Comparison of Two Chosen Clusters

Finally, to look at differences between clusters, a user is able to scroll over a specific cluster and determine how different clusters compare to the entire data set by their individual attribute as seen in Figure 4.20.

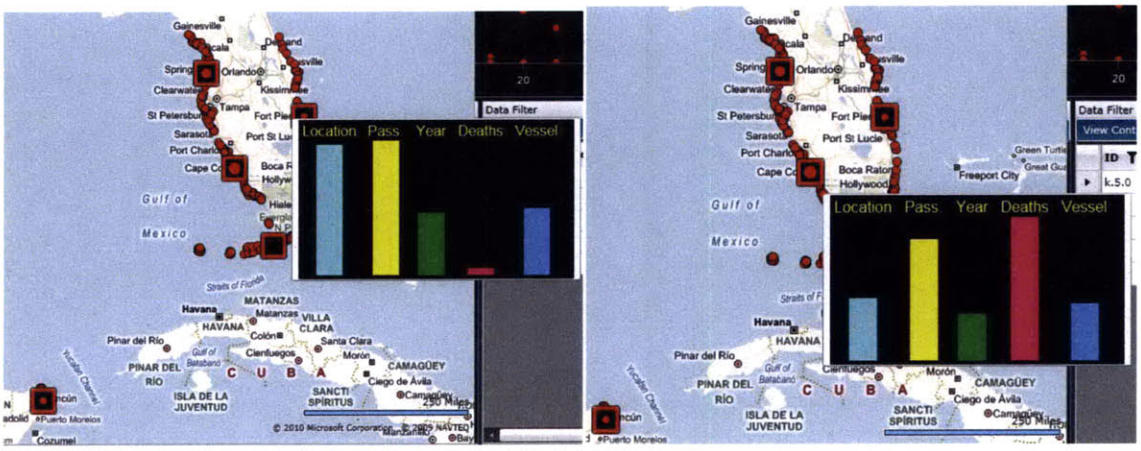


Figure 4.20 Comparisons of Individual Variables Between Two Chosen Clusters

Comparing these clusters, we can see that they are very different when looking at the average number of deaths per year, while they are pretty similar with respect to the types of vessels used and the average number of passengers per year. Overall, with the application of data mining techniques we take visual analytics to a new level. Rather than simple scatter plots of data points on a map, we can easily see a variety of interaction between clusters and study what characterizes the certain landing points represented by the clusters.

[This Page Intentionally Left Blank]

Chapter 5

Contributions and Future Work

5.1 Thesis Contributions

The goal of this thesis is to apply data mining to the medical and visual analytics fields. We show how supervised learning techniques can help predict a patient's length of stay in an Emergency Department upon minutes of his arrival while only knowing a few key elements about the patient and the condition of the ED. The best performing model was the Decision Tree that predicted length of stay with respect to a certain time range centered on the average. We have also shown the application of methods that automatically find the optimal number clusters in a dataset and computed differences between individual clusters.

This research makes the following contributions:

- Shows that Decision Tree Analysis is a viable method for predicting emergency department length of stay.

- Shows that different chief complaints correspond to different length of stays and are dependent on age, ED capacity, and the time of the patient's arrival.
- Demonstrates the use of mathematical algorithms to automatically find the optimal number of clusters for datasets.
- Demonstrates the use of angles to show the pairwise differences between separate clusters and to show how a single cluster differs from the dataset as a whole.
- Shows that there are many uses of data mining in a variety of fields.

5.2 Future Work

There are many opportunities for future work regarding this thesis. First, additional algorithms could be implemented with the Emergency Department data to compare and validate the results of the patient length of stays. Artificial Neural Networks have been used in ED length of stays and could be further explored to encompass more chief complaints and different sets of variables. Finally, there could be analysis into more specific breakdowns of chief complaints such as by anatomical locations. For example, instead of just looking at Fracture as a single chief complaint, it could be broken down by its qualifiers so that a leg fracture, a rib fracture and an arm fracture could all be explored individually.

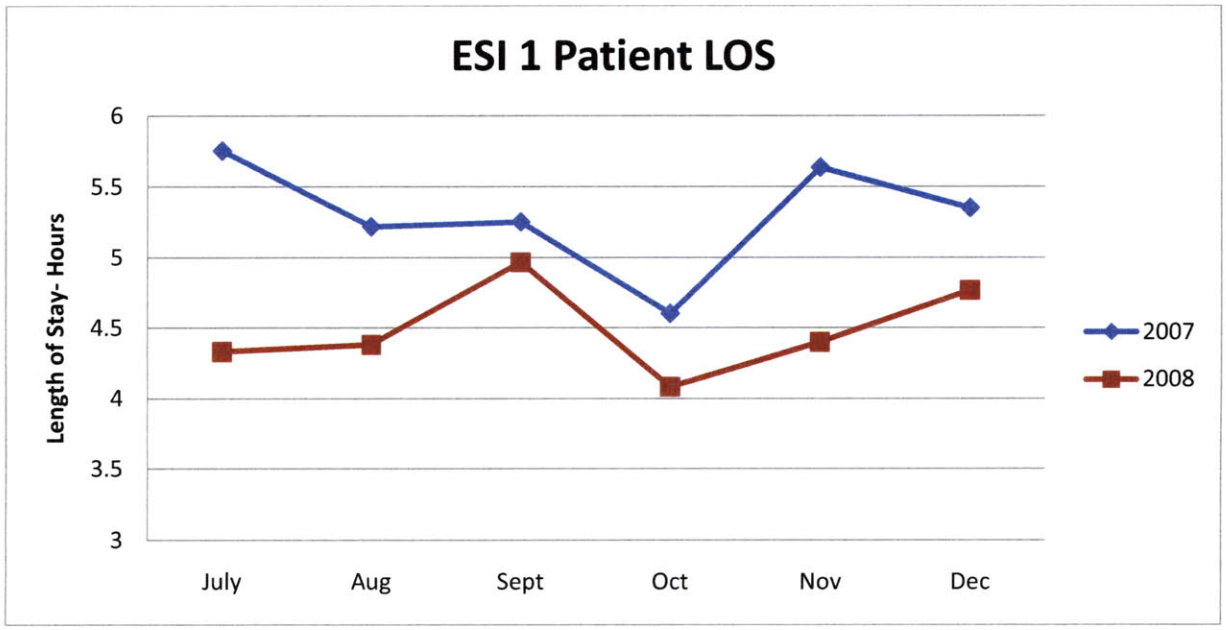
When using data mining as the underlying analysis for visual analytics, there could be implementation of a variety of clustering algorithms to find the optimal number of clusters such as density based clustering, hierarchical clustering, or spectral clustering. These clustering algorithms may provide complementary information to a user and allow him to explore the data more thoroughly. There could also be expansions on current algorithms to make them more robust for use with noisy datasets or datasets with highly indistinguishable clusters. This would show the versatility of a method and allow for higher applicability in the data mining field.

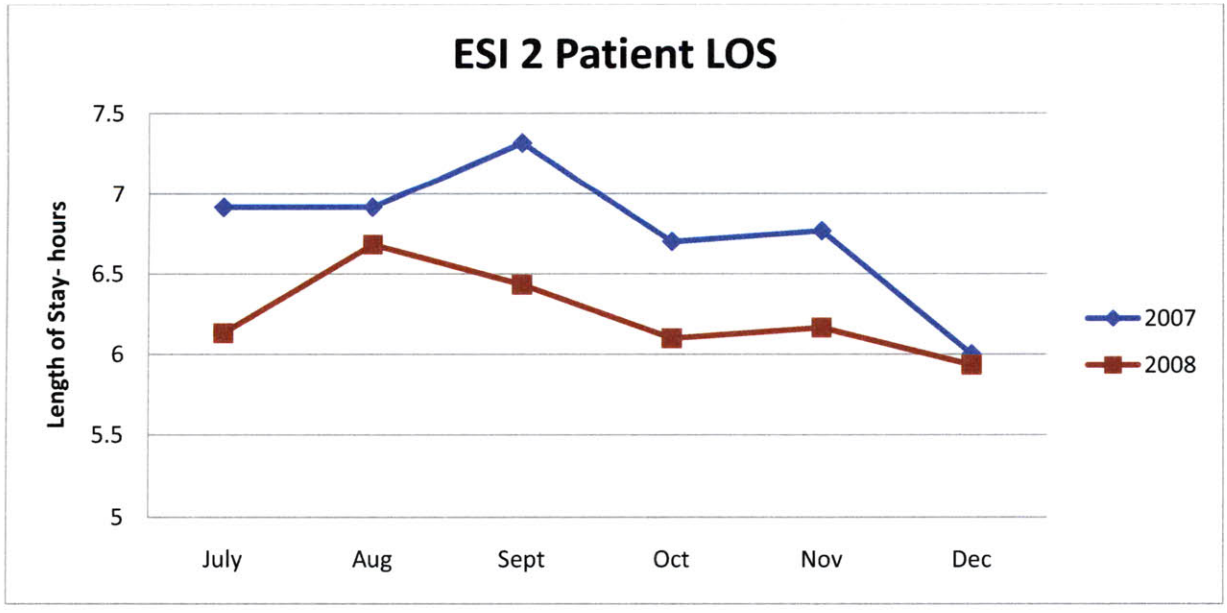
Appendix A – Glossary of Acronyms

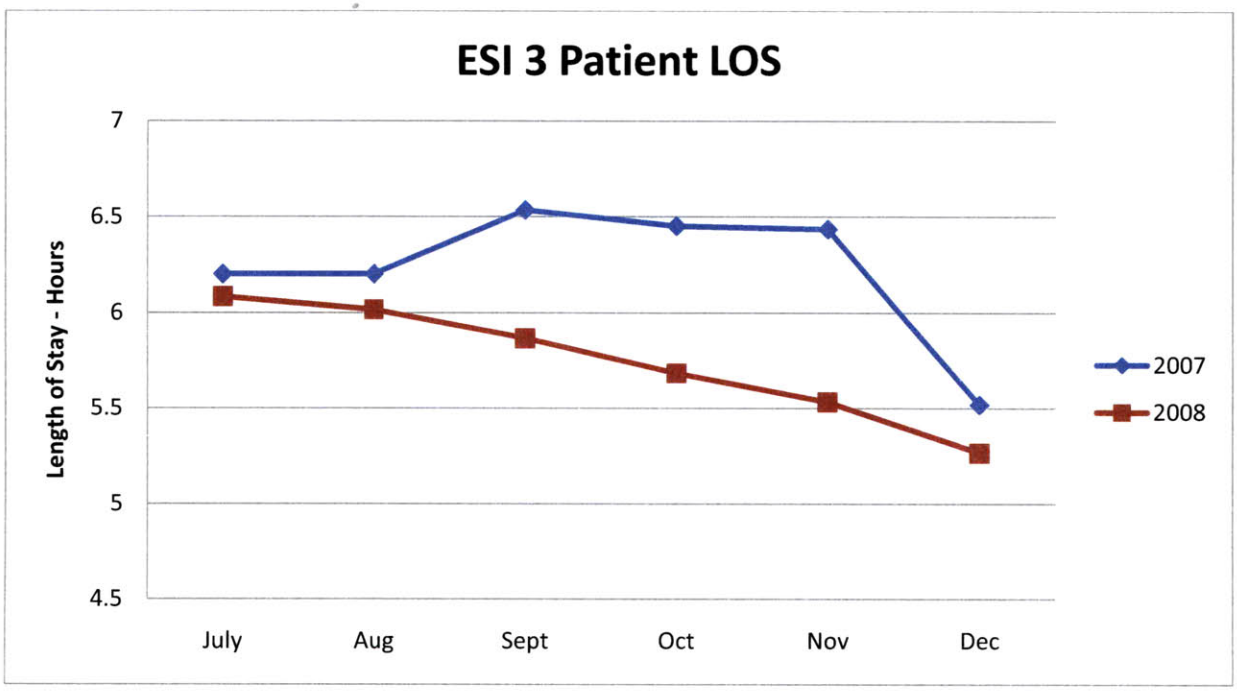
ANN	Artificial Neural Networks
BIDMC	Beth Israel Deaconess Medical Center
C	Capacity
CAT	Computed Tomography
ED	Emergency Department
ESI	Emergency Severity Index
FS	False Short
FL	False Long
IC	Initial Capacity
ICD-9	International Classification of Diseases, 9th Edition
ICMD	Inter-Cluster Minimum Distance
ICU	Intensive Care Unit
IEEE	Institute of Electronic and Electronics Engineers
kNN	*k* Nearest Neighbors
LF	Low Fill
LOS	Length of Stay
MA	Middle Aged
MICD	Mean Inter-Cluster Distance
MRI	Magnetic Resonance Imaging
NEVAC	North East Visual Analytics Center
O	Old
OC	Over Capacity
PTCA	Percuntaneous Transluminal Coronary Angioplasty
SVM	Support Vector Machines
TOD	Time of Day
TS	True Short
TL	True Long
USCG	United States Coast Guard

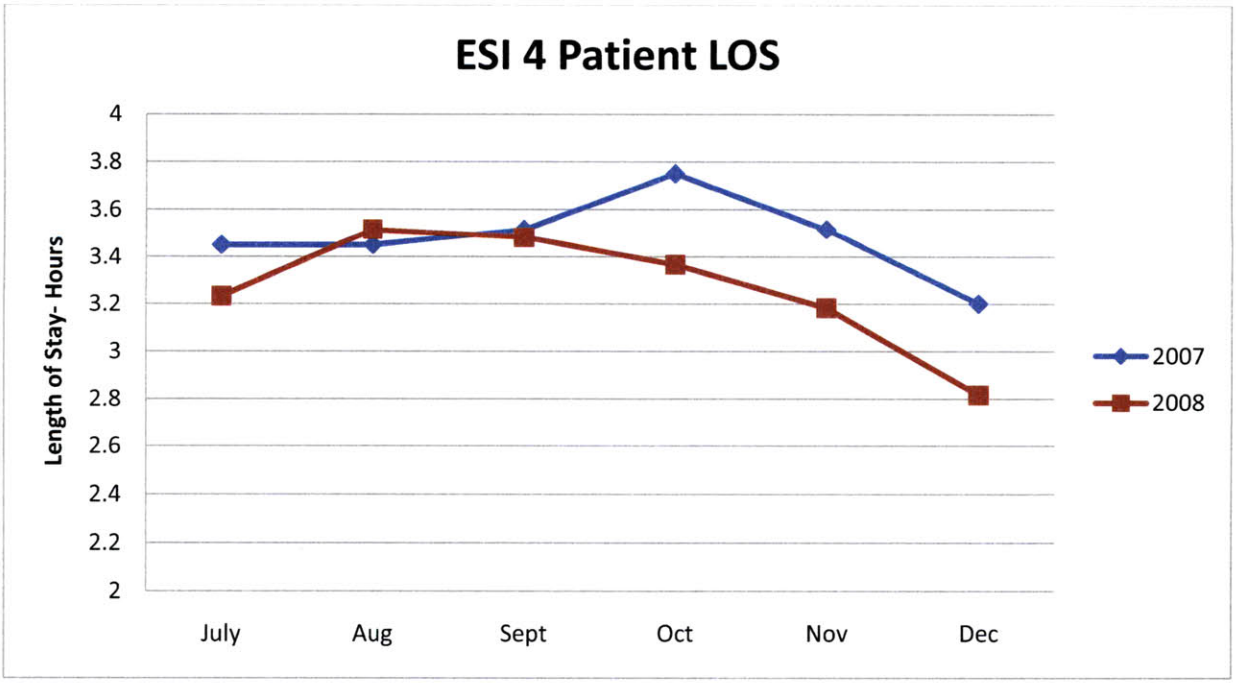
VAST	Visual Analytics Science and Technology
WR	Waiting Room
Y	Young

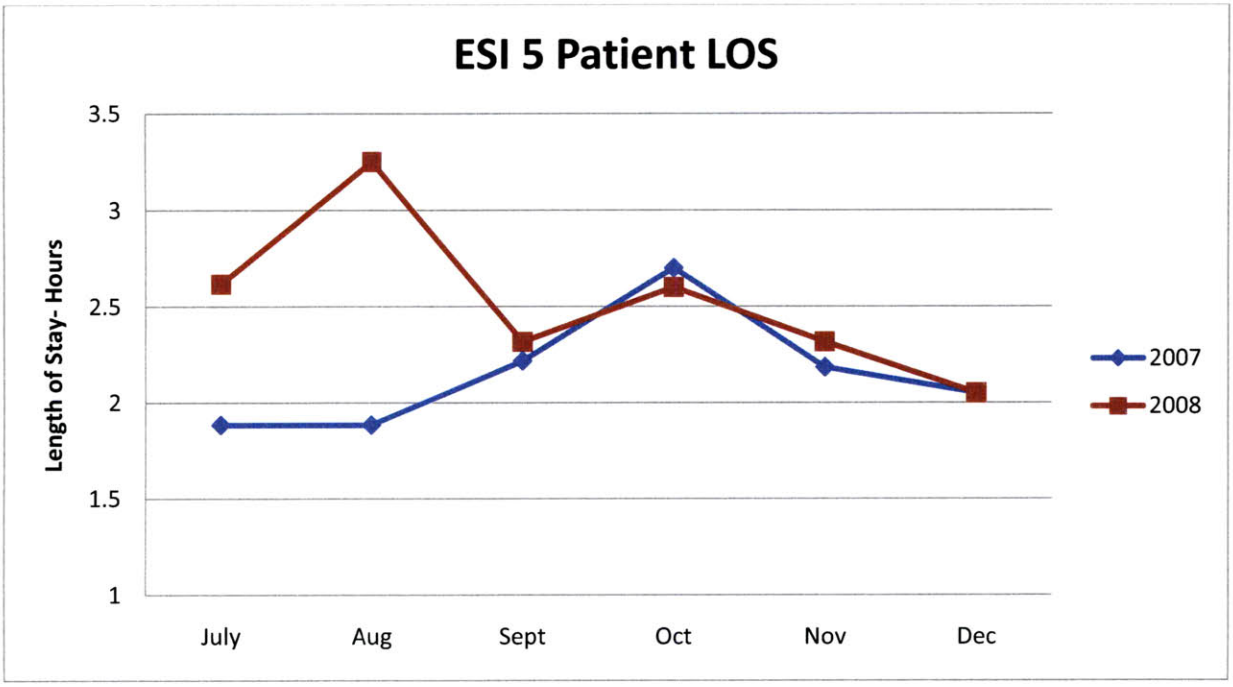
Appendix B – LOS Comparisons for 2007 and 2008 by ESI






ESI 3 Patient LOS
Length of Stay - Hours
7
6.5
6
5.5
5
4.5
July
Aug
Sept
Oct
Nov
Dec
2007
2008


ESI 4 Patient LOS
Length of Stay- Hours
4
3.8
3.6
3.4
3.2
3
2.8
2.6
2.4
2.2
2
July
Aug
Sept
Oct
Nov
Dec
2007
2008


ESI 5 Patient LOS
Length of Stay- Hours
3.5
3
2.5
2
1.5
1
July
Aug
Sept
Oct
Nov
Dec
2007
2008

[This Page Intentionally Left Blank]

Appendix C – Long/Short Breakdown by Chief Complaint

Dermatitis		
	Long	>3
	Short	<=3
	K=	5

Abdominal Pain		
	Long	>6
	Short	<=6
	K=	5

Abscess		
	Long	>5
	Short	<=5
	K=	5

Alcohol Intoxication		
	Long	>8
	Short	<=8
	K=	5

Bleeding		
	Long	>4
	Short	<=4
	K=	5

Cellulitis		
	Long	>5
	Short	<=5
	K=	5

Change of Mental Status		
	Long	>6
	Short	<=6
	K=	5

Fall		
	Long	>5
	Short	<=5
	K=	5

Fever		
	Long	>5
	Short	<=5
	K=	5

Fracture		
	Long	>5
	Short	<=5
	K=	5

Infection		
	Long	>4
	Short	<=4
	K=	5

Laceration		
	Long	>3
	Short	<=3
	K=	5

MVA		
	Long	>3
	Short	<=3
	K=	5

Nausea		
	Long	>6
	Short	<=6
	K=	5

Cough		
	Long	>4
	Short	<=4
	K=	5

Diarrhea		
	Long	>6
	Short	<=6
	K=	5

Dizziness		
	Long	>6
	Short	<=6
	K=	5

Evaluation		
	Long	>5
	Short	<=5
	K=	5

Back Pain		
	Long	>5
	Short	<=5
	K=	5

Chest Pain		
	Long	>8
	Short	<=8
	K=	5

Flank Pain		
	Long	>5
	Short	<=5
	K=	5

Head Pain		
	Long	>6
	Short	<=6
	K=	5

Overdose		
	Long	>8
	Short	<=8
	K=	5

Paresthesia		
	Long	>5
	Short	<=5
	K=	5

Seizure		
	Long	>6
	Short	<=6
	K=	5

Sore Throat		
	Long	>3
	Short	<=3
	K=	5

Stroke		
	Long	>4
	Short	<=4
	K=	5

Suicide		
	Long	>10
	Short	<=10
	K=	5

Swelling		
	Long	>5
	Short	<=5
	K=	5

Syncope		
	Long	>4
	Short	<=4
	K=	5

Dyspnea		
	Long	>5
	Short	<=5
	K=	5

Trauma		
	Long	>3
	Short	<=3
	K=	5

Assault		
	Long	>4
	Short	<=4
	K=	5

Wound		
	Long	>3
	Short	<=3
	K=	5

[This Page Intentionally Left Blank]

Appendix D – kNN Individual Chief Complaint Results on the Test Set

Name	Accuracy	Sensitivity	Specificity
Abdominal Pain	55.699%	65.544%	45.855%
Abscess	65.615%	66.000%	62.264%
Alcohol Intoxication	64.728%	66.667%	62.791%
Assault	57.70%	65.753%	52.055%
Back Pain	58.103%	63.793%	52.414%
Bleeding	47.890%	52.743%	43.038%
Cellulitis	56.480%	70.370%	42.593%
Change of Mental Status	52.500%	61.000%	44.000%
Chest Pain	60.758%	63.814%	57.702%
Cough	61.650%	71.845%	51.456%
Dermatitis	67.284%	69.136%	65.432%
Diarrhea	56.838%	67.521%	46.154%
Dizziness	47.241%	52.414%	53.103%
Dyspnea	52.690%	49.051%	56.329%
Evaluation	54.250%	52.000%	56.500%
Fall	53.985%	50.129%	57.841%
Fever	60.387%	57.971%	62.802%
Flank Pain	62.602%	61.789%	63.415%
Fracture	51.020%	51.020%	51.020%
Head Pain	50.000%	55.851%	44.149%
Infection	55.263%	58.480%	49.171%
Laceration	59.848%	65.152%	54.545%
Motor Vehicle Accident	57.346%	62.567%	43.725%
Nausea	56.731%	55.769%	57.692%
Overdose	54.412%	50.000%	58.824%
Paresthesia	41.250%	55.000%	27.500%
Seizure	48.276%	44.828%	53.333%
Sore Throat	49.231%	58.462%	40.000%
Stroke	48.889%	53.333%	44.444%
Suicidal Ideation	49.565%	52.174%	46.957%
Swelling	47.551%	35.714%	59.184%
Syncope	51.571%	52.941%	50.000%
Trauma	57.585%	54.180%	60.991%
Wound	59.804%	71.569%	48.039%

Dermatitis

Balanced				TS	56	Sens	69.14%
Accuracy	67.28			FS	28	Spec	65.43%
Matrix	56	25	Short	TL	53		
	28	53	Long	FL	25		

Abdominal Pain

Balanced				TS	506	Sens	65.54%
Accuracy	55.7			FS	418	Spec	45.85%
Matrix	506	266	Short	TL	354		
	418	354	Long	FL	266		

Abscess

Balanced				TS	33	Sens	66.00%
Accuracy	65.62			FS	20	Spec	62.26%
Matrix	33	17	Short	TL	33		
	20	33	Long	FL	17		

Alcohol Intoxication

Balanced				TS	86	Sens	66.67%
Accuracy	64.73			FS	48	Spec	62.79%
Matrix	86	43	Short	TL	81		
	48	81	Long	FL	43		

Bleeding

Balanced				TS	125	Sens	52.74%
Accuracy	47.89			FS	135	Spec	43.04%
Matrix	125	112	Short	TL	102		
	135	102	Long	FL	112		

Cellulitis

Balanced				TS	38	Sens	70.37%
Accuracy	56.48			FS	31	Spec	42.59%
Matrix	38	16	Short	TL	23		
	31	23	Long	FL	16		

Change of Mental Status

Balanced				TS	61	Sens	61.00%
Accuracy	52.5			FS	56	Spec	44.00%
Matrix	61	39	Short	TL	44		
	56	44	Long	FL	39		

Cough

Balanced				TS	74	Sens	71.84%
Accuracy	61.65			FS	50	Spec	51.46%
Matrix	74	29	Short	TL	53		
	50	53	Long	FL	29		

Diarrhea

Balanced				TS	79	Sens	67.52%
Accuracy	56.84			FS	63	Spec	46.15%
Matrix	79	38	Short	TL	54		
	63	54	Long	FL	38		

Dizziness

Balanced				TS	76	Sens	52.41%
Accuracy	47.24			FS	68	Spec	53.10%
Matrix	76	69	Short	TL	77		
	68	77	Long	FL	69		

Evaluation

Balanced				TS	104	Sens	52.00%
Accuracy	54.25			FS	87	Spec	56.50%
Matrix	104	96	Short	TL	113		
	87	113	Long	FL	96		

Fall

Balanced				TS	195	Sens	50.13%
Accuracy	53.98			FS	164	Spec	57.84%
Matrix	195	194	Short	TL	225		
	164	225	Long	FL	194		

Fever

Balanced				TS	120	Sens	57.97%
Accuracy	60.39			FS	77	Spec	62.80%
Matrix	120	87	Short	TL	130		
	77	130	Long	FL	87		

Fracture

Balanced				TS	25	Sens	51.02%
Accuracy	51.02			FS	24	Spec	51.02%
Matrix	25	24	Short	TL	25		
	24	25	Long	FL	24		

Infection

Balanced				TS	100	Sens	58.48%
Accuracy	55.26			FS	92	Spec	49.17%
Matrix	100	71	Short	TL	89		
	92	89	Long	FL	71		

Laceration

Balanced				TS	86	Sens	65.15%
Accuracy	59.85			FS	60	Spec	54.55%
Matrix	86	46	Short	TL	72		
	60	72	Long	FL	46		

MVA

Balanced				TS	160	Sens	65.84%
Accuracy	54.53			FS	138	Spec	43.21%
Matrix	160	83	Short	TL	105		
	138	105	Long	FL	83		

Nausea

Balanced				TS	58	Sens	55.77%
Accuracy	56.73			FS	44	Spec	57.69%
Matrix	58	46	Short	TL	60		
	44	60	Long	FL	46		

Overdose

Balanced				TS	17	Sens	50.00%
Accuracy	54.41			FS	14	Spec	58.82%
Matrix	17	17	Short	TL	20		
	14	20	Long	FL	17		

Paresthesia

Balanced				TS	22	Sens	55.00%
Accuracy	41.25			FS	29	Spec	27.50%
Matrix	22	18	Short	TL	11		
	29	11	Long	FL	18		

Seizure

Balanced				TS	39	Sens	44.83%
Accuracy	48.28			FS	42	Spec	53.33%
Matrix	39	48	Short	TL	48		
	42	48	Long	FL	48		

Sore Throat

Balanced				TS	38	Sens	58.46%
Accuracy	49.23			FS	39	Spec	40.00%
Matrix	38	27	Short	TL	26		
	39	26	Long	FL	27		

Stroke

Balanced				TS	24	Sens	53.33%
Accuracy	48.89			FS	25	Spec	44.44%
Matrix	24	21	Short	TL	20		
	25	20	Long	FL	21		

Suicide

Balanced				TS	60	Sens	52.17%
Accuracy	49.57			FS	61	Spec	46.96%
Matrix	60	55	Short	TL	54		
	61	54	Long	FL	55		

Swelling

Balanced				TS	35	Sens	35.71%
Accuracy	47.55			FS	40	Spec	59.18%
Matrix	35	63	Short	TL	58		
	40	58	Long	FL	63		

Syncope

Balanced				TS	54	Sens	52.94%
Accuracy	51.57			FS	51	Spec	50.00%
Matrix	54	48	Short	TL	51		
	51	51	Long	FL	48		

Trauma

Balanced				TS	175	Sens	54.18%
Accuracy	57.59			FS	126	Spec	60.99%
Matrix	175	148	Short	TL	197		
	126	197	Long	FL	148		

Wound

Balanced				TS	73	Sens	71.57%
Accuracy	59.8			FS	53	Spec	48.04%
Matrix	73	29	Short	TL	49		
	53	49	Long	FL	29		

Assault

Balanced Accuracy	57.7							
Matrix	48	25	Short	TS	48	Sens	65.75%	
	35	38	Long	FS	35	Spec	52.05%	
				TL	38			
				FL	25			

Back Pain

Balanced Accuracy	58.1							
Matrix	185	105	Short	TS	185	Sens	63.79%	
	138	152	Long	FS	138	Spec	52.41%	
				TL	152			
				FL	105			

Chest Pain

Balanced Accuracy	60.76							
Matrix	261	148	Short	TS	261	Sens	63.81%	
	173	236	Long	FS	173	Spec	57.70%	
				TL	236			
				FL	148			

Flank Pain

Balanced Accuracy	62.6							
Matrix	76	47	Short	TS	76	Sens	61.79%	
	45	78	Long	FS	45	Spec	63.41%	
				TL	78			
				FL	47			

Head Pain

Balanced Accuracy	50							
Matrix	105	83	Short	TS	105	Sens	55.85%	
	105	83	Long	FS	105	Spec	44.15%	
				TL	83			
				FL	83			

Dyspnea

Balanced Accuracy	52.69							
Matrix	155	161	Short	TS	155	Sens	49.05%	
	138	178	Long	FS	138	Spec	56.33%	
				TL	178			
				FL	161			

Appendix E – Logistic Regression Individual Chief Complaint Results

Name	Accuracy	Sensitivity	Specificity
Abdominal Pain	57.124%	61.4%	52.8%
Abscess	74.528%	71.7%	77.4%
Alcohol Intoxication	69.380%	66.7%	72.1%
Assault	58.14%	61.2%	56.2%
Back Pain	57.069%	64.1%	50.0%
Bleeding	56.751%	62.4%	51.1%
Cellulitis	68.519%	74.1%	63.0%
Change of Mental Status	61.500%	58.0%	65.0%
Chest Pain	60.758%	56.7%	64.8%
Cough	65.593%	69.9%	61.2%
Dermatitis	68.519%	66.7%	70.4%
Diarrhea	62.821%	69.2%	56.4%
Dizziness	59.655%	55.9%	63.2%
Dyspnea	56.962%	50.9%	63.3%
Evaluation	59.250%	48.0%	70.5%
Fall	54.884%	57.3%	52.4%
Fever	62.560%	63.3%	61.8%
Flank Pain	63.821%	66.7%	61.0%
Fracture	62.245%	51.0%	51.0%
Head Pain	56.649%	51.1%	62.2%
Infection	60.235%	61.4%	59.1%
Laceration	64.015%	62.9%	65.2%
Motor Vehicle Accident	60.700%	65.0%	56.4%
Nausea	65.865%	61.5%	70.2%
Overdose	63.253%	64.7%	61.8%
Paresthesia	61.250%	62.5%	60.0%
Seizure	55.172%	65.5%	44.8%
Sore Throat	59.231%	52.3%	66.2%
Stroke	64.440%	64.4%	64.4%
Suicidal Ideation	61.739%	68.7%	54.8%
Swelling	58.163%	53.1%	63.3%
Syncope	59.504%	56.9%	62.7%
Trauma	58.204%	58.2%	58.2%
Wound	67.647%	63.7%	71.6%

[This Page Intentionally Left Blank]

Appendix F – LOS Decision Tree Distributions on the Test Set

					Avg Range	1.875
					Avg Accuracy	80.0%
Fracture	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	NA	NA			0	0
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	4-5	75.0%	1	4	0.0234375	0.03125
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-5	66.0%	1	3	0.0154688	0.023438
1,O,IC	2-5	100.0%	3	3	0.0234375	0.070313
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	NA	NA			0	0
2,Y,IC	2-4	66.0%	2	3	0.0154688	0.046875
2,Y,C	2-4	100.0%	2	3	0.0234375	0.046875
2,Y,OC	NA	NA			0	0
2,MA,LF	4-5	72.7%	1	11	0.0624938	0.085938
2,MA,IC	4-7	100.0%	3	5	0.0390625	0.117188
2,MA,C	4-6	100.0%	2	3	0.0234375	0.046875
2,MA,OC	3-4	66.0%	1	4	0.020625	0.03125
2,O,LF	5-7	66.0%	2	12	0.061875	0.1875
2,O,IC	3-5	83.3%	2	6	0.0390609	0.09375
2,O,C	3-6	100.0%	3	8	0.0625	0.1875
2,O,OC	NA	NA			0	0
3,Y,LF	3-5	66.0%	2	3	0.0154688	0.046875
3,Y,IC	3-4	66.0%	1	3	0.0154688	0.023438
3,Y,C	4-5	100.0%	1	2	0.015625	0.015625
3,Y,OC	4-5	80.0%	1	5	0.03125	0.039063
3,MA,LF	4-5	66.0%	1	6	0.0309375	0.046875
3,MA,IC	5-7	75.0%	2	4	0.0234375	0.0625
3,MA,C	3-7	100.0%	4	5	0.0390625	0.15625
3,MA,OC	3-5	75.0%	2	4	0.0234375	0.0625
3,O,LF	4-6	71.4%	2	7	0.0390469	0.109375
3,O,IC	4-6	75.0%	2	8	0.046875	0.125
3,O,C	4-6	91.7%	2	12	0.0859378	0.1875
3,O,OC	6-7	75.0%	1	4	0.0234375	0.03125

					Avg Range	3.418033
					Avg Accuracy	80.4%
Abscess	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-3	100.0%	1	2	0.0163934	0.016393
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-3	75.0%	1	4	0.0245902	0.032787
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-5	75.0%	2	8	0.0491803	0.131148
2,Y,IC	4-5	77.8%	1	9	0.0573787	0.07377
2,Y,C	3-6	100.0%	3	2	0.0163934	0.04918
2,Y,OC	NA	NA			0	0
2,MA,LF	2-6	81.8%	4	22	0.1475262	0.721311
2,MA,IC	2-6	83.8%	4	12	0.0824557	0.393443
2,MA,C	5-9	83.8%	4	6	0.0412279	0.196721
2,MA,OC	4-8	100.0%	4	2	0.0163934	0.065574
2,O,LF	NA	NA			0	0
2,O,IC	6-9	100.0%	3	3	0.0245902	0.07377
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	2-5	80.0%	3	5	0.0327869	0.122951
3,Y,IC	2-9	77.8%	7	9	0.0573713	0.516393
3,Y,C	3-7	77.8%	4	9	0.0573713	0.295082
3,Y,OC	6-8	71.4%	2	7	0.040973	0.114754
3,MA,LF	2-4	100.0%	2	3	0.0245902	0.04918
3,MA,IC	5-10	75.0%	5	4	0.0245902	0.163934
3,MA,C	5-8	72.7%	3	11	0.0655672	0.270492
3,MA,OC	4-8	75.0%	4	4	0.0245902	0.131148
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	5.937355
					Avg Accuracy	79.5%
Alcohol Intoxication	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	4-9	82.9%	5	117	0.2250418	1.357309
1,Y,IC	4-8	80.0%	4	25	0.0464037	0.232019
1,Y,C	5-7	100.0%	2	3	0.0069606	0.013921
1,Y,OC	NA	NA			0	0
1,MA,LF	4-10	76.2%	6	42	0.0742455	0.584687
1,MA,IC	5-9	80.0%	4	7	0.012993	0.064965
1,MA,C	4-12	75.0%	8	4	0.0069606	0.074246
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	4-12	80.0%	8	5	0.0092807	0.092807
2,Y,IC	NA	NA			0	0
2,Y,C	4-7	75.0%	3	4	0.0069606	0.027842
2,Y,OC	NA	NA			0	0
2,MA,LF	4-11	78.1%	7	32	0.0579861	0.519722
2,MA,IC	2-11	76.5%	9	17	0.0301622	0.354988
2,MA,C	3-10	75.0%	7	12	0.0208817	0.194896
2,MA,OC	NA	NA			0	0
2,O,LF	10-12	100.0%	2	2	0.0046404	0.009281
2,O,IC	7-13	75.0%	6	4	0.0069606	0.055684
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	4-13	75.0%	9	15	0.0261021	0.313225
3,Y,IC	4-9	78.6%	5	14	0.0255216	0.162413
3,Y,C	3-7	90.0%	4	10	0.0208817	0.092807
3,Y,OC	3-6	100.0%	3	3	0.0069606	0.020882
3,MA,LF	4-11	75.0%	7	32	0.0556845	0.519722
3,MA,IC	4-13	77.8%	9	18	0.0324835	0.37587
3,MA,C	6-13	76.2%	7	42	0.0742455	0.682135
3,MA,OC	10-13	81.8%	3	11	0.0208796	0.076566
3,O,LF	3-6	75.0%	3	4	0.0069606	0.027842
3,O,IC	5-11	80.0%	6	5	0.0092807	0.069606
3,O,C	10-11	100.0%	2	3	0.0069606	0.013921
3,O,OC	NA	NA			0	0

					Avg Range	5.473404
					Avg Accuracy	80.4%
Assault	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	1-8	80.8%	7	52	0.2233787	1.93617
1,Y,IC	2-4	75.0%	2	47	0.1875	0.5
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-10	80.0%	7	15	0.0638298	0.558511
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-9	100.0%	7	3	0.0159574	0.111702
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-7	88.9%	5	9	0.0425489	0.239362
2,Y,IC	2-5	100.0%	3	4	0.0212766	0.06383
2,Y,C	2-9	75.0%	7	4	0.0159574	0.148936
2,Y,OC	NA	NA			0	0
2,MA,LF	3-6	100.0%	3	4	0.0212766	0.06383
2,MA,IC	2-4	100.0%	2	2	0.0106383	0.021277
2,MA,C	NA	NA			0	0
2,MA,OC	NA	NA			0	0
2,O,LF	NA	NA			0	0
2,O,IC	NA	NA			0	0
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	1-3	80.0%	2	5	0.0212766	0.053191
3,Y,IC	3-4	75.0%	1	4	0.0159574	0.021277
3,Y,C	2-8	75.0%	6	8	0.0319149	0.255319
3,Y,OC	7-9	100.0%	2	2	0.0106383	0.021277
3,MA,LF	2-9	78.6%	7	14	0.0585096	0.521277
3,MA,IC	6-18	75.0%	12	8	0.0319149	0.510638
3,MA,C	4-16	85.7%	12	7	0.0319133	0.446809
3,MA,OC	NA	NA			0	0
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	2.946309
					Avg Accuracy	80.1%
Cellulitis	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-6	77.7%	3	9	0.0469329	0.181208
1,Y,IC		NA			0	0
1,Y,C		NA			0	0
1,Y,OC		NA			0	0
1,MA,LF	2-5	83.3%	3	6	0.0335557	0.120805
1,MA,IC		NA			0	0
1,MA,C		NA			0	0
1,MA,OC		NA			0	0
1,O,LF		NA			0	0
1,O,IC		NA			0	0
1,O,C		NA			0	0
1,O,OC		NA			0	0
2,Y,LF	5-7	100.0%	2	3	0.0201342	0.040268
2,Y,IC	5-7	75.0%	2	4	0.0201342	0.053691
2,Y,C	5-7	100.0%	2	2	0.0134228	0.026846
2,Y,OC		NA			0	0
2,MA,LF	3-6	82.1%	3	28	0.154357	0.563758
2,MA,IC	5-8	71.4%	3	14	0.067106	0.281879
2,MA,C	4-8	80.0%	4	5	0.0268456	0.134228
2,MA,OC		NA			0	0
2,O,LF	3-5	88.9%	2	9	0.0536678	0.120805
2,O,IC	4-7	85.7%	3	7	0.0402664	0.14094
2,O,C	5-7	75.0%	2	8	0.0402685	0.107383
2,O,OC	5-7	75.0%	2	7	0.0352349	0.09396
3,Y,LF	3-8	71.4%	5	3	0.0143799	0.100671
3,Y,IC	2-5	83.3%	3	3	0.0167779	0.060403
3,Y,C	4-8	76.4%	3	3	0.0153866	0.060403
3,Y,OC	NA	NA			0	0
3,MA,LF	2-4	80.0%	2	5	0.0268456	0.067114
3,MA,IC	2-6	75.0%	4	8	0.0402685	0.214765
3,MA,C	3-7	75.0%	4	12	0.0604027	0.322148
3,MA,OC	4-8	100.0%	4	6	0.0402685	0.161074
3,O,LF		NA			0	0
3,O,IC		NA			0	0
3,O,C	5-6	75.0%	2	7	0.0352349	0.09396
3,O,OC		NA			0	0

					Avg Range	1.879781
					Avg Accuracy	80.4%
Dermatitis	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-3	81.8%	1	11	0.0491754	0.060109
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-3	78.6%	1	14	0.0601082	0.076503
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-3	100.0%	1	2	0.010929	0.010929
1,O,IC	3-4	100.0%	1	2	0.010929	0.010929
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-4	75.0%	2	28	0.1147541	0.306011
2,Y,IC	3-5	87.5%	2	15	0.0717213	0.163934
2,Y,C	2-4	100.0%	2	5	0.0273224	0.054645
2,Y,OC	3-4	100.0%	1	2	0.010929	0.010929
2,MA,LF	2-4	75.0%	1	20	0.0819672	0.10929
2,MA,IC	3-6	75.0%	3	11	0.045082	0.180328
2,MA,C	NA	NA			0	0
2,MA,OC	NA	NA			0	0
2,O,LF	3-6	81.8%	3	11	0.0491694	0.180328
2,O,IC	5-6	75.0%	1	4	0.0163934	0.021858
2,O,C	6-7	75.0%	1	5	0.0204918	0.027322
2,O,OC	NA	NA			0	0
3,Y,LF	2-3	100.0%	1	5	0.0273224	0.027322
3,Y,IC	2-3	100.0%	1	8	0.0437158	0.043716
3,Y,C	2-4	69.2%	2	13	0.0491798	0.142077
3,Y,OC	2-3	62.5%	1	8	0.0273224	0.043716
3,MA,LF	NA	NA			0	0
3,MA,IC	2-7	100.0%	5	9	0.0491803	0.245902
3,MA,C	2-5	70.0%	3	10	0.0382514	0.163934
3,MA,OC	NA	NA			0	0
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	4.018462
					Avg Accuracy	80.3%
Fall	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-6	78.6%	4	42	0.0338455	0.172308
1,Y,IC	2-8	88.8%	6	9	0.0081969	0.055385
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-6	76.9%	3	39	0.030768	0.12
1,MA,IC	3-7	85.7%	4	7	0.0061535	0.028718
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	3-7	77.1%	4	70	0.0553826	0.287179
1,O,IC	3-5	85.1%	2	14	0.0122252	0.028718
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-7	80.8%	4	26	0.0215384	0.106667
2,Y,IC	3-6	83.3%	3	12	0.010256	0.036923
2,Y,C	4-5	75.0%	2	4	0.0030769	0.008205
2,Y,OC	NA	NA			0	0
2,MA,LF	3-8	80.0%	5	90	0.0738462	0.461538
2,MA,IC	3-6	77.8%	3	45	0.0358985	0.138462
2,MA,C	3-6	81.8%	3	22	0.0184597	0.067692
2,MA,OC	NA	NA			0	0
2,O,LF	3-8	81.2%	5	153	0.1273745	0.784615
2,O,IC	3-8	80.0%	5	70	0.0574359	0.358974
2,O,C	4-7	87.9%	3	66	0.0594812	0.203077
2,O,OC	4-8	85.7%	4	7	0.0061535	0.028718
3,Y,LF	3-5	75.0%	2	12	0.0092308	0.024615
3,Y,IC	3-8	84.6%	5	13	0.0112813	0.066667
3,Y,C	3-4	77.7%	1	9	0.0071723	0.009231
3,Y,OC	3-5	75.0%	2	4	0.0030769	0.008205
3,MA,LF	3-7	78.3%	4	23	0.0184613	0.094359
3,MA,IC	3-7	78.3%	4	23	0.0184613	0.094359
3,MA,C	4-8	76.5%	4	34	0.0266665	0.139487
3,MA,OC	4-8	89.5%	4	19	0.0174352	0.077949
3,O,LF	4-7	82.9%	3	35	0.029741	0.107692
3,O,IC	3-7	76.9%	3	39	0.030768	0.12
3,O,C	3-7	77.1%	4	61	0.0482394	0.250256
3,O,OC	4-9	81.5%	5	27	0.0225637	0.138462

					Avg Range	3.182371
					Avg Accuracy	79.6%
Infection	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-6	84.6%	4	12	0.0308608	0.145897
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-7	88.8%	4	18	0.0485836	0.218845
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-6	81.8%	4	11	0.0273529	0.133739
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-4	75.0%	2	20	0.0455927	0.121581
2,Y,IC	2-5	81.3%	3	16	0.0395137	0.145897
2,Y,C	2-6	88.9%	4	9	0.0243164	0.109422
2,Y,OC	5-6	75.0%	1	5	0.0113982	0.015198
2,MA,LF	3-6	77.8%	3	45	0.1063723	0.410334
2,MA,IC	3-7	76.5%	4	17	0.0395134	0.206687
2,MA,C	NA	NA			0	0
2,MA,OC	NA	NA			0	0
2,O,LF	3-6	75.0%	3	27	0.0615502	0.246201
2,O,IC	4-7	75.0%	3	16	0.0364742	0.145897
2,O,C	5-7	100.0%	2	8	0.0243161	0.048632
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	75.0%	2	8	0.0182371	0.048632
3,Y,IC	3-6	76.6%	3	8	0.0186188	0.072948
3,Y,C	3-8	78.6%	5	14	0.033434	0.212766
3,Y,OC	2-6	75.0%	4	4	0.0091185	0.048632
3,MA,LF	2-5	78.6%	3	14	0.033434	0.12766
3,MA,IC	2-6	85.7%	4	7	0.0182362	0.085106
3,MA,C	3-6	75.0%	3	33	0.075228	0.300912
3,MA,OC	4-7	100.0%	3	6	0.0182371	0.054711
3,O,LF	3-6	87.5%	3	8	0.0212766	0.072948
3,O,IC	3-6	81.8%	3	11	0.0273529	0.100304
3,O,C	4-7	75.0%	3	12	0.0273556	0.109422
3,O,OC	NA	NA			0	0

					Avg Range	1.668464
					Avg Accuracy	80.4%
Laceration	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-3	77.5%	1	40	0.083558	0.107817
1,Y,IC	2-3	80.0%	1	5	0.0107817	0.013477
1,Y,C	2-3	83.3%	1	6	0.0134717	0.016173
1,Y,OC	NA	NA			0	0
1,MA,LF	2-3	76.2%	1	21	0.0431264	0.056604
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-4	87.5%	2	40	0.0943396	0.215633
2,Y,IC	2-4	77.8%	2	18	0.0377466	0.097035
2,Y,C	2-5	83.3%	3	7	0.015717	0.056604
2,Y,OC	3-4	100.0%	1	3	0.0080863	0.008086
2,MA,LF	2-4	77.4%	2	31	0.0646822	0.167116
2,MA,IC	2-4	76.2%	2	21	0.0431264	0.113208
2,MA,C	2-4	80.0%	2	10	0.0215633	0.053908
2,MA,OC	NA	NA			0	0
2,O,LF	2-4	90.0%	2	10	0.0242588	0.053908
2,O,IC	3-4	75.0%	1	4	0.0080863	0.010782
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	88.2%	2	17	0.0404288	0.091644
3,Y,IC	2-4	90.3%	2	31	0.0754695	0.167116
3,Y,C	2-3	71.0%	1	39	0.0745941	0.105121
3,Y,OC	2-4	87.5%	2	16	0.0377358	0.086253
3,MA,LF	2-4	87.5%	2	8	0.0188679	0.043127
3,MA,IC	2-3	75.0%	1	12	0.0242588	0.032345
3,MA,C	3-5	73.3%	2	15	0.0296482	0.080863
3,MA,OC	2-4	75.0%	2	8	0.0161725	0.043127
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	2-4	75.0%	2	9	0.0181941	0.048518
3,O,OC	NA	NA			0	0

					Avg Range	3.025
					Avg Accuracy	78.9%
Motor Vehicle Acciddent	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-7	76.9%	4	52	0.0833083	0.433333
1,Y,IC	2-7	75.0%	5	8	0.0125	0.083333
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-5	79.4%	3	34	0.0562488	0.2125
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-5	78.7%	3	47	0.07708	0.29375
2,Y,IC	2-4	78.3%	2	23	0.0374996	0.095833
2,Y,C	3-4	77.7%	2	9	0.0145688	0.0375
2,Y,OC	NA	NA			0	0
2,MA,LF	2-5	71.7%	2	60	0.0895875	0.25
2,MA,IC	3-7	79.3%	4	29	0.0479165	0.241667
2,MA,C	4-6	81.8%	3	11	0.0187481	0.06875
2,MA,OC	4-5	75.0%	1	4	0.00625	0.008333
2,O,LF	2-5	90.9%	3	11	0.0208313	0.06875
2,O,IC	2-6	75.0%	4	8	0.0125	0.066667
2,O,C	2-6	100.0%	4	7	0.0145833	0.058333
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	90.9%	2	22	0.0416625	0.091667
3,Y,IC	2-5	77.3%	3	22	0.0354154	0.1375
3,Y,C	2-6	77.3%	4	22	0.0354154	0.183333
3,Y,OC	3-6	80.0%	3	10	0.0166667	0.0625
3,MA,LF	3-6	76.0%	3	25	0.0395833	0.15625
3,MA,IC	3-6	94.0%	3	16	0.0313167	0.1
3,MA,C	3-6	80.8%	3	27	0.0454275	0.16875
3,MA,OC	3-6	75.0%	3	12	0.01875	0.075
3,O,LF	2-5	80.0%	3	5	0.0083333	0.03125
3,O,IC	3-6	80.0%	3	5	0.0083333	0.03125
3,O,C	2-5	72.7%	3	11	0.016665	0.06875
3,O,OC	NA	NA			0	0
						0

					Avg Range	7.333333
					Avg Accuracy	81.4%
Overdose	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-10	83.3%	7	6	0.0617259	0.518519
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-10	71.4%	7	12	0.1058074	1.037037
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-12	80.0%	9	5	0.0493827	0.555556
2,Y,IC	3-21	100.0%	18	7	0.0864198	1.555556
2,Y,C	3-8	80.0%	5	3	0.0296296	0.185185
2,Y,OC	NA	NA			0	0
2,MA,LF	3-9	77.8%	6	9	0.0864111	0.666667
2,MA,IC	8-14	100.0%	6	3	0.037037	0.222222
2,MA,C	NA	NA			0	0
2,MA,OC	NA	NA			0	0
2,O,LF	5-7	100.0%	2	3	0.037037	0.074074
2,O,IC	NA	NA			0	0
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	3-8	80.0%	5	6	0.0592593	0.37037
3,Y,IC	3-8	80.0%	5	4	0.0395062	0.246914
3,Y,C	4-13	75.0%	9	8	0.0740741	0.888889
3,Y,OC	NA	NA			0	0
3,MA,LF	3-5	100.0%	2	3	0.037037	0.074074
3,MA,IC	3-8	75.0%	5	4	0.037037	0.246914
3,MA,C	3-10	75.0%	7	8	0.0740741	0.691358
3,MA,OC	NA	NA			0	0
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	2.513158
					Avg Accuracy	79.9%
Stroke	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	NA	NA			0	0
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	NA	NA			0	0
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-6	83.3%	2	5	0.0548224	0.131579
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	NA	NA			0	0
2,Y,IC	NA	NA			0	0
2,Y,C	NA	NA			0	0
2,Y,OC	NA	NA			0	0
2,MA,LF	2-8	100.0%	6	6	0.0789474	0.473684
2,MA,IC	NA	NA			0	0
2,MA,C	NA	NA			0	0
2,MA,OC	NA	NA			0	0
2,O,LF	3-6	83.3%	3	12	0.1315737	0.473684
2,O,IC	4-6	76.9%	2	13	0.1315788	0.342105
2,O,C	4-6	71.4%	2	7	0.0657816	0.184211
2,O,OC	5-7	75.0%	2	4	0.0394737	0.105263
3,Y,LF	NA	NA			0	0
3,Y,IC	NA	NA			0	0
3,Y,C	NA	NA			0	0
3,Y,OC	NA	NA			0	0
3,MA,LF	NA	NA			0	0
3,MA,IC	NA	NA			0	0
3,MA,C	NA	NA			0	0
3,MA,OC	NA	NA			0	0
3,O,LF	4-7	75.0%	3	3	0.0296053	0.118421
3,O,IC	3-5	71.4%	2	6	0.0563684	0.157895
3,O,C	3-5	76.9%	2	13	0.1315737	0.342105
3,O,OC	3-5	85.7%	2	7	0.0789434	0.184211

					Avg Range	2.350949
					Avg Accuracy	79.5%
Trauma	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-5	75.0%	3	32	0.0325203	0.130081
1,Y,IC	3-5	90.0%	2	10	0.0121951	0.0271
1,Y,C	4-6	75.0%	2	4	0.004065	0.01084
1,Y,OC	NA	NA			0	0
1,MA,LF	2-4	83.3%	2	30	0.033874	0.081301
1,MA,IC	3-6	83.3%	3	6	0.0067748	0.02439
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-4	79.6%	2	93	0.1002585	0.252033
2,Y,IC	2-4	75.8%	2	33	0.033872	0.089431
2,Y,C	3-6	77.8%	3	27	0.0284524	0.109756
2,Y,OC	3-7	77.8%	4	9	0.0094841	0.04878
2,MA,LF	2-4	79.6%	2	95	0.1024146	0.257453
2,MA,IC	2-4	83.3%	2	31	0.0350031	0.084011
2,MA,C	2-4	78.9%	2	19	0.020313	0.051491
2,MA,OC	NA	NA			0	0
2,O,LF	2-6	79.5%	4	38	0.040935	0.205962
2,O,IC	3-6	75.0%	3	8	0.0081301	0.03252
2,O,C	3-6	100.0%	3	6	0.0081301	0.02439
2,O,OC	NA	NA			0	0
3,Y,LF	2-3	79.1%	1	43	0.0460648	0.058266
3,Y,IC	2-4	84.0%	2	42	0.0478049	0.113821
3,Y,C	2-4	78.4%	2	52	0.0552623	0.140921
3,Y,OC	2-6	77.2%	4	23	0.0240627	0.124661
3,MA,LF	2-4	76.0%	2	25	0.0257453	0.067751
3,MA,IC	2-4	80.6%	2	32	0.0349659	0.086721
3,MA,C	2-5	73.2%	3	41	0.04065	0.166667
3,MA,OC	2-5	75.0%	3	17	0.0172764	0.069106
3,O,LF	3-6	85.7%	3	7	0.0081297	0.028455
3,O,IC	4-5	75.0%	1	4	0.004065	0.00542
3,O,C	3-7	100.0%	4	11	0.0149051	0.059621
3,O,OC	NA	NA			0	0

					Avg Range	3.068966
					Avg Accuracy	79.7%
Wound	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-4	76.5%	2	17	0.0560341	0.146552
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	1-4	80.0%	4	18	0.062069	0.310345
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-5	100.0%	3	3	0.012931	0.038793
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	1-4	81.8%	3	23	0.0811047	0.297414
2,Y,IC	2-4	81.8%	2	11	0.0387892	0.094828
2,Y,C	1-3	100.0%	2	9	0.0387931	0.077586
2,Y,OC	NA	NA			0	0
2,MA,LF	1-4	76.9%	3	40	0.1326207	0.517241
2,MA,IC	2-6	71.4%	4	15	0.0461767	0.258621
2,MA,C	2-6	66.7%	4	10	0.0287371	0.172414
2,MA,OC	NA	NA			0	0
2,O,LF	2-5	100.0%	3	11	0.0474138	0.142241
2,O,IC	4-6	100.0%	2	3	0.012931	0.025862
2,O,C	4-7	66.6%	3	6	0.0172241	0.077586
2,O,OC	NA	NA			0	0
3,Y,LF	1-4	75.0%	3	6	0.0193966	0.077586
3,Y,IC	3-6	75.0%	4	8	0.0258621	0.137931
3,Y,C	2-5	70.0%	3	10	0.0301724	0.12931
3,Y,OC	3-5	80.0%	2	5	0.0172414	0.043103
3,MA,LF	2-3	100.0%	1	6	0.0258621	0.025862
3,MA,IC	2-6	77.7%	4	9	0.0301422	0.155172
3,MA,C	2-6	76.9%	4	13	0.0431017	0.224138
3,MA,OC	NA	NA			0	0
3,O,LF	2-5	80.0%	3	5	0.0172414	0.064655
3,O,IC	NA	NA			0	0
3,O,C	3-6	75.0%	3	4	0.012931	0.051724
3,O,OC	NA	NA			0	0

					Avg Range	3.140553
					Avg Accuracy	79.4%
Dyspnea	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-8	75.0%	5	17	0.0146889	0.097926
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-5	73.8%	2	76	0.0645737	0.175115
1,MA,IC	4-6	88.9%	2	9	0.0092157	0.020737
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	3-6	82.5%	3	80	0.0760369	0.276498
1,O,IC	3-5	85.7%	2	7	0.0069113	0.016129
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	4-6	76.7%	2	30	0.0264988	0.069124
2,Y,IC	4-5	80.0%	1	5	0.0046083	0.00576
2,Y,C	3-5	77.8%	2	9	0.0080637	0.020737
2,Y,OC	NA	NA			0	0
2,MA,LF	3-6	72.2%	3	22	0.0183046	0.076037
2,MA,IC	3-8	81.4%	5	26	0.0243795	0.14977
2,MA,C	4-8	80.8%	4	43	0.0400078	0.198157
2,MA,OC	5-8	100.0%	3	5	0.0057604	0.017281
2,O,LF	4-8	74.3%	4	136	0.1163521	0.626728
2,O,IC	3-6	84.0%	3	78	0.0754839	0.269585
2,O,C	4-7	79.0%	3	43	0.0391409	0.148618
2,O,OC	3-6	75.0%	3	8	0.0069124	0.02765
3,Y,LF	3-5	80.0%	2	10	0.0092166	0.023041
3,Y,IC	3-7	81.8%	4	11	0.0103676	0.050691
3,Y,C	3-6	84.6%	3	13	0.0126705	0.044931
3,Y,OC	3-7	80.0%	4	5	0.0046083	0.023041
3,MA,LF	3-5	80.0%	2	21	0.0193548	0.048387
3,MA,IC	3-6	82.6%	3	23	0.0218897	0.079493
3,MA,C	3-7	83.7%	4	43	0.0414742	0.198157
3,MA,OC	3-6	80.0%	3	10	0.0092166	0.034562
3,O,LF	3-6	84.6%	3	39	0.0380115	0.134793
3,O,IC	3-5	66.7%	2	30	0.0230392	0.069124
3,O,C	3-6	83.7%	3	49	0.0472331	0.169355
3,O,OC	3-6	85.0%	3	20	0.0195853	0.069124

					Avg Range	5.185468
					Avg Accuracy	78.3%
Evaluation	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-9	75.8%	7	34	0.0492447	0.455067
1,Y,IC	4-18	100.0%	14	4	0.0076482	0.107075
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-9	69.4%	7	37	0.0491258	0.49522
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-6	83.3%	4	13	0.020713	0.099426
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-6	82.9%	4	42	0.0665897	0.321224
2,Y,IC	2-8	76.2%	6	22	0.0320493	0.25239
2,Y,C	2-8	75.0%	6	21	0.0301147	0.240918
2,Y,OC	NA	NA			0	0
2,MA,LF	1-7	78.3%	6	70	0.1047457	0.803059
2,MA,IC	3-8	84.4%	5	33	0.0532354	0.315488
2,MA,C	2-9	76.2%	7	22	0.0320493	0.294455
2,MA,OC	2-8	100.0%	6	4	0.0076482	0.045889
2,O,LF	2-7	80.0%	5	36	0.0550669	0.344168
2,O,IC	4-6	72.2%	2	19	0.0262294	0.072658
2,O,C	5-8	75.0%	3	13	0.0186424	0.07457
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	80.0%	2	11	0.016826	0.042065
3,Y,IC	2-6	81.8%	4	12	0.0187732	0.091778
3,Y,C	3-6	85.7%	3	15	0.0245822	0.086042
3,Y,OC	4-9	75.0%	5	9	0.0129063	0.086042
3,MA,LF	2-7	81.8%	5	12	0.0187709	0.114723
3,MA,IC	2-7	73.9%	5	25	0.0353442	0.239006
3,MA,C	2-8	75.0%	6	29	0.041587	0.332696
3,MA,OC	3-9	71.4%	6	8	0.0109247	0.091778
3,O,LF	2-5	75.0%	3	9	0.0129063	0.051625
3,O,IC	4-5	85.7%	1	8	0.0131105	0.015296
3,O,C	4-7	85.7%	3	8	0.0131105	0.045889
3,O,OC	2-7	85.7%	5	7	0.0114717	0.066922

					Avg Range	3.26506
					Avg Accuracy	79.5%
Bleeding	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-4	78.6%	1	15	0.0236657	0.03012
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-6	73.1%	4	25	0.0366817	0.200803
1,MA,IC	3-6	100.0%	3	4	0.0080321	0.024096
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	3-6	75.0%	3	23	0.0346386	0.138554
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-7	75.9%	4	30	0.0456988	0.240964
2,Y,IC	4-8	75.0%	4	13	0.0195783	0.104418
2,Y,C	4-8	88.9%	4	10	0.0178474	0.080321
2,Y,OC	3-6	75.0%	3	5	0.0075301	0.03012
2,MA,LF	3-6	82.1%	3	40	0.0659036	0.240964
2,MA,IC	3-6	80.0%	3	31	0.0497992	0.186747
2,MA,C	4-7	81.3%	3	17	0.0277359	0.10241
2,MA,OC	NA	NA			0	0
2,O,LF	3-7	82.7%	4	53	0.0880034	0.425703
2,O,IC	3-6	70.3%	3	21	0.0296446	0.126506
2,O,C	2-5	88.9%	3	10	0.0178474	0.060241
2,O,OC	NA	NA			0	0
3,Y,LF	3-5	93.3%	2	16	0.0299855	0.064257
3,Y,IC	3-4	76.9%	1	14	0.0216185	0.028112
3,Y,C	4-7	90.0%	3	21	0.0379518	0.126506
3,Y,OC	5-7	83.3%	2	7	0.0117131	0.028112
3,MA,LF	4-8	76.9%	4	14	0.0216185	0.11245
3,MA,IC	3-7	75.0%	4	21	0.0316265	0.168675
3,MA,C	3-7	76.0%	4	26	0.0396787	0.208835
3,MA,OC	4-6	75.0%	2	5	0.0075301	0.02008
3,O,LF	3-7	76.5%	4	18	0.0276398	0.144578
3,O,IC	4-7	72.7%	3	23	0.0335855	0.138554
3,O,C	3-6	81.5%	3	28	0.045812	0.168675
3,O,OC	3-7	85.7%	4	8	0.0137671	0.064257

					Avg Range	4.03163
					Avg Accuracy	79.4%
Dizziness	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-5	75.0%	2	8	0.0145985	0.038929
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	4-9	81.3%	5	16	0.0316302	0.194647
1,MA,IC	5-7	100.0%	2	2	0.0048662	0.009732
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	3-10	75.0%	7	8	0.0145985	0.136253
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-9	80.6%	6	31	0.0607932	0.452555
2,Y,IC	3-5	80.0%	2	10	0.0194647	0.048662
2,Y,C	3-6	83.3%	3	6	0.012165	0.043796
2,Y,OC	NA	NA			0	0
2,MA,LF	4-9	75.0%	5	60	0.1094891	0.729927
2,MA,IC	3-7	80.6%	4	31	0.0608234	0.301703
2,MA,C	3-5	84.6%	2	13	0.0267591	0.06326
2,MA,OC	4-7	80.0%	3	5	0.0097324	0.036496
2,O,LF	4-8	77.8%	4	72	0.1362394	0.70073
2,O,IC	4-7	81.8%	3	22	0.0437912	0.160584
2,O,C	5-8	75.0%	3	8	0.0145985	0.058394
2,O,OC	3-6	100.0%	3	2	0.0048662	0.014599
3,Y,LF	3-5	85.7%	2	7	0.0145961	0.034063
3,Y,IC	3-7	87.5%	4	8	0.0170316	0.077859
3,Y,C	4-7	84.6%	3	13	0.0267591	0.094891
3,Y,OC	NA	NA			0	0
3,MA,LF	3-6	81.8%	3	11	0.0218956	0.080292
3,MA,IC	3-8	75.0%	5	16	0.0291971	0.194647
3,MA,C	4-7	76.2%	3	21	0.0389292	0.153285
3,MA,OC	8-10	100.0%	2	2	0.0048662	0.009732
3,O,LF	4-7	100.0%	3	3	0.0072993	0.021898
3,O,IC	5-8	80.0%	3	10	0.0194647	0.072993
3,O,C	4-9	75.0%	5	20	0.0364964	0.243309
3,O,OC	4-8	87.5%	4	6	0.0127737	0.058394

					Avg Range	2.905229
					Avg Accuracy	80.8%
Nausea	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	4-6	75.0%	2	16	0.0392157	0.104575
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	5-8	76.5%	3	17	0.0424833	0.166667
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-7	76.5%	3	17	0.0424833	0.166667
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-6	88.0%	3	25	0.0718954	0.245098
2,Y,IC	5-6	80.0%	1	10	0.0261438	0.03268
2,Y,C	4-5	83.3%	1	12	0.0326667	0.039216
2,Y,OC	NA	NA			0	0
2,MA,LF	3-8	80.0%	5	45	0.1176471	0.735294
2,MA,IC	5-9	75.0%	4	18	0.0441176	0.235294
2,MA,C	6-8	75.0%	2	12	0.0294118	0.078431
2,MA,OC	5-6	100.0%	1	3	0.0098039	0.009804
2,O,LF	5-7	80.0%	2	20	0.0522876	0.130719
2,O,IC	6-8	80.0%	2	10	0.0261438	0.065359
2,O,C	5-8	100.0%	3	3	0.0098039	0.029412
2,O,OC	NA	NA			0	0
3,Y,LF	4-7	77.8%	3	9	0.0228765	0.088235
3,Y,IC	4-6	77.8%	2	9	0.0228765	0.058824
3,Y,C	4-8	85.7%	4	14	0.0392137	0.183007
3,Y,OC	4-5	75.0%	1	4	0.0098039	0.013072
3,MA,LF	4-5	100.0%	1	4	0.0130719	0.013072
3,MA,IC	4-6	84.6%	2	13	0.0359412	0.084967
3,MA,C	4-7	70.6%	3	17	0.0392111	0.166667
3,MA,OC	7-9	92.3%	2	13	0.0392124	0.084967
3,O,LF	NA	NA			0	0
3,O,IC	3-7	85.7%	4	8	0.0224157	0.104575
3,O,C	6-9	85.7%	3	7	0.0196046	0.068627
3,O,OC	NA	NA			0	0

					Avg Range	2.678431
					Avg Accuracy	80.8%
Change of Mental Status	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	6-8	60.0%	2	5	0.0117647	0.039216
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	4-5	70.0%	1	10	0.027451	0.039216
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-6	78.6%	2	14	0.0431365	0.109804
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	4-8	80.0%	4	5	0.0156863	0.078431
2,Y,IC	5-7	100.0%	2	10	0.0392157	0.078431
2,Y,C	NA	NA			0	0
2,Y,OC	NA	NA			0	0
2,MA,LF	4-7	85.7%	3	21	0.0705847	0.247059
2,MA,IC	6-9	73.3%	3	15	0.0431353	0.176471
2,MA,C	4-7	83.3%	3	6	0.0196071	0.070588
2,MA,OC	NA	NA			0	0
2,O,LF	4-7	81.8%	3	55	0.1764529	0.647059
2,O,IC	4-7	75.9%	3	29	0.0862722	0.341176
2,O,C	5-8	86.7%	3	15	0.0509824	0.176471
2,O,OC	NA	NA			0	0
3,Y,LF	NA	NA			0	0
3,Y,IC	NA	NA			0	0
3,Y,C	NA	NA			0	0
3,Y,OC	NA	NA			0	0
3,MA,LF	4-6	80.0%	2	5	0.0156863	0.039216
3,MA,IC	NA	NA			0	0
3,MA,C	6-9	90.0%	3	10	0.0352941	0.117647
3,MA,OC	4-5	60.0%	2	5	0.0117647	0.039216
3,O,LF	4-6	81.8%	2	11	0.0352906	0.086275
3,O,IC	5-7	77.8%	2	9	0.0274482	0.070588
3,O,C	5-8	86.4%	3	22	0.0745067	0.258824
3,O,OC	5-7	75.0%	2	8	0.0235294	0.062745

					Avg Range	3.089147
					Avg Accuracy	80.5%
Swelling	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-6	75.0%	4	8	0.0232558	0.124031
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-8	80.0%	5	20	0.0620155	0.387597
1,MA,IC	4-5	100.0%	1	3	0.0116279	0.011628
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	2-6	75.0%	4	4	0.0116279	0.062016
1,O,IC	2-4	100.0%	2	2	0.0077519	0.015504
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-6	74.6%	4	13	0.0375891	0.20155
2,Y,IC	2-6	83.3%	4	6	0.0193721	0.093023
2,Y,C	NA	NA			0	0
2,Y,OC	4-7	100.0%	3	3	0.0116279	0.034884
2,MA,LF	2-5	69.7%	3	33	0.0891384	0.383721
2,MA,IC	5-7	76.5%	2	17	0.0503872	0.131783
2,MA,C	4-6	85.7%	2	7	0.0232547	0.054264
2,MA,OC	2-4	100.0%	2	4	0.0155039	0.031008
2,O,LF	4-6	76.7%	2	30	0.0891512	0.232558
2,O,IC	4-7	85.7%	3	14	0.0465093	0.162791
2,O,C	5-9	85.4%	4	7	0.0231733	0.108527
2,O,OC	NA	NA			0	0
3,Y,LF	2-6	100.0%	4	3	0.0116279	0.046512
3,Y,IC	2-4	75.0%	2	4	0.0116279	0.031008
3,Y,C	3-4	80.0%	1	5	0.0155039	0.01938
3,Y,OC	5-8	100.0%	3	3	0.0116279	0.034884
3,MA,LF	2-3	83.3%	1	6	0.0193721	0.023256
3,MA,IC	3-6	81.0%	3	21	0.0658895	0.244186
3,MA,C	3-7	78.6%	4	14	0.0426349	0.217054
3,MA,OC	5-8	90.0%	3	10	0.0348837	0.116279
3,O,LF	3-6	100.0%	3	3	0.0116279	0.034884
3,O,IC	4-6	100.0%	2	4	0.0155039	0.031008
3,O,C	2-7	80.0%	5	10	0.0310078	0.193798
3,O,OC	5-9	75.0%	4	4	0.0116279	0.062016

					Avg Range	2.266932
					Avg Accuracy	80.9%
Cough	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-4	100.0%	2	6	0.0239044	0.047809
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-4	86.7%	2	15	0.0517948	0.119522
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-5	100.0%	1	5	0.0199203	0.01992
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-6	76.0%	3	25	0.0756972	0.298805
2,Y,IC	2-4	81.8%	2	11	0.035853	0.087649
2,Y,C	2-4	75.0%	2	4	0.0119522	0.031873
2,Y,OC	NA	NA			0	0
2,MA,LF	3-5	70.7%	2	41	0.1155351	0.326693
2,MA,IC	4-6	73.3%	2	15	0.0438227	0.119522
2,MA,C	4-5	80.0%	1	10	0.0318725	0.039841
2,MA,OC	NA	NA			0	0
2,O,LF	3-6	84.4%	3	32	0.1075697	0.38247
2,O,IC	4-8	88.9%	4	9	0.0318693	0.143426
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	85.7%	2	7	0.0239032	0.055777
3,Y,IC	3-4	75.0%	1	4	0.0119522	0.015936
3,Y,C	2-4	75.0%	2	12	0.0358566	0.095618
3,Y,OC	4-5	100.0%	1	3	0.0119522	0.011952
3,MA,LF	2-6	85.7%	4	7	0.0239032	0.111554
3,MA,IC	3-6	90.9%	3	11	0.039841	0.131474
3,MA,C	3-4	81.8%	1	11	0.035853	0.043825
3,MA,OC	NA	NA			0	0
3,O,LF	4-6	75.0%	2	4	0.0119522	0.031873
3,O,IC	3-5	85.7%	2	7	0.0239032	0.055777
3,O,C	5-7	75.0%	2	8	0.0239044	0.063745
3,O,OC	5-7	100.0%	2	4	0.0159363	0.031873

					Avg Range	4.172285
					Avg Accuracy	80.2%
Diarrhea	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	4-8	69.2%	4	13	0.0337075	0.194757
1,Y,IC	4-7	100.0%	3	2	0.0074906	0.022472
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	4-9	80.0%	5	19	0.0569288	0.355805
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	3-7	85.7%	4	7	0.0224708	0.104869
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-8	77.8%	5	18	0.0524292	0.337079
2,Y,IC	4-8	80.0%	4	10	0.0299625	0.149813
2,Y,C	NA	NA			0	0
2,Y,OC	NA	NA			0	0
2,MA,LF	4-9	82.9%	5	41	0.1273303	0.76779
2,MA,IC	4-7	76.5%	3	17	0.0486888	0.191011
2,MA,C	7-10	81.8%	3	11	0.0337045	0.123596
2,MA,OC	NA	NA			0	0
2,O,LF	3-7	75.8%	4	33	0.0936236	0.494382
2,O,IC	4-9	75.0%	5	16	0.0449438	0.299625
2,O,C	7-10	83.3%	3	6	0.0187258	0.067416
2,O,OC	3-5	100.0%	2	2	0.0074906	0.014981
3,Y,LF	4-6	87.5%	2	8	0.0262172	0.059925
3,Y,IC	3-6	75.0%	3	4	0.011236	0.044944
3,Y,C	4-9	77.8%	5	9	0.026218	0.168539
3,Y,OC	NA	NA			0	0
3,MA,LF	3-7	85.7%	4	7	0.0224708	0.104869
3,MA,IC	3-8	81.8%	5	11	0.0337045	0.205993
3,MA,C	5-9	77.2%	4	9	0.0260157	0.134831
3,MA,OC	6-10	100.0%	4	3	0.011236	0.044944
3,O,LF	3-4	100.0%	1	4	0.0149813	0.014981
3,O,IC	5-11	85.7%	6	7	0.0224708	0.157303
3,O,C	4-7	80.0%	3	10	0.0299625	0.11236
3,O,OC	NA	NA			0	0

					Avg Range	8.850427
					Avg Accuracy	80.1%
Suicidal Ideation	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	7-12	77.8%	5	18	0.0598231	0.384615
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	7-14	76.0%	7	25	0.0811966	0.747863
1,MA,IC	8-16	100.0%	8	3	0.0128205	0.102564
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	5-11	81.0%	6	21	0.0726474	0.538462
2,Y,IC	6-19	81.8%	13	11	0.0384577	0.611111
2,Y,C	5-10	85.7%	5	7	0.0256397	0.149573
2,Y,OC	6-9	100.0%	3	3	0.0128205	0.038462
2,MA,LF	6-15	80.0%	9	30	0.1025641	1.153846
2,MA,IC	5-16	78.9%	10	19	0.0640641	0.811966
2,MA,C	6-10	77.7%	4	11	0.0365256	0.188034
2,MA,OC	NA	NA			0	0
2,O,LF	4-6	100.0%	2	3	0.0128205	0.025641
2,O,IC	NA	NA			0	0
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	4-18	85.7%	14	7	0.0256397	0.418803
3,Y,IC	6-18	77.8%	12	9	0.0299115	0.461538
3,Y,C	4-13	78.6%	9	14	0.0470077	0.538462
3,Y,OC	4-13	75.0%	9	8	0.025641	0.307692
3,MA,LF	9-21	75.0%	12	8	0.025641	0.410256
3,MA,IC	8-21	86.6%	13	15	0.0555128	0.833333
3,MA,C	7-19	75.0%	12	16	0.0512821	0.820513
3,MA,OC	10-22	83.3%	12	6	0.0213667	0.307692
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	3.756303
					Avg Accuracy	80.5%
Seizure	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	5-11	83.3%	6	6	0.0210076	0.151261
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	4-8	77.3%	4	22	0.0714261	0.369748
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	6-11	100.0%	5	3	0.012605	0.063025
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	4-8	78.6%	4	14	0.0462176	0.235294
2,Y,IC	4-7	87.5%	3	8	0.0294118	0.10084
2,Y,C	4-7	100.0%	3	6	0.0252101	0.07563
2,Y,OC	NA	NA			0	0
2,MA,LF	3-8	75.0%	5	36	0.1134454	0.756303
2,MA,IC	3-7	79.2%	4	24	0.0798252	0.403361
2,MA,C	3-6	100.0%	3	6	0.0252101	0.07563
2,MA,OC	6-9	100.0%	3	3	0.012605	0.037815
2,O,LF	4-7	81.8%	3	11	0.0378113	0.138655
2,O,IC	4-7	75.0%	3	8	0.0252101	0.10084
2,O,C	5-7	100.0%	2	3	0.012605	0.02521
2,O,OC	NA	NA			0	0
3,Y,LF	3-7	77.8%	4	8	0.0261412	0.134454
3,Y,IC	3-5	87.5%	3	8	0.0294118	0.10084
3,Y,C	3-7	85.7%	4	14	0.0504176	0.235294
3,Y,OC	3-5	83.3%	2	5	0.0175063	0.042017
3,MA,LF	5-6	80.0%	1	5	0.0168067	0.021008
3,MA,IC	3-7	75.0%	4	12	0.0378151	0.201681
3,MA,C	3-8	85.7%	5	14	0.0504176	0.294118
3,MA,OC	4-5	75.0%	2	12	0.0378151	0.10084
3,O,LF	4-5	100.0%	1	4	0.0168067	0.016807
3,O,IC	4-7	66.7%	3	3	0.0084025	0.037815
3,O,C	4-7	10.0%	3	3	0.0012605	0.037815
3,O,OC	NA	NA			0	0

					Avg Range	2.160194
					Avg Accuracy	81.1%
Syncope	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-7	66.6%	4	9	0.0290971	0.174757
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-5	85.7%	2	7	0.0291248	0.067961
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-7	87.5%	3	8	0.0339806	0.116505
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-4	78.3%	2	23	0.0873777	0.223301
2,Y,IC	3-5	80.0%	2	5	0.0194175	0.048544
2,Y,C	5-7	100.0%	2	3	0.0145631	0.029126
2,Y,OC	NA	NA			0	0
2,MA,LF	2-4	85.7%	2	14	0.0582495	0.135922
2,MA,IC	4-6	77.8%	2	9	0.0339772	0.087379
2,MA,C	5-7	75.0%	2	4	0.0145631	0.038835
2,MA,OC	NA	NA			0	0
2,O,LF	4-6	74.4%	2	39	0.1408354	0.378641
2,O,IC	2-4	77.8%	2	4	0.015101	0.038835
2,O,C	4-6	66.0%	2	9	0.028835	0.087379
2,O,OC	NA	NA			0	0
3,Y,LF	2-4	80.0%	2	5	0.0194175	0.048544
3,Y,IC	2-4	87.6%	2	8	0.0340117	0.07767
3,Y,C	3-6	100.0%	3	7	0.0339806	0.101942
3,Y,OC	3-5	83.3%	2	6	0.0242709	0.058252
3,MA,LF	NA	NA			0	0
3,MA,IC	NA	NA			0	0
3,MA,C	3-5	91.7%	2	12	0.0534	0.116505
3,MA,OC	4-6	75.0%	2	4	0.0145631	0.038835
3,O,LF	3-5	80.0%	2	10	0.038835	0.097087
3,O,IC	3-5	100.0%	2	7	0.0339806	0.067961
3,O,C	3-5	80.0%	2	10	0.038835	0.097087
3,O,OC	4-6	100.0%	2	3	0.0145631	0.029126

					Avg Range	2.700637
					Avg Accuracy	80.7%
Sore Throat	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	1-4	83.3%	3	12	0.0636917	0.229299
1,Y,IC	4-8	100.0%	4	3	0.0191083	0.076433
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-4	80.0%	2	5	0.0254777	0.063694
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	1-4	86.4%	3	44	0.242028	0.840764
2,Y,IC	2-4	75.0%	2	12	0.0573248	0.152866
2,Y,C	2-5	81.9%	3	9	0.0469318	0.171975
2,Y,OC	NA	NA			0	0
2,MA,LF	1-4	71.4%	3	14	0.0636688	0.267516
2,MA,IC	2-4	75.0%	2	4	0.0191083	0.050955
2,MA,C	2-4	80.0%	2	5	0.0254777	0.063694
2,MA,OC	NA	NA			0	0
2,O,LF	1-4	72.2%	3	18	0.0828	0.343949
2,O,IC	NA	NA			0	0
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	2-5	85.7%	3	7	0.0382146	0.133758
3,Y,IC	2-4	83.3%	2	6	0.0318459	0.076433
3,Y,C	3-5	72.2%	2	8	0.0368	0.101911
3,Y,OC	3-5	100.0%	2	4	0.0254777	0.050955
3,MA,LF	2-4	75.0%	2	4	0.0191083	0.050955
3,MA,IC	2-4	75.0%	2	2	0.0095541	0.025478
3,MA,C	NA	NA			0	0
3,MA,OC	NA	NA			0	0
3,O,LF	NA	NA			0	0
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	2.207547
					Avg Accuracy	80.2%
Paresthesia	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	NA	NA			0	0
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	3-5	66.7%	2	9	0.0565981	0.169811
1,MA,IC	NA	NA			0	0
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	NA	NA			0	0
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	4-7	66.0%	3	6	0.0373585	0.169811
2,Y,IC	4-7	83.3%	3	6	0.0471679	0.169811
2,Y,C	7-9	91.7%	2	12	0.1037774	0.226415
2,Y,OC	NA	NA			0	0
2,MA,LF	3-5	76.2%	2	21	0.1509425	0.396226
2,MA,IC	5-6	80.0%	1	5	0.0377358	0.04717
2,MA,C	5-8	100.0%	3	3	0.0283019	0.084906
2,MA,OC	NA	NA			0	0
2,O,LF	4-8	85.7%	4	7	0.0566009	0.264151
2,O,IC	4-6	80.0%	2	5	0.0377358	0.09434
2,O,C	4-9	100.0%	2	2	0.0188679	0.037736
2,O,OC	NA	NA			0	0
3,Y,LF	5-6	100.0%	1	2	0.0188679	0.018868
3,Y,IC	NA	NA			0	0
3,Y,C	4-6	75.0%	2	4	0.0283019	0.075472
3,Y,OC	NA	NA			0	0
3,MA,LF	4-6	80.0%	2	5	0.0377358	0.09434
3,MA,IC	4-6	83.3%	2	6	0.0471679	0.113208
3,MA,C	4-6	75.0%	2	4	0.0283019	0.075472
3,MA,OC	4-6	75.0%	2	4	0.0283019	0.075472
3,O,LF	4-6	80.0%	2	5	0.0377358	0.09434
3,O,IC	NA	NA			0	0
3,O,C	NA	NA			0	0
3,O,OC	NA	NA			0	0

					Avg Range	3.544355
					Avg Accuracy	80.5%
Flank Pain	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-7	80.0%	4	10	0.0322581	0.16129
1,Y,IC	NA	NA			0	0
1,Y,C	5-8	100.0%	3	2	0.0080645	0.024194
1,Y,OC	NA	NA			0	0
1,MA,LF	5-9	76.5%	4	17	0.052419	0.274194
1,MA,IC	7-10	100.0%	3	3	0.0120968	0.03629
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	4-10	84.6%	6	13	0.044352	0.314516
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	3-5	84.6%	2	13	0.044352	0.104839
2,Y,IC	5-8	75.0%	3	11	0.0332661	0.133065
2,Y,C	5-6	66.0%	1	3	0.0079839	0.012097
2,Y,OC	NA	NA			0	0
2,MA,LF	3-6	82.6%	3	46	0.1532097	0.556452
2,MA,IC	4-9	76.5%	5	17	0.052419	0.342742
2,MA,C	4-8	81.8%	4	9	0.0296891	0.145161
2,MA,OC	3-6	100.0%	3	3	0.0120968	0.03629
2,O,LF	5-8	68.8%	3	16	0.0443548	0.193548
2,O,IC	7-10	75.0%	3	4	0.0120968	0.048387
2,O,C	NA	NA			0	0
2,O,OC	NA	NA			0	0
3,Y,LF	3-4	85.7%	1	7	0.0241923	0.028226
3,Y,IC	3-7	71.4%	4	7	0.0201589	0.112903
3,Y,C	4-8	75.0%	4	8	0.0241935	0.129032
3,Y,OC	4-8	81.8%	4	11	0.0362867	0.177419
3,MA,LF	3-6	90.5%	3	11	0.0401411	0.133065
3,MA,IC	3-6	76.9%	3	13	0.0403105	0.157258
3,MA,C	4-9	78.6%	5	14	0.044354	0.282258
3,MA,OC	5-8	80.0%	3	5	0.016129	0.060484
3,O,LF	3-7	100.0%	4	3	0.0120968	0.048387
3,O,IC	NA	NA			0	0
3,O,C	3-7	100.0%	4	2	0.0080645	0.032258
3,O,OC	NA	NA			0	0

					Avg Range	5.178715
					Avg Accuracy	79.9%
Head Pain	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-8	76.9%	6	26	0.040159	0.313253
1,Y,IC	NA	NA			0	0
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	1-7	79.4%	6	34	0.0542157	0.409639
1,MA,IC	6-8	75.0%	2	4	0.0060241	0.016064
1,MA,C	NA	NA			0	0
1,MA,OC	NA	NA			0	0
1,O,LF	5-9	75.0%	4	8	0.0120482	0.064257
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-8	85.4%	6	48	0.0823229	0.578313
2,Y,IC	5-9	76.0%	4	25	0.0381526	0.200803
2,Y,C	3-9	75.0%	6	20	0.0301205	0.240964
2,Y,OC	NA	NA			0	0
2,MA,LF	2-8	81.8%	6	77	0.1264934	0.927711
2,MA,IC	4-7	81.5%	3	27	0.0441759	0.162651
2,MA,C	3-8	78.6%	5	28	0.0441759	0.281124
2,MA,OC	NA	NA			0	0
2,O,LF	3-7	80.0%	4	20	0.0321285	0.160643
2,O,IC	4-10	80.0%	6	10	0.0160643	0.120482
2,O,C	4-6	75.0%	2	4	0.0060241	0.016064
2,O,OC	NA	NA			0	0
3,Y,LF	3-8	78.6%	5	14	0.022088	0.140562
3,Y,IC	1-6	85.7%	5	35	0.060238	0.351406
3,Y,C	3-8	75.0%	5	5	0.0075301	0.050201
3,Y,OC	NA	NA			0	0
3,MA,LF	3-8	76.5%	5	17	0.0261042	0.170683
3,MA,IC	3-8	77.3%	5	22	0.0341353	0.220884
3,MA,C	3-9	76.5%	6	34	0.0522084	0.409639
3,MA,OC	5-9	75.0%	4	12	0.0180723	0.096386
3,O,LF	3-5	80.0%	2	5	0.0080321	0.02008
3,O,IC	3-10	85.7%	7	7	0.0120476	0.098394
3,O,C	4-8	77.8%	4	9	0.0140548	0.072289
3,O,OC	4-8	85.7%	4	7	0.0120476	0.056225

					Avg Range	3.842814
					Avg Accuracy	80.5%
Back Pain	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-5	82.6%	3	23	0.0284401	0.103293
1,Y,IC	2-5	75.0%	3	4	0.004491	0.017964
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	1-5	82.4%	4	68	0.0838293	0.407186
1,MA,IC	4-6	100.0%	4	8	0.011976	0.047904
1,MA,C	2-5	100.0%	3	3	0.004491	0.013473
1,MA,OC	NA	NA			0	0
1,O,LF	2-8	72.2%	6	18	0.0194605	0.161677
1,O,IC	NA	NA			0	0
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-6	80.0%	4	50	0.0598802	0.299401
2,Y,IC	2-6	80.7%	4	26	0.0314141	0.155689
2,Y,C	2-5	88.9%	3	9	0.0119749	0.040419
2,Y,OC	2-4	75.0%	2	4	0.004491	0.011976
2,MA,LF	2-6	76.3%	4	135	0.1541789	0.808383
2,MA,IC	2-6	79.5%	4	44	0.0523916	0.263473
2,MA,C	2-9	80.6%	4	31	0.0374228	0.185629
2,MA,OC	4-6	85.7%	2	7	0.0089816	0.020958
2,O,LF	3-7	75.0%	4	32	0.0359281	0.191617
2,O,IC	3-7	88.2%	4	17	0.0224537	0.101796
2,O,C	3-6	85.7%	3	7	0.0089816	0.031437
2,O,OC	NA	NA			0	0
3,Y,LF	2-6	89.5%	4	19	0.0254481	0.113772
3,Y,IC	2-5	81.8%	3	9	0.0110223	0.040419
3,Y,C	2-6	89.5%	4	19	0.0254481	0.113772
3,Y,OC	5-7	80.0%	2	10	0.011976	0.02994
3,MA,LF	2-6	87.9%	4	19	0.0250015	0.113772
3,MA,IC	3-5	76.5%	2	17	0.0194609	0.050898
3,MA,C	2-6	76.3%	4	38	0.0434099	0.227545
3,MA,OC	3-7	81.8%	4	22	0.0269434	0.131737
3,O,LF	3-6	80.0%	3	5	0.005988	0.022455
3,O,IC	3-6	80.0%	3	10	0.011976	0.04491
3,O,C	5-10	80.0%	5	11	0.0131737	0.082335
3,O,OC	6-8	100.0%	2	3	0.004491	0.008982

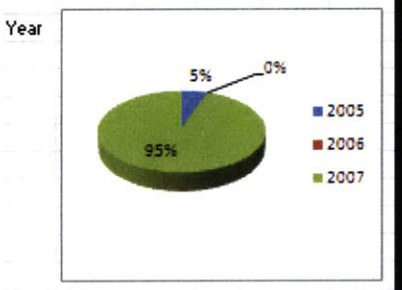
					Avg Range	9.14058
					Avg Accuracy	79.9%
Chest Pain	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	2-5	83.9%	3	31	0.0188404	0.067391
1,Y,IC	5-9	83.3%	4	6	0.003623	0.017391
1,Y,C	NA	NA			0	0
1,Y,OC	NA	NA			0	0
1,MA,LF	2-11	83.3%	9	126	0.0760839	0.821739
1,MA,IC	3-9	81.8%	6	11	0.0065211	0.047826
1,MA,C	3-9	75.0%	4	4	0.0021739	0.011594
1,MA,OC	NA	NA			0	0
1,O,LF	2-9	81.3%	7	59	0.0347587	0.299275
1,O,IC	4-12	83.3%	8	6	0.003623	0.034783
1,O,C	NA	NA			0	0
1,O,OC	NA	NA			0	0
2,Y,LF	2-7	78.6%	5	56	0.0318835	0.202899
2,Y,IC	2-6	84.2%	4	19	0.0115941	0.055072
2,Y,C	3-8	75.0%	5	12	0.0065217	0.043478
2,Y,OC	3-9	100.0%	6	3	0.0021739	0.013043
2,MA,LF	1-19	77.4%	18	190	0.1065101	2.478261
2,MA,IC	3-20	77.0%	17	87	0.0485498	1.071739
2,MA,C	3-22	85.2%	19	54	0.0333313	0.743478
2,MA,OC	2-16	76.2%	14	21	0.0115941	0.213043
2,O,LF	3-9	85.2%	6	149	0.0920237	0.647826
2,O,IC	3-7	79.6%	4	54	0.0311557	0.156522
2,O,C	2-7	80.8%	5	26	0.0152157	0.094203
2,O,OC	4-7	100.0%	3	7	0.0050725	0.015217
3,Y,LF	2-7	75.0%	5	24	0.0130435	0.086957
3,Y,IC	2-6	77.8%	4	18	0.0101452	0.052174
3,Y,C	2-8	78.8%	6	33	0.0188387	0.143478
3,Y,OC	3-8	75.0%	5	20	0.0108696	0.072464
3,MA,LF	2-8	78.0%	6	50	0.0282609	0.217391
3,MA,IC	2-10	75.4%	8	61	0.0333334	0.353623
3,MA,C	2-17	79.1%	5	91	0.0521733	0.32971
3,MA,OC	3-12	78.0%	9	41	0.0231739	0.267391
3,O,LF	3-6	75.0%	3	28	0.0152174	0.06087
3,O,IC	2-14	81.5%	12	27	0.0159417	0.234783
3,O,C	3-9	78.0%	6	50	0.0282609	0.217391
3,O,OC	4-10	75.0%	6	16	0.0086957	0.069565

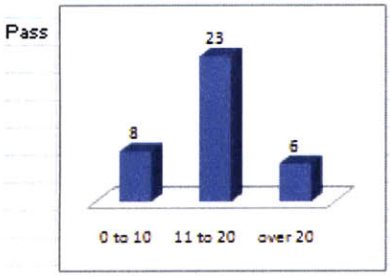
					Avg Range	4.945376
					Avg Accuracy	80.5%
Chest Pain	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-8	90.7%	5	80	0.045024	0.248293
1,Y,IC	5-9	77.8%	4	9	0.0043447	0.022346
1,Y,C	5-9	77.8%	4	9	0.0043447	0.022346
1,Y,OC	NA	NA			0	0
1,MA,LF	3-8	75.9%	5	133	0.0626858	0.412787
1,MA,IC	6-9	87.5%	3	8	0.0043451	0.014898
1,MA,C	9-12	83.3%	3	6	0.0031035	0.011173
1,MA,OC	NA	NA			0	0
1,O,LF	4-10	79.7%	6	59	0.0291741	0.219739
1,O,IC	8-10	100.0%	2	3	0.0018622	0.003724
1,O,C	3-9	75.0%	2	4	0.0018622	0.004966
1,O,OC	NA	NA			0	0
2,Y,LF	3-8	83.6%	5	128	0.0664154	0.397269
2,Y,IC	3-8	79.1%	5	43	0.0211023	0.133457
2,Y,C	4-9	63.1%	5	63	0.0246799	0.195531
2,Y,OC	7-11	100.0%	4	3	0.0018622	0.007449
2,MA,LF	3-9	87.2%	6	223	0.1206498	0.83054
2,MA,IC	4-8	75.3%	4	73	0.0341392	0.181254
2,MA,C	5-10	78.9%	5	57	0.0279304	0.176909
2,MA,OC	6-11	76.9%	5	13	0.0062071	0.040348
2,O,LF	4-9	78.1%	5	96	0.0465549	0.297952
2,O,IC	6-10	82.1%	4	39	0.0198631	0.096834
2,O,C	4-9	78.3%	5	23	0.0111731	0.071384
2,O,OC	5-7	100.0%	2	2	0.0012415	0.002483
3,Y,LF	3-8	83.3%	5	42	0.0217248	0.130354
3,Y,IC	4-8	84.5%	4	27	0.0141587	0.067039
3,Y,C	3-8	75.4%	5	61	0.0285538	0.189323
3,Y,OC	3-8	78.1%	5	32	0.0155183	0.099317
3,MA,LF	3-7	78.5%	4	65	0.0316567	0.16139
3,MA,IC	4-9	84.9%	5	83	0.0437412	0.257604
3,MA,C	4-9	78.2%	5	87	0.0422093	0.270019
3,MA,OC	6-11	77.8%	5	36	0.017381	0.111732
3,O,LF	4-9	86.7%	5	30	0.0161378	0.09311
3,O,IC	6-9	81.3%	3	16	0.0080695	0.029795
3,O,C	5-9	76.1%	4	46	0.0217236	0.114215
3,O,OC	5-9	75.0%	4	12	0.0055866	0.029795

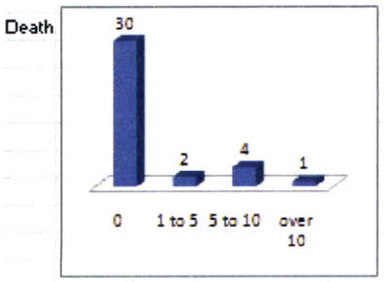
					Avg Range	3.940397
					Avg Accuracy	80.1%
Fever	Act. Range	> 75%	Hour Range	Total	W. Accuracy	W. Hour
1,Y,LF	3-5	81.0%	2	21	0.0375265	0.092715
1,Y,IC	NA	NA		0	0	0
1,Y,C	NA	NA		0	0	0
1,Y,OC	NA	NA		0	0	0
1,MA,LF	3-9	80.0%	6	25	0.0441501	0.331126
1,MA,IC	3-7	100.0%	4	3	0.0066225	0.02649
1,MA,C	NA	NA		0	0	0
1,MA,OC	NA	NA		0	0	0
1,O,LF	3-6	80.0%	3	15	0.0264901	0.099338
1,O,IC	4-5	100.0%	1	6	0.013245	0.013245
1,O,C	NA	NA		0	0	0
1,O,OC	NA	NA		0	0	0
2,Y,LF	4-9	77.5%	5	31	0.0530155	0.342163
2,Y,IC	4-7	75.0%	3	8	0.013245	0.05298
2,Y,C	5-8	85.7%	3	7	0.0132428	0.046358
2,Y,OC	NA	NA		0	0	0
2,MA,LF	3-8	83.3%	5	42	0.0772596	0.463576
2,MA,IC	4-9	81.0%	5	21	0.0375265	0.231788
2,MA,C	4-9	81.8%	5	11	0.0198656	0.121413
2,MA,OC	7-10	25.0%	3	4	0.0022075	0.02649
2,O,LF	3-6	77.4%	3	31	0.0529737	0.205298
2,O,IC	4-8	80.0%	4	10	0.01766	0.0883
2,O,C	5-7	75.0%	2	12	0.0198675	0.05298
2,O,OC	4-6	100.0%	2	3	0.0066225	0.013245
3,Y,LF	4-5	75.0%	1	8	0.013245	0.01766
3,Y,IC	3-5	87.6%	2	14	0.0270636	0.06181
3,Y,C	4-7	93.3%	3	15	0.030894	0.099338
3,Y,OC	4-7	80.0%	3	5	0.00883	0.033113
3,MA,LF	3-6	80.0%	3	20	0.0353201	0.13245
3,MA,IC	4-7	82.1%	4	28	0.0507709	0.247241
3,MA,C	4-8	75.0%	4	32	0.0529801	0.282561
3,MA,OC	4-8	81.8%	4	11	0.0198656	0.09713
3,O,LF	4-9	73.2%	5	21	0.0339291	0.231788
3,O,IC	4-10	80.0%	6	15	0.0264901	0.198675
3,O,C	4-9	75.0%	5	24	0.0397351	0.264901
3,O,OC	4-7	90.0%	3	10	0.0198675	0.066225

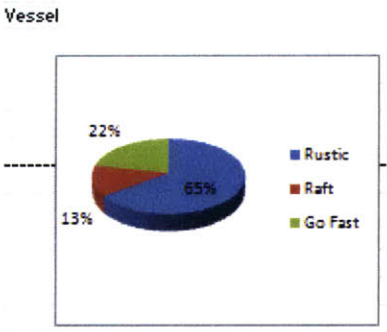
Appendix G – Cluster Differences

Cluster 1	Angle	2005	2006	2007
Long/Lat	2.937			
Year	26.31	2	0	35
Pass	37.68	7.5	0	15.26
Death	28.04	4	0	1.171
		Rustic	Raft	GoFast
Vessel	7.622	24	5	8

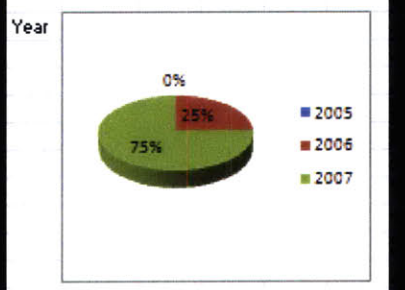


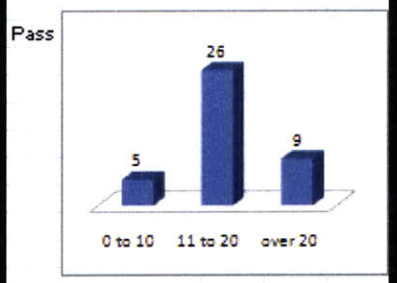


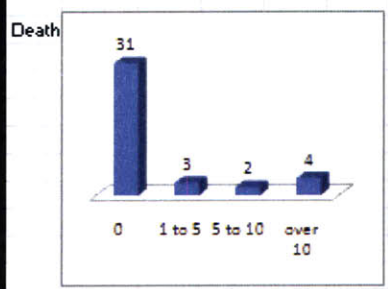


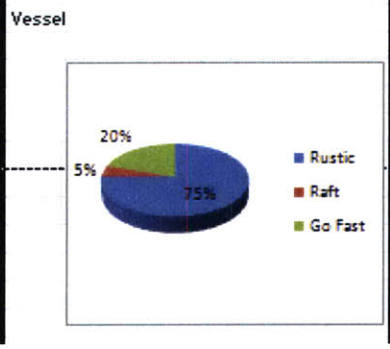


Cluster 2	Angle	2005	2006	2007
Long/Lat	2.998			
Year	11.54	0	10	30
Pass	32.24	0	15.4	16.97
Death	54.04	0	0.2	3.1
		Rustic	Raft	GoFas
Vessel	13.1	30	2	8

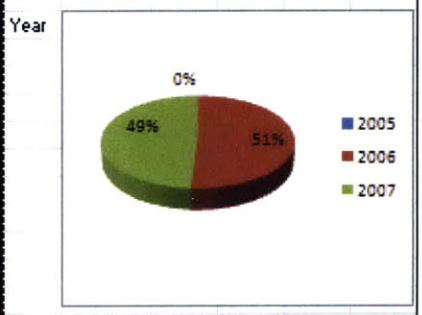


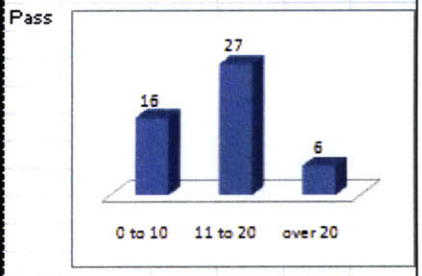


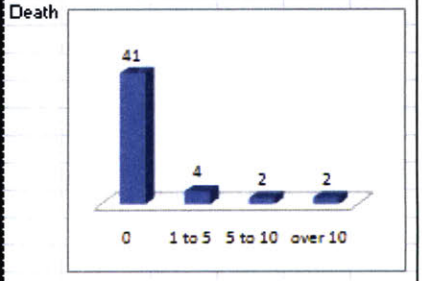


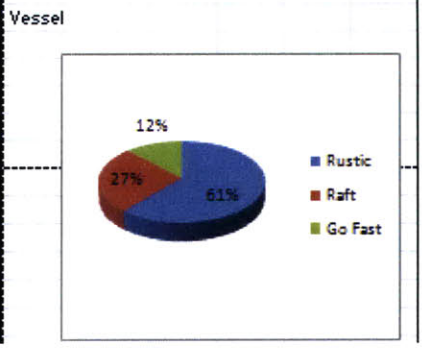


Cluster 3	Angle	2005	2006	2007
Long/Lat	1.854			
Year	22.07	0	25	24
Pass	32.42	0	13.4	12.67
Death	59.56	0	1.6	0.583
		Rustic	Raft	GoFast
Vessel	6.568	30	13	66

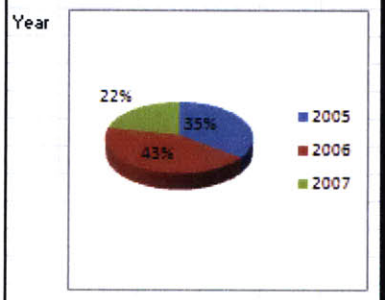


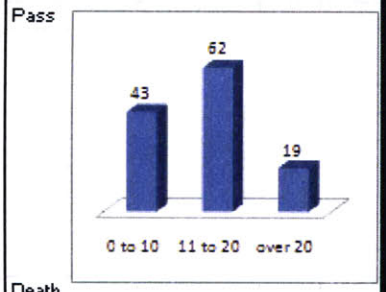


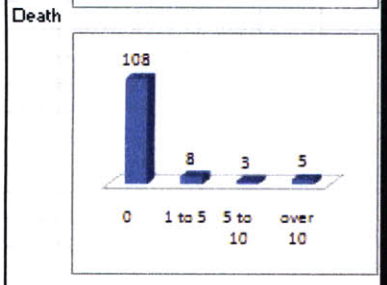


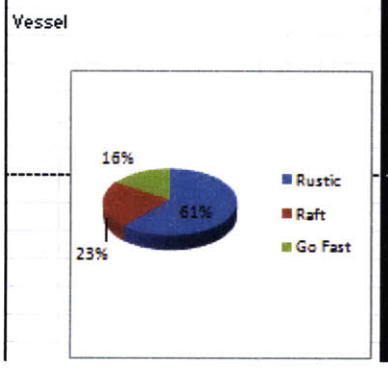


Cluster 4	Angle	2005	2006	2007
Long/Lat	0.959			
Year	43.55	44	53	27
Pass	3.051	12.8	14.17	13.52
Death	26.38	1.636	0.962	0.296
		Rustic	Raft	GoFas
Vessel	3.427	76	28	20

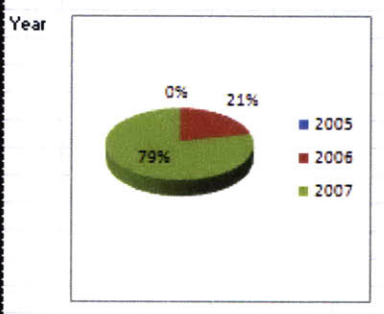


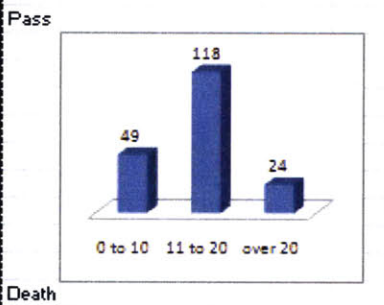


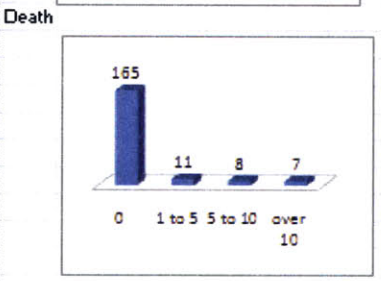


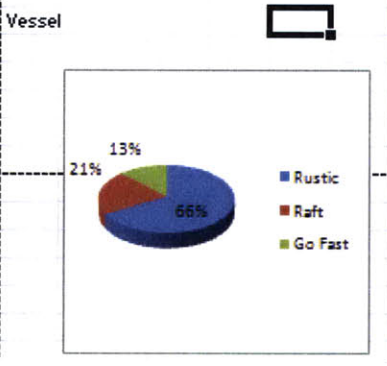


Cluster 5	Angle	2005	2006	2007
Long/Lat	2.241			
Year	13.76	0	41	150
Pass	32.3	0	12.44	14.48
Death	53.06	0	0.146	1.28
		Rustic	Raft	GoFas
Vessel	2.224	127	40	24

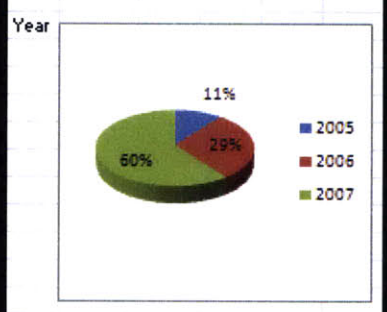


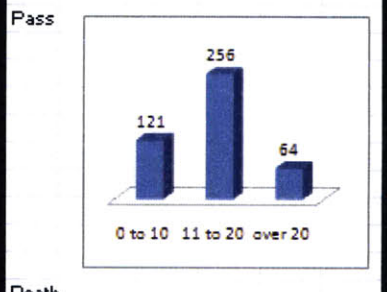


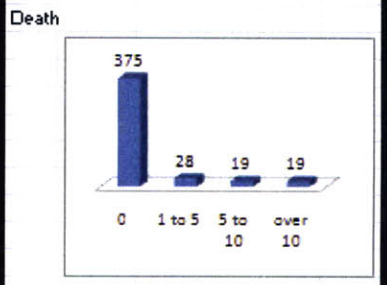


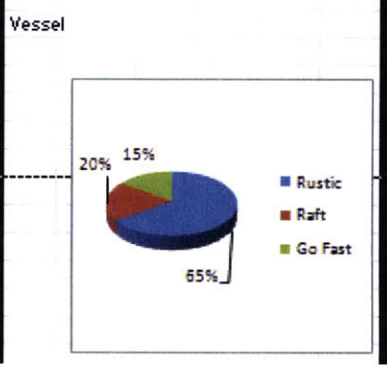


Total Data	Angle	2005	2006	2007
Long/Lat	0			
Year	0	46	129	266
Pass	0	12.57	13566	14.6
Death	0	1.739	0.767	1.308
		Rustic	Raft	GoFas
Vessel	0	287	88	66









Appendix H – Pairwise Comparisons of Clusters

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37

		cos(ang)	arc cos	degrees
Long Lat	angle	1	0	0
Year	angle	1	0	0
Num pass	angle	1	0	0
Num deaths	angle	1	0	0
Vessel	angle	1	0	0

Total Difference
0

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37
2	-82.73	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
			0.05	0	0.95							0.649	0.14	0.22	
			0	0.25	0.75							0.75	0.05	0.2	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0.06
Year	angle	0.94714	0.33	18.7
Num pass	angle	0.66452	0.84	48.4
Num deaths	angle	0.28047	1.29	73.7
Vessel	angle	0.98963	0.14	8.26

Total Difference
149

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37
3	-82.03	26.4314	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
			0.05	0	0.95							0.649	0.14	0.22	
			0	0.51	0.49							0.612	0.27	0.12	

		cosine angle	arc cos	degrees	Total Difference
Long Lat	angle	0.99982	0.02	1.08	197
Year	angle	0.6914	0.81	46.3	
Num pass	angle	0.61648	0.91	51.9	
Num deaths	angle	0.09627	1.47	84.5	
Vessel	angle	0.97175	0.24	13.7	

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37
4	-81.06	24.7284	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
			0.05	0	0.95							0.649	0.14	0.22	
			0.35	0.43	0.22							0.613	0.23	0.16	

		cosine angle	arc cos	degrees	Total Difference
Long Lat	angle	0.9994	0.03	1.98	149
Year	angle	0.39826	1.16	66.5	
Num pass	angle	0.7599	0.71	40.5	
Num deaths	angle	0.8607	0.53	30.6	
Vessel	angle	0.98728	0.16	9.15	

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
1	-80.32	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37
5	-86.74	21.2497	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
			0.05	0	0.95							0.649	0.14	0.22	
			0	0.21	0.79							0.665	0.21	0.13	

		cosine angle	arc cos	degrees
Long Lat	angle	0.99592	0.09	5.18
Year	angle	0.96304	0.27	15.6
Num pass	angle	0.68074	0.82	47.1
Num deaths	angle	0.27923	1.29	73.8
Vessel	angle	0.98597	0.17	9.61

Total Difference
151

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
1	-80.319	27.5682	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0.06
Year	angle	0.94714	0.33	18.7
Num pass	angle	0.66452	0.84	48.4
Num deaths	angle	0.28047	1.29	73.7
Vessel	angle	0.98963	0.14	8.26

Total Difference
149

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
			0	0.27	0.81							0.811	0.05	0.22	
			0	0.25	0.75							0.75	0.05	0.2	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0
Year	angle	1	0	0
Num pass	angle	1	0	0
Num deaths	angle	1	0	0
Vessel	angle	1	0	0

Total Difference
0

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
3	-82.028	26.4314	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
			0	0.27	0.81							0.811	0.05	0.22	
			0	0.51	0.49							0.612	0.27	0.12	

		cosine angle	arc cos	degrees
Long Lat	angle	0.9998	0.02	1.14
Year	angle	0.88512	0.48	27.7
Num pass	angle	0.99708	0.08	4.38
Num deaths	angle	0.40231	1.16	66.3
Vessel	angle	0.94176	0.34	19.7

Total Difference
119

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
4	-81.06	24.7284	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
			0	0.27	0.81							0.811	0.05	0.22	
			0.35	0.43	0.22							0.613	0.23	0.16	

		cosine angle	arc cos	degrees
Long Lat	angle	0.99937	0.04	2.04
Year	angle	0.57273	0.96	55.1
Num pass	angle	0.83499	0.58	33.4
Num deaths	angle	0.18614	1.38	79.3
Vessel	angle	0.96162	0.28	15.9

Total Difference
186

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
2	-82.732	28.4938	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
5	-86.742	21.2497	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
			0	0.27	0.81							0.811	0.05	0.22	
			0	0.21	0.79							0.665	0.21	0.13	

		cosine angle	arc cos	degrees
Long Lat	angle	0.99582	0.09	5.24
Year	angle	0.99849	0.05	3.15
Num pass	angle	0.99963	0.03	1.56
Num deaths	angle	0.99878	0.05	2.83
Vessel	angle	0.96973	0.25	14.1

Total Difference
26.9

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
1	-80.319	27.568	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37

		cosine angle	arc cos	degrees
Long Lat	angle	0.9998	0.02	1.08
Year	angle	0.6914	0.81	46.3
Num pass	angle	0.6165	0.91	51.9
Num deaths	angle	0.0963	1.47	84.5
Vessel	angle	0.9717	0.24	13.7

Total Difference
197

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
2	-82.732	28.494	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
			0	0.68	0.65							0.81	0.35	0.16	
			0	0.25	0.75							0.75	0.05	0.2	

		cosine angle	arc cos	degrees
Long Lat	angle	0.9998	0.02	1.14
Year	angle	0.8851	0.48	27.7
Num pass	angle	0.9971	0.08	4.38
Num deaths	angle	0.4023	1.16	66.3
Vessel	angle	0.9418	0.34	19.7

Total Difference
119

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
			0	0.68	0.65							0.81	0.35	0.16	
			0	0.51	0.49							0.61	0.27	0.12	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0
Year	angle	1	0	0
Num pass	angle	1	0	0
Num deaths	angle	1	0	0
Vessel	angle	1	0	0

Total Difference
0

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
4	-81.06	24.728	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
			0	0.68	0.65							0.81	0.35	0.16	
			0.35	0.43	0.22							0.61	0.23	0.16	

		cosine angle	arc cos	degrees
Long Lat	angle	0.9999	0.02	0.89
Year	angle	0.7695	0.69	39.7
Num pass	angle	0.8371	0.58	33.2
Num deaths	angle	0.5234	1.02	58.4
Vessel	angle	0.9967	0.08	4.68

Total Difference
137

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
			0	0.68	0.65							0.81	0.35	0.16	
			0	0.21	0.79							0.66	0.21	0.13	

		cosine angle	arc cos	degrees
Long Lat	angle	0.9974	0.07	4.1
Year	angle	0.8582	0.54	30.9
Num pass	angle	0.9946	0.1	5.95
Num deaths	angle	0.447	1.11	63.4
Vessel	angle	0.9948	0.1	5.85

Total Difference
110

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
1	-80.32	27.57	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37

		cosine angle	arc cos	degrees
Long Lat	angle	0.999	0.03	1.98
Year	angle	0.398	1.16	66.5
Num pass	angle	0.76	0.71	40.5
Num deaths	angle	0.861	0.53	30.6
Vessel	angle	0.987	0.16	9.15

Total Difference
149

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
2	-82.73	28.49	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
			1.19	1.43	0.73							2.05	0.76	0.54	
			0	0.25	0.75							0.75	0.05	0.2	

		cosine angle	arc cos	degrees
Long Lat	angle	0.999	0.04	2.04
Year	angle	0.573	0.96	55.1
Num pass	angle	0.835	0.58	33.4
Num deaths	angle	0.186	1.38	79.3
Vessel	angle	0.962	0.28	15.9

Total Difference
186

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
3	-82.03	26.43	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
			1.19	1.43	0.73							2.05	0.76	0.54	
			0	0.51	0.49							0.61	0.27	0.12	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0.02	0.89
Year	angle	0.769	0.69	39.7
Num pass	angle	0.837	0.58	33.2
Num deaths	angle	0.523	1.02	58.4
Vessel	angle	0.997	0.08	4.68

Total Difference
137

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
			1.19	1.43	0.73							2.05	0.76	0.54	
			0.35	0.43	0.22							0.61	0.23	0.16	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0
Year	angle	1	0	0
Num pass	angle	1	0	0
Num deaths	angle	1	0	0
Vessel	angle	1	0	0

Total Difference
0

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
4	-81.06	24.73	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
5	-86.74	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
			1.19	1.43	0.73							2.05	0.76	0.54	
			0	0.21	0.79							0.66	0.21	0.13	

		cosine angle	arc cos	degrees
Long Lat	angle	0.998	0.06	3.2
Year	angle	0.541	1	57.3
Num pass	angle	0.833	0.59	33.6
Num deaths	angle	0.21	1.36	77.9
Vessel	angle	0.997	0.08	4.53

Total Difference
176

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
1	-80.319	27.568	2	0	35	7.5	0	15.3	4	0	1.17	24	5	8	37

		cosine angle	arc cos	degrees
Long Lat	angle	0.9959	0.09	5.18
Year	angle	0.963	0.27	15.6
Num pass	angle	0.6807	0.82	47.1
Num deaths	angle	0.2792	1.29	73.8
Vessel	angle	0.986	0.17	9.61

Total Difference
151

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
2	-82.732	28.494	0	10	30	0	15.4	17	0	0.2	3.1	30	2	8	40
			0	1.11	4.05							3.43	1.018	0.65	
			0	0.25	0.75							0.75	0.05	0.2	

		cosine angle	arc cos	degrees
Long Lat	angle	0.9958	0.09	5.24
Year	angle	0.9985	0.05	3.15
Num pass	angle	0.9996	0.03	1.56
Num deaths	angle	0.9988	0.05	2.83
Vessel	angle	0.9697	0.25	14.1

Total Difference
26.9

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
3	-82.028	26.431	0	25	24	0	13.4	12.7	0	1.6	0.58	30	13	6	49
			0	1.11	4.05							3.43	1.08	0.65	
			0	0.51	0.49							0.61	0.27	0.12	

		cosine angle	arc cos	degrees	Total Difference
Long Lat	angle	0.9974	0.07	4.1	110
Year	angle	0.8582	0.54	30.9	
Num pass	angle	0.9946	0.1	5.95	
Num deaths	angle	0.447	1.11	63.4	
Vessel	angle	0.9948	0.1	5.85	

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
4	-81.06	24.728	44	53	27	12.8	14.2	13.5	1.64	0.96	0.3	76	28	20	124
			0	1.11	4.05							3.43	1.08	0.65	
			0.35	0.43	0.22							0.61	0.23	0.16	

		cosine angle	arc cos	degrees	Total Difference
Long Lat	angle	0.9984	0.06	3.2	176
Year	angle	0.5409	1	57.3	
Num pass	angle	0.833	0.59	33.6	
Num deaths	angle	0.2101	1.36	77.9	
Vessel	angle	0.9969	0.08	4.53	

	Centroid		Year			Year Avg Pass			Year Avg Deaths			Vessel			Total
Cluster	Land Long	Land Lat	2005	2006	2007	2005	2006	2007	2005	2006	2007	Rustic	Raft	Go Fast	
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
5	-86.742	21.25	0	41	150	0	12.4	14.5	0	0.15	1.28	127	40	24	191
			0	1.11	4.05							3.43	1.08	0.65	
			0	0.21	0.79							0.66	0.21	0.13	

		cosine angle	arc cos	degrees
Long Lat	angle	1	0	0
Year	angle	1	0	0
Num pass	angle	1	0	0
Num deaths	angle	1	0	0
Vessel	angle	1	0	0

Total Difference
0

[This Page Intentionally Left Blank]

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 4th ed. New York, United States of America: Springer, 2001.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," 1996.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 4th ed. New York, United States of America: Springer, 2001.

[4] Trevor Hastie, Robert Tibshirani, and Friedman Jerome, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 4th ed. New York, United States of America: Springer, 2001.

[5] Joseph M. Hilbe, *Logistic Regression Models*, 1st ed. United States of America: Chapman and Hall Press, 2009.

[6] StatGun Statistics Consulting. (2007) Data Analysis and Statistics Professionals. [Online]. http://www.statgun.com/tutorials/logistic-regression.html

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 4th ed. New York, United States of America: Springer, 2001.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 4th ed. New York, United States of America: Springer, 2001.

[9] Laurynt Hyafil, "Constructing Optimal Binary Decision Trees," *Information Processing Letters*, vol. 5, no. 1, pp. 15-17, 1976.

[10] Lawarance C. Hamilton, *Statistics with STATA*, 10th ed. Canada: Cengage Learning, 2009.

[11] Jerome Friedman, Charles Stone, R.A. Olshen Leo Breiman, *Classification and Regression Trees*. Baton Rouge, United States of America: Chapman and Hall, 1998.

[12] David J Vukich, "Emergency Medicine: Coming of Age," *Jacksonville Medicine*, pp. 20-23, March 1999.

[13] Jacob Goldstein, "Emergency Rooms Visits Hit Record High," *Wall Street Journal*, p. 8, August 2008.

[14] Beth Israel Deaconess Medical Center. (2010, March) BIDMC- Emergency Medicine. [Online]. http://www.bidmc.org/CentersandDepartments/Departments/EmergencyMedicine.aspx

[15] Caitlin Loureiro, Peter Luff, Tendai Nyakurimawa, Iulia Pirvu Brian Gannon, "Beth Israel Deaconess Medical Center Emergency Department Patient Care Process," McCallum Graduate School, Bentley College, Busines Process Project 2006.

[16] Clay Noyes, "Analysis and Optimization of the Emergency Department at Beth Israel Deaconess Medical Center via Simulation," Draper Laboratory, Cambridge, Thesis 2006.

[17] Lei Lei, Junjie Yin, Wei Zhang, Wu Naijun, Elia El-Darzi Peng Liu, "Healthcare Data Mining:Prediction Inpatient Length of Stay," in *3rd International IEEE Conference Intelligent Systems*, 2006, pp. 261-266.

[18] Thomas Ryan, Francis Hegarty, Neil O'Hare Michael Rowan, "The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors," *Artificial Intelligence in Medicine*, vol. 40, no. 3, pp. 211-221, July 2007.

[19] Raj Gopalan, John P Kichak, and Matthew Hager, "Artificial Neural Networks to Predict Patient's Length of Stay," *Medinfo 2007: Proceedings of the 12th World Congress on Health* , vol. 129, pp. 1825-1827, 2007.

[20] Duy-Tien Nguyen, Marinus J Eijkemans, Ewout W Steyerberg, Hugo W Tilanus, Diederik Gommers, Gerhard Wullink, Jan Bakker, Geert Kazemier Mark Van Houdenhoven, "Optimizing intensive care capacity using individual length-of-stay prediction models," *The Critical Care Forum*, vol. 11, no. 2, March 2007.

[21] Dong Won Cho, Jane Christman,John G. Csernansky, Dale A. Huntley, "Predicting Length of Stay in an Acute Psychiatric Hospital ," *American Psychiatric Association*, vol. 49, pp. 1049-1053, August 1998.

[22] Constantine E. Anagnostopoulosab, Daniel G. Swistela, Joseph J. DeRose Jra Ioannis K. Toumpoulisab, "Does EuroSCORE predict length of stay and specific postoperative complications after cardiac surgery?," *European Journal of Cardio-Thoracic Surgery*, vol. 27, no. 1, pp. 128-123, January 2005.

[23] RL Bruno, R Zorowitz,J Walker T Galski, "Predicting length of stay, functional outcome, and aftercare in the rehabilitation of stroke patients. The dominant role of higher-order cognition ," *Stroke*, vol. 24, pp. 1794-1800, 1993.

[24] Paul Jen-Hwa Hu Tsang-Hsiang Cheng, "A data-driven approach to manage the length of stay for appendectomy patients," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, no. 6, pp. 1339-1347, November 2009.

[25] Jesse Wren, Ian Jones, Clare Bates Congdon, and Dominik Aronsky, "Estimating Patient's Length of Stay in the Emergency Department with an Artificial Neural Network," in *AMIA 2005 Symposium Proceedings*, 2005, p. 1155.

[26] IEEE. (2009, October) IEEE Symposium on Visual Analytics Science and Technology. [Online]. http://vis.computer.org/VisWeek2009/vast/

[27] IEEE. (2009, October) IEEE Symposium on Visual Analytics Science and Technology. [Online]. http://vis.computer.org/VisWeek2009/vast/

[28] Palantir Technologies Inc. (2010, March) Palantir. [Online]. http://www.palantirtech.com/

[29] NIST. (2008, November) VAST Challenge 2008. [Online]. http://vac.nist.gov/challenge2008.html

[30] Guenther Walther, Trevor Hastie Robert Tibshirani, "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *J.R. Statist. Soc*, vol. 63, pp. 411-423, 2001.

[31] Yong-Woon PARK, Dong-Jo PARK Do-Jong KIM, "A Novel Validity Index for Determination of the Optimal Number of Clusters," *Iece Trans inf. & Syst.*, vol. 84, no. 2, pp. 281-285, February 2001.

[32] Deniz Yuret Ergun Bicici, "Locally Scalled Density Based Clustering," Koc University, Istanbul, Turkey, 2007.

[33] Kenneth Hild, Deniz Erdogmus, Jose Pricipe, Torbjorn Eltoft Robert Jenssen, "Clustering Using Renyi's Entropy," University of Florida, Gainesville, 2003.

[34] Pang-Ning Tan, Steinbach Michael, and Kumar Vipin, *Introduction to Data Mining*, 2nd ed. New York, United States of America: Addison Wesley, 2006.