

2014

Probabilistic Model To Identify Movement Patterns In Geospatial Data

Jeffrey Nii Anu

North Carolina Agricultural and Technical State University

Follow this and additional works at: <https://digital.library.ncat.edu/theses>

Recommended Citation

Anu, Jeffrey Nii, "Probabilistic Model To Identify Movement Patterns In Geospatial Data" (2014). *Theses*. 151.

<https://digital.library.ncat.edu/theses/151>

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at Aggie Digital Collections and Scholarship. It has been accepted for inclusion in Theses by an authorized administrator of Aggie Digital Collections and Scholarship. For more information, please contact iyanna@ncat.edu.

Probabilistic Model to Identify Movement Patterns in Geospatial Data

Jeffrey Nii Northey Anu

North Carolina A&T State University

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Computer Systems Technology

Major: Information Technology

Major Professor: Dr. Rajeev Agrawal

Greensboro, North Carolina

2014

The Graduate School
North Carolina Agricultural and Technical State University

This is to certify that the Master's Thesis of

Jeffrey Nii Nortey Anu

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2014

Approved by:

Dr. Rajeev Agrawal
Major Professor

Dr. Sambit Bhattacharya
Committee Member

Dr. Clay Gloster, Jr.
Committee Member

Dr. Clay Gloster, Jr.
Department Chair

Dr. Sanjiv Sarin
Dean, The Graduate School

© Copyright by
Jeffrey Nii Nortey Anu
2014

Biographical Sketch

Jeffrey Nii Nortey Anu pursued and completed his Higher National Diploma (HND) in Electrical/Electronic Engineering from Accra Polytechnic Ghana in 2004. After relocating to the United States in 2008, he enrolled at North Carolina Agricultural and Technical State University (A&T) in 2010 and earned his Bachelor of Science degree in Information Technology in August 2012.

During the fall semester of 2012, he joined the School of Technology graduate program in Information Technology at North Carolina Agricultural and Technical State University. While pursuing his graduate studies, he worked as a graduate teaching assistant for computer systems technology department. He also worked as research assistant for the EGAGE 2BE engineers program.

Jeffrey Nii Nortey Anu has a research paper accepted and published for the 2014 IEEE Southeast Conference held in Lexington, Kentucky. He also won first place for his poster presentation on ranking tourist attractions using time series GPS data of cabs during the Ronald E. McNair 28th Annual Celebration and 12th Research Symposium.

Dedication

I would like to dedicate my thesis work to my beloved wife Maame Adwoa Anu and children (Jeffrey and Jonathan) for all the times they missed a husband and father.

Acknowledgements

It's not by my might or power that I have been able to come this far and for that I give thanks to the Almighty God. I would like to acknowledge the vital role Dr. Rajeev Agrawal played during my studies. Not only was he my committee chair and academic advisor; he acted as a teacher setting a high standard of expectation and leading me to venture into areas of studies I felt uncomfortable in. He made it a point to avail himself irrespective of his location and taking time to explain himself to me.

I would like to thank Dr. Sambit Bhattacharya of the Department of Mathematics and Computer Science, Fayetteville State University for being another force influencing and guiding my decisions throughout my research paper. He readily made available scripts and program that made easy otherwise tedious task. Additional thanks also go to Dr. Clay Gloster, Jr (Professor and Chair Department of Computer Systems Technology) for his thought provoking questions that inquire more information about my work.

To my fellow thesis mates Muhammad Suleiman, Yolanda Baker and Curtis Jackson, thank you for your reviews, suggestions and information. Many thanks also goes to my work colleagues Michael Webster, Ashley Mansfield, Renee Baker, Shaun Finney, Tremaine McInnis Christine Kuehl, Kameisha Allen and Ricardo Villanueva at BioLife Plasma Services for helping me make time for my studies. And to Maame Anu and Yolanda Baker thank you for reviewing my work.

Table of Contents

List of Figures	ix
List of Tables	xi
Abstract	1
CHAPTER 1 Introduction.....	2
1.1 Introduction.....	2
1.2 Motivation and Problem Statement	3
CHAPTER 2 Literature Review	4
2.1 Machine Learning Algorithm	4
2.2 Hidden Markov Model	5
2.3 Taxonomy of Patterns.....	5
2.3.1 Visualization of patterns.....	8
2.3.2 Summary taxonomy of patterns.....	11
2.4 Clustering Algorithms	16
2.4.1 K-means.....	16
2.4.2 Density-based spatial clustering of applications with noise (DBSCAN).....	18
2.4.3 Affinity propagation.....	18
CHAPTER 3 Methodology.....	20
3.1 Proposed Approach to Ranking Tourist Attractions.....	20
3.1.1 Yellow cab GPS dataset	21
3.1.2 San Francisco landmarks.....	21
3.1.3 Data Cleansing.....	21
3.2 Data Description	23
3.3 Determining Distance using Haversine Formula.....	24

3.4 Approach Using Probability Model to Predict Destination	25
3.4.1 Definitions of terms	26
3.4.2 Data Description	26
3.4.3 Conditional Probability.	27
CHAPTER 4 Experiments and Results.....	29
4.1 Ranking Tourist Attractions Using Time Series GPS Data of Cabs	29
4.1.1 All GPS Coordinates (Experiment 1).....	29
4.1.2 Pickup and Drop-off coordinate of GPS (Experiment 2).....	29
4.1.3 Plotting of landmarks on a map.....	30
4.1.4 Results of Experiment 1.	32
4.1.5 Results of Experiment 2.	34
4.1.6 Summary of findings for ranking tourist attractions using time series GPS data of cabs	36
4.2 Probability Model to Predict Destination	38
4.2.1 Experiment 1 probability calculation and results	38
4.2.2 Experiment 2 probability calculation and results.....	44
4.2.3 Experiment 3 probability calculation and results.....	50
CHAPTER 5 Discussion and Future Research.....	55
References.....	56

List of Figures

Figure 1. This figure displays the classification of some common movement patterns	6
Figure 2. Normalized sample trajectory of eye movement.....	7
Figure 3. Sample trajectory of human movement (pedestrian).....	7
Figure 4. K-means clustering on the digits dataset (PCA-reduced data) Centroids are marked with white cross	17
Figure 5. This image is a graphical representation of DBSCAN.....	18
Figure 6. This image is a graphical representation of the concept of Affinity Propagation	19
Figure 7. Logical work flow map of work done	20
Figure 8. A snapshot of a cab text file named new_abboip	21
Figure 9. A snapshot of cab file new_abboip with the order of data reversed	22
Figure 10. A typical representation of the earth	24
Figure 11. GPS Visualizer input form	31
Figure 12. A visual of the landmarks in the San Francisco Bay area differentiated by colors.....	32
Figure 13. Horizontal bar graph representation of results from experiment 1.....	33
Figure 14. Horizontal bar graph representation of results from experiment 2.....	35
Figure 15. A map showing the varying sizes of point representing the landmarks	37
Figure 16. A map showing the numerical representation of the landmarks	37
Figure 17. A representation of routes generated with colored route for experiment 1	39
Figure 18. A graph showing probability verses route ID and cell sequence	43
Figure 19. Representation of routes generated with colored route for experiment 2	45
Figure 20. A graph showing probability verses route ID and cell sequence in experiment 2	49
Figure 21. Satellite image showing a section of downtown Greensboro with four routes	50

Figure 22. A grid representation of the satellite image with the four routes 51

Figure 23. A graph showing probability verses route ID and cell sequence in experiment 3 54

List of Tables

Table 1	Examples some commonly visualized patterns	8
Table 2	List of the 25 landmarks with their GPS coordinates	23
Table 3	Results of ranking landmarks using all GPS coordinates	34
Table 4	Results of ranking landmarks using start and end coordinates	36
Table 5	Observation table of route number, cell sequence and frequency experiment 1	38
Table 6	A grid table showing the outcome of possible events in experiment 1	40
Table 7	Probability table showing outcome of 12 events	41
Table 8	Observation table showing route number, cell sequence and frequency experiment 2..	44
Table 9	A grid table showing the outcome of possible events in experiment 2	46
Table 10	Probability table showing outcome of 17 events	47
Table 11	Observation table of route numbers, cell sequence and frequency experiment 3.....	52
Table 12	A grid table showing the outcome of possible events in experiment 3	53
Table 13	Probability table showing outcome of 4 events	53

Abstract

The task of trying to determine the movement pattern of objects based on available databases is a daunting one. Tracking the movement of these dynamic objects is important in different areas to understand the higher order patterns of movement that carry special meaning for a target application. However this is still a largely unsolved problem and recent work has focused on the relationships of moving point objects with stationary objects or landmarks on a map.

Global Position System (GPS) is a widely used satellite-based navigation system. Popular use of these devices has produced large collections of data, some of which have been archived. These archived data sets and sometimes real time GPS data are now readily available over the internet and their analysis through computational methods can generate meaningful insights. These insights when applied appropriately can be used in everyday life. The purpose of this research is to make the case that automated analysis can provide insight that can otherwise be difficult to achieve due to the large volume and noisy characteristics of GPS data. We present experiments that have been performed on one of these archived databases which contain GPS traces of 536 yellow cabs in the San Francisco Bay area. Using data analysis, we determine the most visited tourist destinations within the San Francisco Bay area during the time period of the captured data.

We also propose a probabilistic framework, which determines the probability of a new routing pattern using previous patterns. We use simulated routing patterns built on the same data format as that of the San Francisco cab data to predict the possible routes to be taken by a vehicle. All the probability calculations performed are done using Bayes' theorem of conditional probability formula.

CHAPTER 1

Introduction

1.1 Introduction

From sailors using the stars at night to sojourners using compasses, humans have always searched for better navigation methods. In modern days, devices which rely on global positioning systems (GPS) have become part of our lives. We use GPS devices either as stand-alone devices or integrated into cell phones, vehicles or other electronic devices. GPS consists of three segments: the space segment which is a pattern made of solar-powered satellites orbiting the earth in orbits at an altitude of about 20,000 kilometers and beams radio signals down to our planet. This connects to the user segment (devices). There is the third and last segment which is the control segment that maintains the satellites in orbit around the Earth (Gray, 2013).

In this research we are taking a closer look at archived GPS data of cabs. This GPS data is a collection of 536 Yellow cabs in the San Francisco Bay area. Our proposed work creates application specific classification of patterns of moving objects to static landmarks using the San Francisco cab datasets. We created algorithms which computed principal sub-trajectory patterns from our training data. Using the working information extracted from our GPS dataset, we considered the most visited landmark within the San Francisco Bay area. We also predicted the route most likely to be followed by a vehicle judging from a probability of the usage of that route in the past.

The rest of this thesis work is organized as follows. Section 1.2 discusses the motivation and problem statement our research. Chapter 2 gives a background to related work and summary of literature reviews. Chapter 3 describes the methodology used in the research. This includes designing the studies, describing the dataset used for the experiment and the actual experiments

performed. Chapter 4 talks about the experiments performed and outcome of those experiments. Chapter 5 ends the paper with discussions and future related work.

1.2 Motivation and Problem Statement

Big data is continuously generated from web data, e-commerce applications, retail purchase histories and bank transactions. Each and every one of us is constantly producing and releasing data about ourselves through our everyday activities. But the amount of data being generated keeps on increasing and a full 90 percent of all the data in the world occurred over the last two years (Brandtzæg, 2013).

In reality, a large part of these data sit ideally over the internet in cooperate vaults and learning institutions without its full potential being exploited due to its sheer enormous size. Apart from their primary use, what can we do with all these enormous data sitting pointlessly? In our effort to gather worthwhile information from these data, we found one of such archived data made up of a collection of GPS data of 536 yellow cabs in the San Francisco Bay area gathered for 30 days.

This dataset and other mobility datasets are readily available at Crowdad, a community resource used for archiving wireless data at Dartmouth College. We analyze this GPS dataset and used the information gathered to help identify movement patterns in geospatial data. Analysis made from this research might not necessarily be a determinant for identifying movement patterns in geospatial data, but the results will help in taking a closer look at such datasets and deriving some meaningful conclusions from them.

CHAPTER 2

Literature Review

2.1 Machine Learning Algorithm

Machine learning algorithm is also known in some circles as predictive analytics or data mining is a fast developing field which seeks to help do task by generalizing from examples. Machine learning algorithm thrives on the availability of enormous dataset to give exact or near perfect predictions. The dataset used in machine learning algorithm is relative to the problem at hand. The data are in form of pictures, text, symbols or patterns. Machine learning algorithms can be said to work by recalling information, events, or experiences, although this assumption is not necessarily always the case.

Suppose there is an application that you think machine learning might be good for. The first problem facing you is the variety of learning algorithms available and the one to use. There are literally thousands available, and hundred more published each year. The key to not getting lost in this huge space is to realize that it consists of combinations of just three components. The components are: (a) Representation - Choosing a representation for a learner is tantamount to choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space of the learner. If a classifier is not in the hypothesis space, it cannot be learned, (b) Evaluation. An evaluation function (also called objective function or scoring function) is needed to distinguish good classifiers from bad ones.

The evaluation function used internally by the algorithm may differ from the external one that is required of the classifier to optimize, (c) Optimization - Finally, we need a method to search among the classifiers in the language for the highest-scoring one (Domingos, 2012).

2.2 Hidden Markov Model

Hidden Markov models (HMMs) are the building blocks of computational sequence analysis. They are a formal foundation for making probabilistic models of linear sequence problem and offer a conceptual toolkit for building complex models just by drawing an intuitive picture. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition (Roger, 2008), part-of-speech tagging, musical score following (Birmingham, 2005), partial discharges (Satish & Gururaj, 1993) and bioinformatics.

Analyses of hidden Markov models seek to recover the sequences of states from the observed data. With HMM being a probabilistic model, it uses Bayesian probability theory to manipulate numbers and limits, interpreting the significance of the outcome. Upon describing a system as hidden Markov Model, one can begin to work on pattern recognition problems: Finding the probability of an observed sequence given a HMM (evaluation); and finding the sequence of hidden states that most probably generated an observed sequence (decoding). Another problem that can be worked on is generating a HMM given a sequence of observations (Roger, 2008).

2.3 Taxonomy of Patterns

There have been many contributions to the developing of data mining algorithms and visual analytic techniques for movement analysis. Conceptual frameworks of movement behavior for different moving objects have been developed with comprehensive classification for those movement patterns.

There are many advantages to classifying some of the most common movement patterns. Movement classification may be grouped into generic and behavioral patterns. These classified

movement patterns were then represented graphically in co-location plane and the x-y plane. The generic patterns identified in the classification allow a domain-independent visualization of movement. This makes it easy for researchers from various disciplines, because these generic patterns are applicable to all moving datasets at all spatio-temporal resolutions (Dodge, Weibel, & Lautenschütz, 2008).

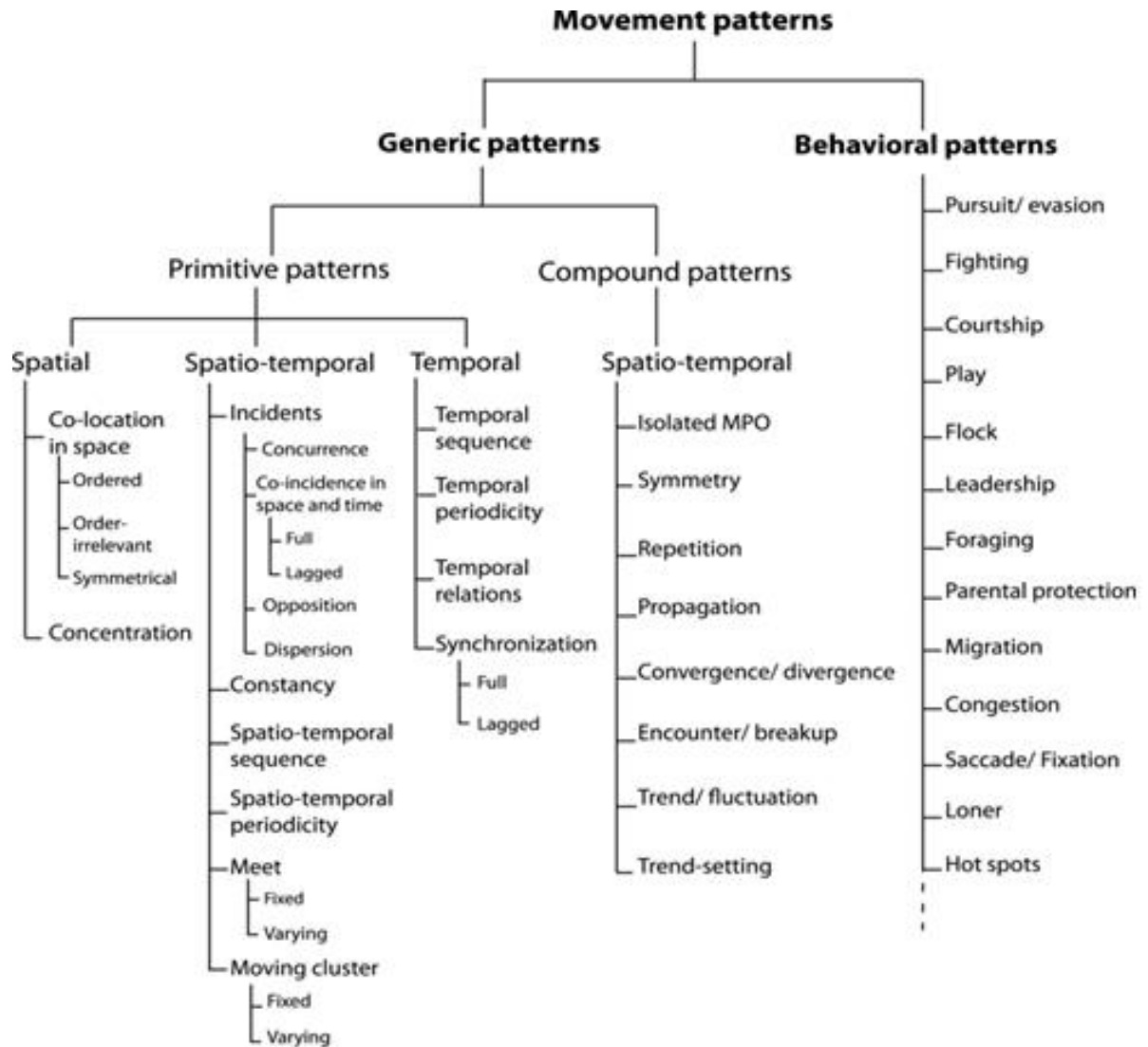


Figure 1. This figure displays the classification of some common movement patterns

With these movements have been classified, it might not necessary be true when used in classifying huge data sets or certain kinds of movements. One of such an example is the use of

eye movement data as a proxy for other kinds of MPO data. This data will not generate conclusive results when used within certain research since eye movement is different from an entire human body motion.

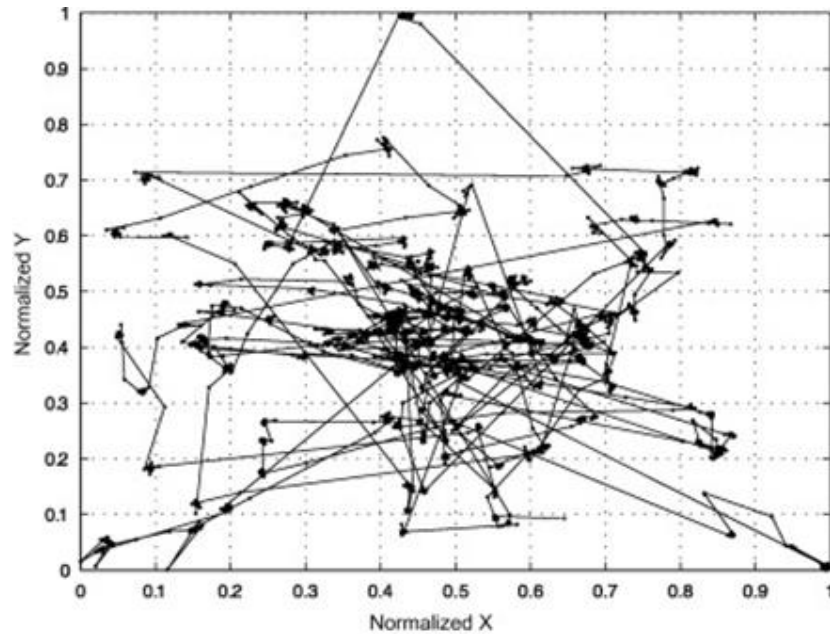


Figure 2. Normalized sample trajectory of eye movement

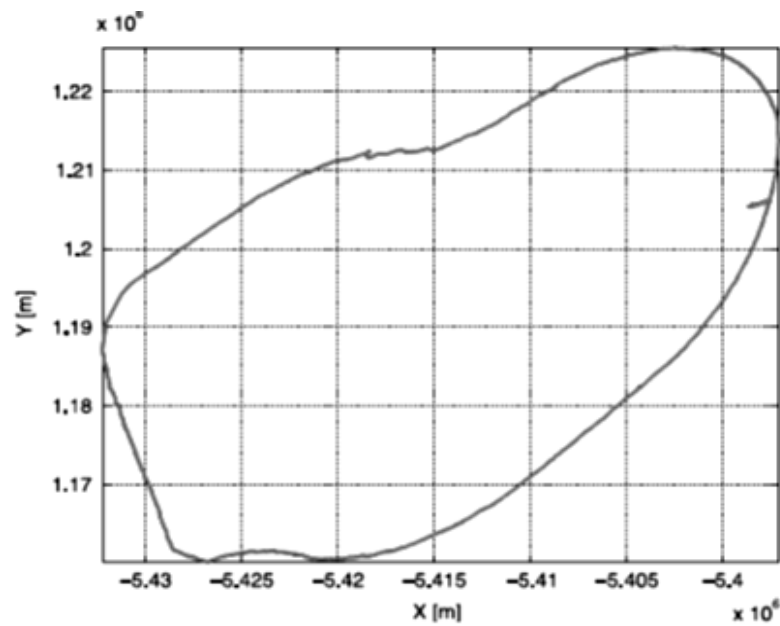


Figure 3. Sample trajectory of human movement (pedestrian)

2.3.1 Visualization of patterns. Getting a visual perspective of the various movement patterns help in understanding them. This section of our research shows visual of common taxonomy of patterns.

Table 1

Examples some commonly visualized patterns

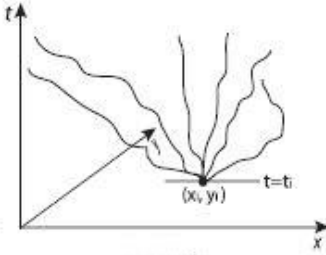
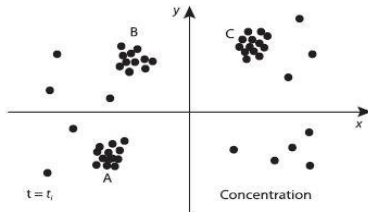
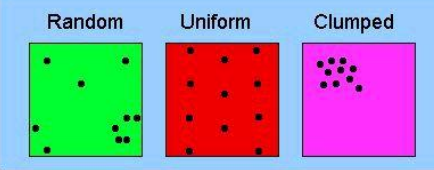
Visualized taxonomy of patterns	Description
	<p>Breakup- This is a divergence pattern where the objects move at the same time from a scene. An example will be ducks fleeing from a pond (Dodge et al., 2008).</p>
	<p>Concentration - This represents spots of high spatial density. Congestion and fixation patterns represent forms of concentration. An example will be spots of concentrations formed as a result of traffic jams (stationary clusters) (Dodge et al., 2008).</p>
	<p>Dispersion - Random dispersion is a pattern which occurs if no special forces are acting on the spatial distribution of people in a population. Uniform dispersion can result from behavior or Allelopathy. Clumped distributions can result from an aggregative behavior or from restricted availability of suitable habitat or microhabitat (McDonald, 2013).</p>

Table 1

Cont.

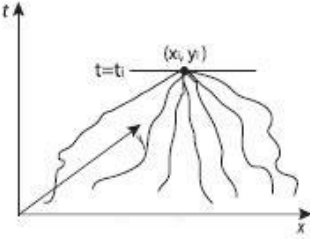
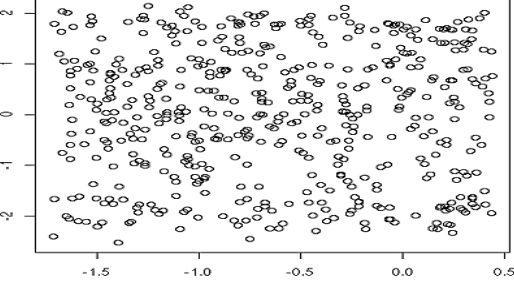
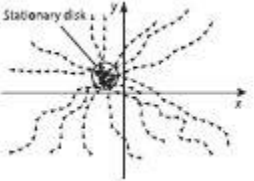
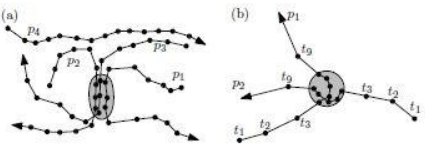
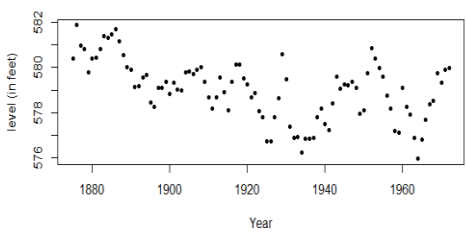
	<p>Encounter – This is a convergence pattern and opposite of breakup. Here the objects arrive at the same time. An example will be a passengers arriving at an airport gate after landing (Dodge et al., 2008).</p>
	<p>Fighting – Fighting is a combination of pursuit and evasion, attack and defense. Very high-speed movements are combined with large amounts of tightly intertwined turning, looping and frequent contact (where trajectories meet) in small distance between objects (Dodge et al., 2008).</p>
	<p>Fixed Meet – A fixed meet pattern consists of a set of MPOs that form a stationary cluster that stay together for the same amount of duration. An example can be students attending a lecture for a certain period of time (Dodge et al., 2008).</p>
	<p>Flock - The flock pattern describes a group of animals moving in the same direction while staying close together, for instance, a flock of sheep (Dodge et al., 2008).</p>
	<p>Fluctuation – This refers to irregular changes in the movement parameters of moving objects, for example, a flock of geese may change their flying formation between V-shape, irregular V-shape, or sometimes lines (Dodge et al., 2008).</p>

Table 1

Cont.

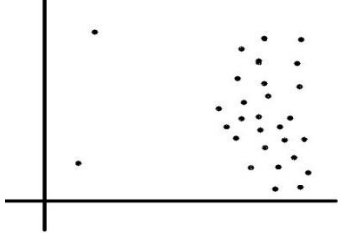
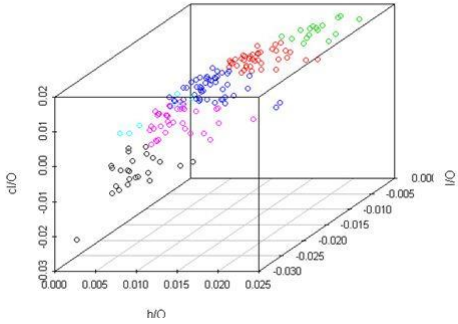
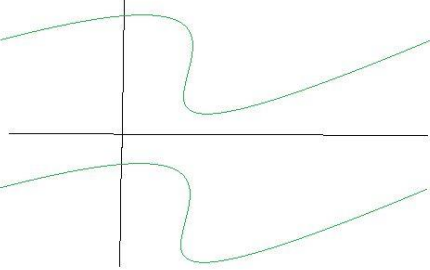
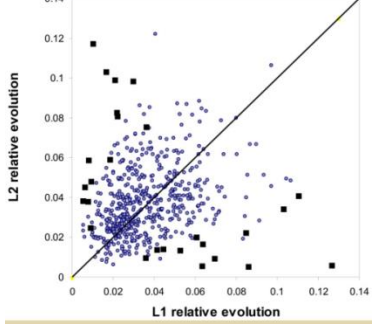
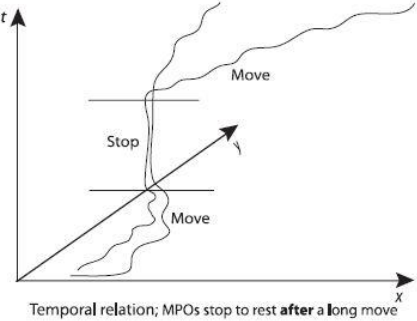
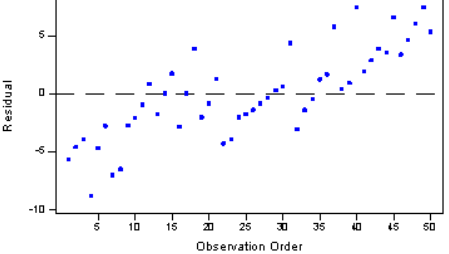
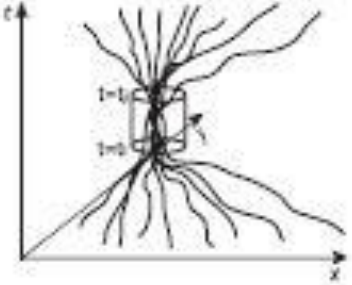
	<p>Loner – This movement describes objects moving in the same direction closely together at the same interval but with few individuals moving at a different interval. An example will be a herd of cattle grazing together at the same spot and one cow grazing at a different spot.</p>
	<p>Mixed cluster – This movement pattern is characterized by different objects moving at the same time. An example is the migration of Sandhill and Whooping Cranes that use the Central Platte River Valley in Nebraska as a staging habitat during their migration north to breeding (Birds, 2007).</p>
	<p>Repetition – This refers to the occurrence of the same patterns or pattern sequence at different time intervals. For instance, in a football match the wingers may repeatedly sprint along the sidelines (Dodge et al., 2008).</p>
	<p>Symmetrical – This refers to sequences of patterns, where the same patterns are arranged in reverse order, such as wild geese heading north in the spring, and south in the fall (Dodge et al., 2008).</p>

Table 1

Cont.

 <p>Temporal relation; MPOs stop to rest after a long move</p>	<p>Temporal relation - These include any temporal relation among various events on the time axis. An instance will be a flock of wild geese usually stopping to rest after a long continuous flight (Dodge et al., 2008).</p>
	<p>Trend - Trend refers to consistent changes in the movement parameters of moving objects. For example, for an airplane circling in a holding pattern the rate of change of the movement direction will remain constant (Dodge et al., 2008).</p>
	<p>Varying meeting - A fixed meet pattern consists of a set of MPOs that form a stationary cluster that change in the meeting region. An example for a varying meet is the rental car drop-off at an airport (Dodge et al., 2008).</p>

2.3.2 Summary taxonomy of patterns. Being able to decide the moving patterns of people and classify them has been an area of interest to both the academic and industrial world. Regarding classifying movement patterns, Somayeh, Robert, and Anna-Katharina (2008) acknowledged in their research paper that there is little agreement on the relevant types of movement patterns and if any, only few isolated definitions of these exist. Their paper intended

to contribute to the development of a toolbox of data mining algorithms and visual analytic techniques for movement analysis by developing first a conceptual framework for movement behavior of different moving objects and secondly a comprehensive classification and review of movement patterns. They show the utilization of their classification by answering the question about the extent of movement patterns of the eye tracking data is a proxy of other types of movement data. They set up a moderated discussion platform to help the further evolution of their proposed classification towards a consolidated taxonomy in a consensus process. Their research was able to use the generic patterns identified in their classification to allow a domain-independent visualization of movement (Dodge et al., 2008).

In reporting leadership patterns among trajectories, spatio-temporal movement patterns in large tracking data sets were investigated. Their paper presented several algorithms for computing patterns which were analyzed both theoretically and experimentally. Using a modified NetLogo Flocking Model, they were able to generate trajectories as a test bed for their pattern detection algorithms. Their paper presented a formal notion of a pattern called 'leadership', describing the event or process of one individual in front leading the movement of a group. Their approach was inspired by movement patterns documented in animal behavior and behavioral ecology literature. One drawback they realized was the overall challenge which lies in relating movement patterns with the surrounding environment to understand where, when and ultimately why the agents move the way they do (Andersson et al., 2007).

Benkert, Gudmundsson, Hubner and Wolle (2006) used the size and movement patterns of animals to determine if they were a flock or not by developing three approximation algorithms base on the size of region the animals were in. All the trajectories which were used in their experiments were created artificially. From their experiments, they concluded that the idea of

projecting trajectories into points in higher dimensional space is very practicable for finding flocks in spatio-temporal data. For the techniques they used in their experiment, they had to relax their definition of term flocks. As a conclusion they saw that the idea of projecting trajectories into points in higher dimensional space was very viable for finding flocks in spatio-temporal data (Benkert, Gudmundsson, Hübner, & Wolle, 2008).

Also Mazzoni (2005) used his research paper to write a software library, called LibFeature, which attempts to make the process of constructing feature vectors from raw data easier by allowing one to specify the commands to produce a feature vector in a high-level language. LibFeature was written in very portable C and designed to compile and run on almost any modern computing platform, including Windows, Mac OS X, Linux, and any modern UNIX system (Mazzoni, 2005).

A proposed concept of spatio-temporal patterns as a systematic and scalable concept to query developments of objects and their relationships was undertaken by Martin Erwig. Based on his earlier work on spatio-temporal predicates, he outlined the design of spatio-temporal patterns as a query mechanism to characterize complex object behaviors in space and time. His research focused on deriving constraints that will allow spatio-temporal patterns to become well designed composable abstractions that are smoothly integrated into spatio-temporal query languages. He observed that most users of spatio-temporal data (such as, scientists) do not have a formal computer education and do not know how to use query languages for complex data like spatio-temporal data. Offering ordinary users access to spatio-temporal data is therefore becoming a more important issue that is addressed by developing a visual query language and a corresponding user interface (Erwig, 2004).

Lee, Paek and Ryu (2004) presented an innovative data mining technique for extracting temporal patterns of moving objects with spatio-temporal attributes which some industry refers to as location-based service (LBS). They used spatial operation to generalize a location of moving point, applying time constraints between locations of moving objects to make valid moving sequences. Location-based service aims to accurately identify individuals' locations and, by applying this information to various marketing and services, provide more personalized and satisfying mobile service to its users. The algorithm used to determine their moving pattern mining consisted of four stages. First, the database is arranged into object identifier and valid time. Second, using spatial operation, moving objects' location information is transformed into an area to discover significant information. Third, time constraints are imposed to extract effective moving sequence. Finally, the frequent moving patterns are extracted from the generated moving sequences. They acknowledged from their research that a more efficient pattern mining techniques are needed to be developed (Lee, 2004).

Based upon their research work, Zheng, Xie and Ma (2010) introduced a social networking service, called GeoLife which aims to understand trajectories, locations and users. And also the correlation between users and their base on the trajectories generated by the users. Using trajectories generated from GPS devices and Wi-Fi usage, they sought in their research to classify the transportation modes (walking, driving, etc.) of the users. They also used the GPS trajectories to determine interesting traveling locations and also share the life experience of the user; bridging the gap between people and their location (Zheng, Xie, & Ma, 2010).

Until a standardized guideline is developed for spatio-temporal trajectories, there are many research problems on which work is needed. Noyon, Devogele and Claramunt (2005) conducted research to explore and develop a trajectory manipulation model that supports not

only the representation of mobile trajectories, but also an intuitive data manipulation language that facilitates the underlying behavior, processes and patterns exhibited by moving points (Noyon, Devogele, & Claramunt, 2005).

Using their paper (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), they examined the attention being generated by data mining and knowledge discovery in databases by researchers, industry, and media. Their article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. It also mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field. Computational theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). Knowledge discovery in databases application areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents. Here data patterns are analyzed and patterns determined are used for future goals.

Zweig and Burges (2011) conducted a research on modeling semantics in text. Their data challenge consisted of 1,040 sentences each of which had four imposter sentences in which a single fixed word in the original sentence has been replaced by an impostor word with similar occurrence statistics. With the aid of a programming language model trained on 19th century novels, they were able to compute 30 alternative words for a given low frequency word in a sentence. Using human judgment, they then picked the four best impostor words, based on a set of provided guidelines. They performed further check on their data by running the same tests on 203 sentence completion questions from a practice SAT exam and achieve similar results (Princeton Review, 11 Practice Tests for the SAT&PSAT, and 2011 Edition). To train language

models for the SAT question task, they used 1.2 billion words of Los Angeles Times data taken from the years 1985 through 2002. Using unaffiliated human answer on a random subset of 100 questions, ninety-one percent were answered correctly (Zweig & Burges, 2011).

The research presented in this paper undertaken by Zheni, Frihida, Ghezala and Claramunt (2009) proposes an integration of the semantic dimension, inspired from the concept of time-path, within a formal representation of space-time trajectories. They introduced an algebraic model that explicitly represents a spatio-temporal trajectory (STT) as an Abstract Data Type, where a series of trajectory states is potentially observed and measured (Zheni, Frihida, Ghezala, & Claramunt, 2009).

Laube, Dennis, Forer and Walker (2007) used their research to discuss standardizations that integrated the extended set of motion descriptors within various temporal and spatial frames of reference. Their proposed lifeline context operators and standardizations are illustrated using high resolution trajectory data obtained from homing pigeons carrying miniature global positioning devices. Their paper also discusses opportunities and shortcomings of analyzing lifeline data from a Geographic Information Science perspective, specifically in the situation where three spatial dimensions are involved and where movement is largely unfettered. Using their pigeon flight data suggested that the selection of the algorithms used to compute lifeline context operators needed some care because not all algorithms are suitable for all data models or data-capture procedures (Laube & Purves, 2006).

2.4 Clustering Algorithms

2.4.1 K-means. K-means algorithm is a method of analyzing data by means of clustering or grouping the data. K-means is also referred to as Lloyd's algorithm. K-means algorithm does

the clustering of data by trying to separate samples into groups of equal variance. This algorithm requires the number of clusters to be specified.

K-means algorithm uses three basic steps to analyze data presented to it. The first step chooses the initial centroids, with the most basic method being to choose k samples from the dataset X . After initialization, K-means consists of looping between the two other steps. The first step assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid.

The difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly. The K-means algorithm aims to choose centroids C that minimize the within cluster sum of squares objective function with a dataset X with n samples.

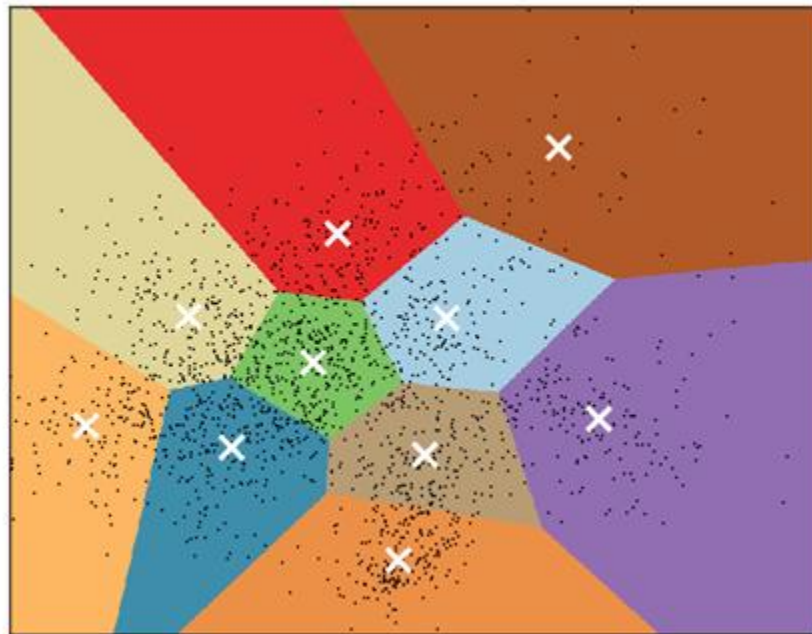


Figure 4. K-means clustering on the digits dataset (PCA-reduced data) Centroids are marked with white cross

2.4.2 Density-based spatial clustering of applications with noise (DBSCAN).

DBSCAN is a data clustering algorithm which views clusters as areas of high density separated by areas of low density. DBSCAN is commonly used for clustering in spatial database because it needs less knowledge of the input parameters of the data.

DBSCAN is use to identify arbitrary shape objects and removal of noise during the clustering process. DBSCAN has problems with handling large databases; it cannot cluster data sets well with large differences in densities. Similarly, it cannot produce correct result on varied densities.

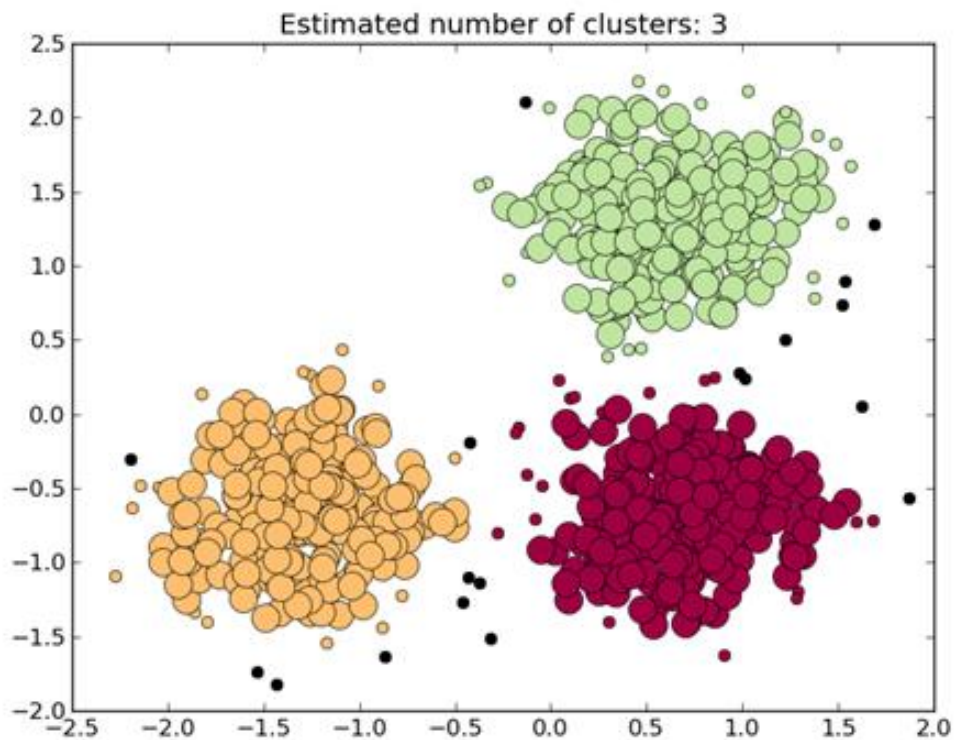


Figure 5. This image is a graphical representation of DBSCAN

2.4.3 Affinity propagation. This is an algorithm that identifies exemplar among data points and forms clusters of data points around these exemplars. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points

until a good set of exemplars and clusters emerges. Affinity Propagation creates clusters by sending messages between pairs of samples until they converge. Using two nodes as exemplars, the Affinity Propagation algorithm takes as input a collection of real-valued similarities between data points to indicate how close the two nodes are.

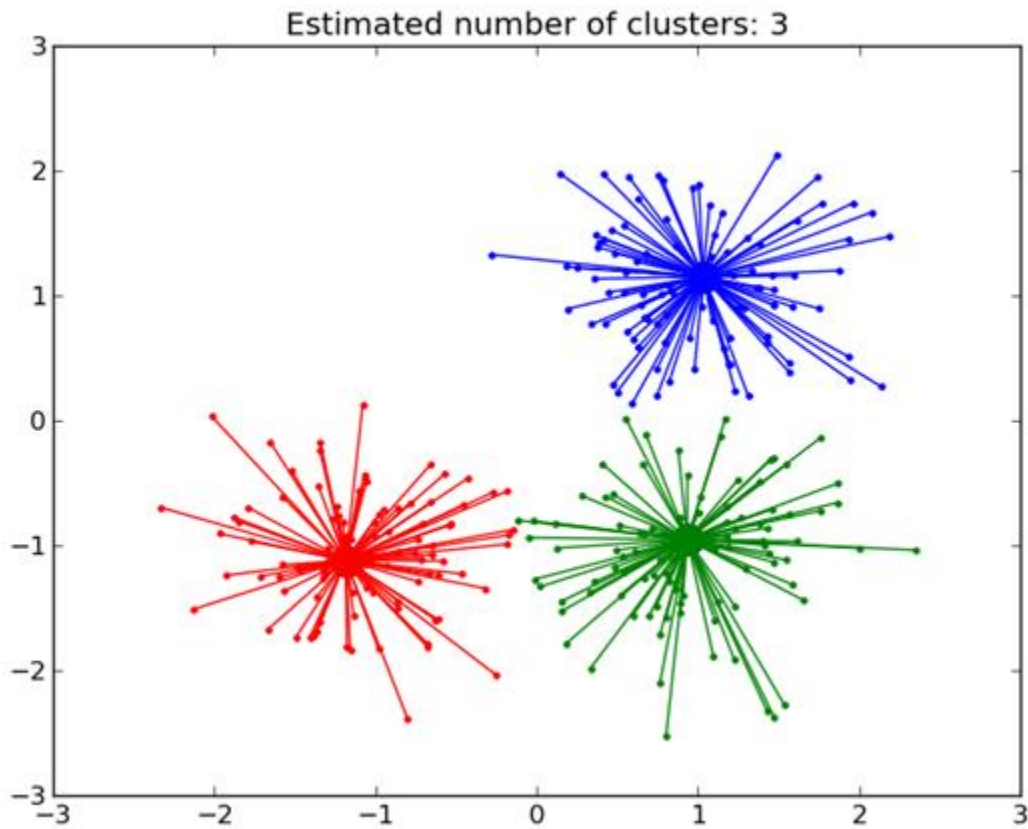


Figure 6. This image is a graphical representation of the concept of Affinity Propagation

CHAPTER 3

Methodology

For the method used in our thesis work, we used real and simulated GPS data to conduct a variety of experiments. This chapter describes the methodologies used in calculating the ranking of the most popular landmarks in the city of San Francisco and prediction of the destinations of a vehicle using the available route information followed in the past.

3.1 Proposed Approach to Ranking Tourist Attractions

Figure 7 shows the logical flow map of processes used in achieving the ranking of the tourist attractions. This logical flow map has two sections going through two processes to achieve our output. The blocks of work are described in details in the subsequent sections.

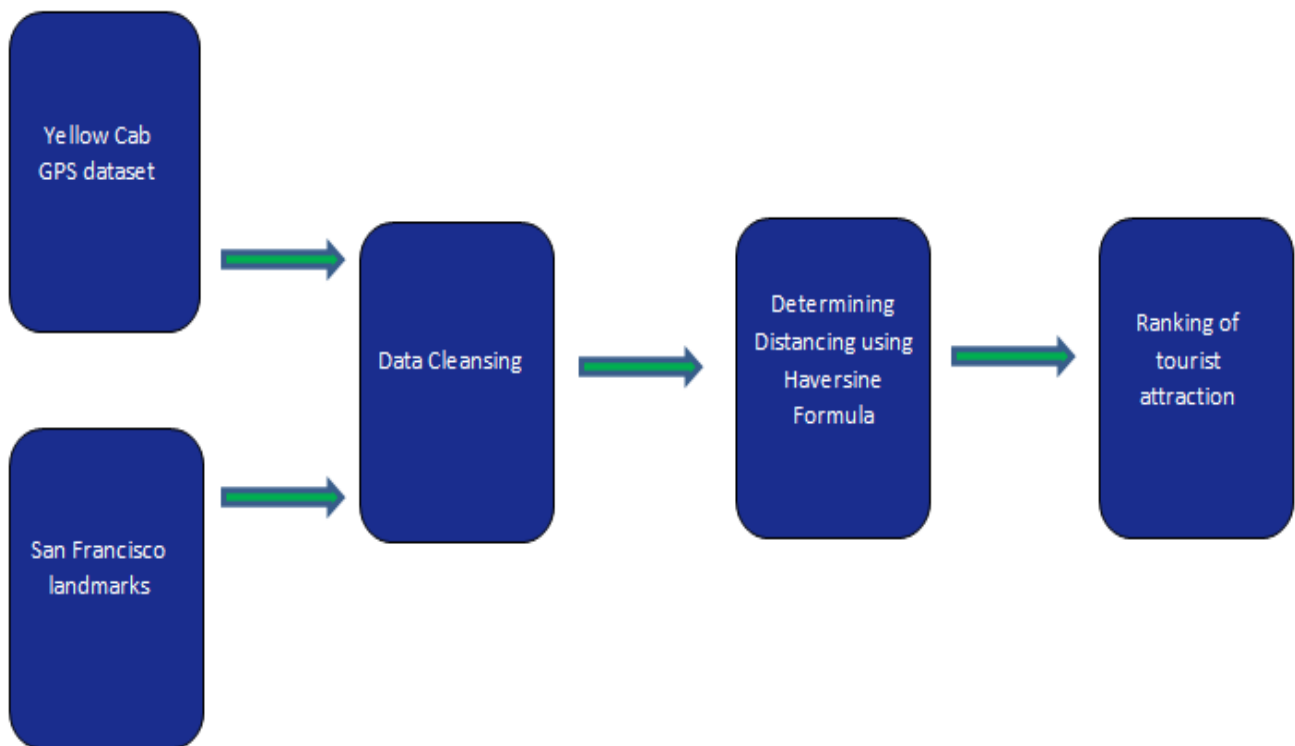
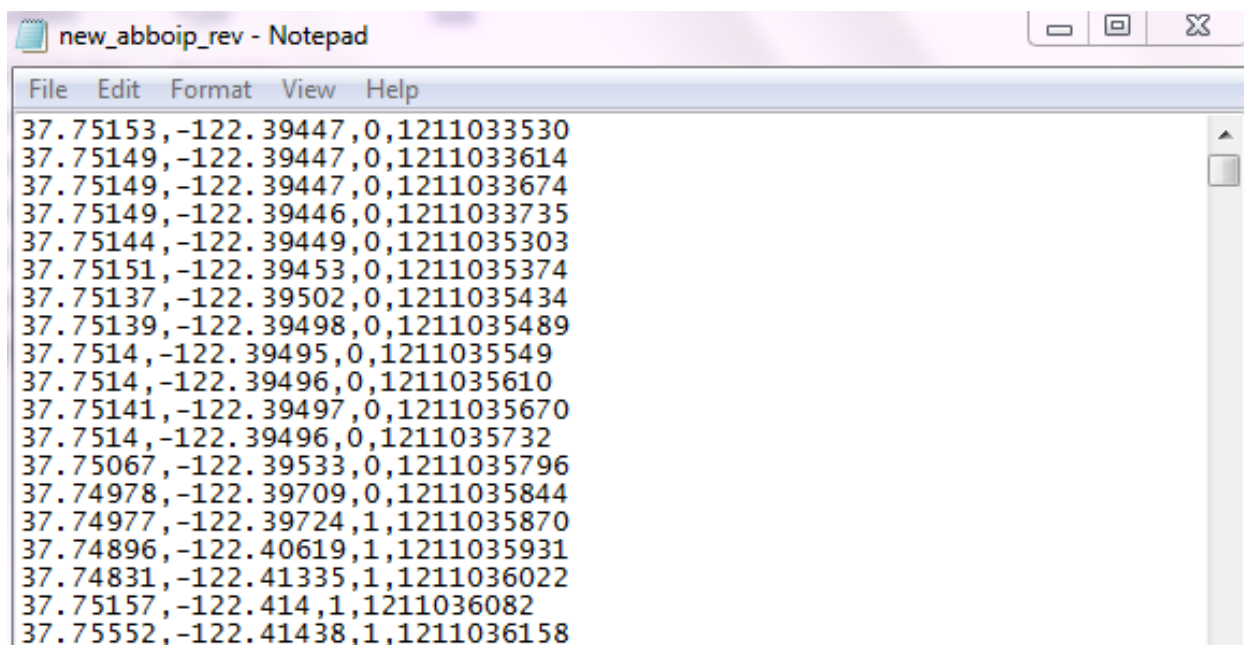


Figure 7. Logical work flow map of work done

3.1.1 Yellow cab GPS dataset. The data used was imported from CRAWDAD, a community resource for archiving wireless data at Dartmouth College (Piorkowski, Sarafijanovic-Djukic, & Grossglauser, 2009). Each file in the data set was given a unique name. For each cab its mobility trace was saved in a separate ASCII file [9]. An example snapshot of data is given below in Figure 8.



```

File Edit Format View Help
37.75153,-122.39447,0,1211033530
37.75149,-122.39447,0,1211033614
37.75149,-122.39447,0,1211033674
37.75149,-122.39446,0,1211033735
37.75144,-122.39449,0,1211035303
37.75151,-122.39453,0,1211035374
37.75137,-122.39502,0,1211035434
37.75139,-122.39498,0,1211035489
37.7514,-122.39495,0,1211035549
37.7514,-122.39496,0,1211035610
37.75141,-122.39497,0,1211035670
37.7514,-122.39496,0,1211035732
37.75067,-122.39533,0,1211035796
37.74978,-122.39709,0,1211035844
37.74977,-122.39724,1,1211035870
37.74896,-122.40619,1,1211035931
37.74831,-122.41335,1,1211036022
37.75157,-122.414,1,1211036082
37.75552,-122.41438,1,1211036158

```

Figure 8. A snapshot of a cab text file named new_abboip

3.1.2 San Francisco landmarks. These are made of twenty-five (25) popular landmarks which were sourced from the internet (Rosenbaum, 2008). We used the physical address location of these landmarks to determine their geographical coordinates in latitudes and longitudes. This was made possible by using the google map webpage that enable a known physical address of a located to be translated into geographical coordinate.

3.1.3 Data Cleansing. Taking a closer look at the cab data imported, we could not conclude if the values in the first line of any of the files represented the starting GPS coordinate captured. We were able to determine the order of the coordinates by converting the values of the

UNIX epoch time stamp assigned to each GPS coordinate. The resultant time stamp was displayed in a Coordinated Universal Time (Year: Month: Day, Hour: Minute: Second) format. From here we were able to reverse the order of the data in each cab file based on the time stamp and arrange them in ascending order based on the earliest time.

	A	B	C	D	E	F	G	H	I
1	37.75153	-122.394	0	1211033530					
2	37.75149	-122.394	0	1211033614					
3	37.75149	-122.394	0	1211033674					
4	37.75149	-122.394	0	1211033735					
5	37.75144	-122.394	0	1211035303					
6	37.75151	-122.395	0	1211035374					
7	37.75137	-122.395	0	1211035434					
8	37.75139	-122.395	0	1211035489					

Figure 9. A snapshot of cab file new_abboip with the order of data reversed

We refined the landmark dataset further to make each location distinguishable. For each location we assigned a description number, color and symbol. This task was performed to help make each landmark recognizable on the map when plotted as shown in Table 2. Having cleansed the two datasets, we measured the distance between each landmark $L_i(\phi_i, \lambda_i)$ and GPS location captured by each cab $C_i(\phi_{1a}... \phi_{1m}, \lambda_{1a}... \lambda_{1n})$ (Borg & Groenen, 2005).

Displayed in table 2 is list of the landmarks assigned with a number description, color, symbol to represent them when plotted on the map and their geographical location.

Table 2

List of the 25 landmarks with their GPS coordinates

name	disc	color	symbol	latitude	longitude
Pier 39, Angel Island State Park	0	red	circle	37.81195	-122.409353
Golden Gate Bridge	1	blue	circle	37.80999	-122.477059
Golden Gate Park	2	yellow	circle	37.7699	-122.486275
Lombard Street	3	green	circle	37.80175	-122.427185
Pier 33, Also pier for Alcatraz Island	4	cyan	circle	37.8087	-122.404961
California Academy of Sciences	5	pink	circle	37.77153	-122.466073
The de Young Museum, San Francisco Museum of Modern Art	6	purple	circle	37.77309	-122.468734
The Cable Car Museum	7	magenta	circle	37.79659	-122.411563
The Exploratorium	8	red	circle	37.80321	-122.39758
The San Francisco Giants at AT&T Park	9	blue	circle	37.78062	-122.389426
Contemporary Jewish Museum	10	yellow	circle	37.79351	-122.403889
San Francisco Symphony at Davies Symphony Hall	11	green	circle	37.77792	-122.420409
San Francisco Zoo	12	cyan	circle	37.7338	-122.503227
Aquarium of the Bay	13	pink	circle	37.80863	-122.409288
Bay Area Discovery Museum	14	purple	circle	37.8357	-122.476855
Cathedral of St Mary of the Assumption	15	magenta	circle	37.78416	-122.425271
City Hall	16	red	circle	37.7794	-122.419566
Fort Mason Center	17	blue	circle	37.806	-122.431708
Grace Cathedral	18	yellow	circle	37.7919	-122.41304
Old St Mary's Cathedral	19	green	circle	37.79273	-122.405677
San Francisco Main Library	20	cyan	circle	37.77918	-122.41583
St Boniface Catholic Church	21	pink	circle	37.78206	-122.412791
Treasure Island	22	purple	circle	37.82296	-122.370263
Beach Chalet & Park Chalet	23	magenta	circle	37.76949	-122.510284
Children's Fairyland	24	red	circle	37.80925	-122.259969

3.2 Data Description

The dataset used in this research contains GPS coordinates of cabs in San Francisco California, USA. The GPS coordinates is of 536 cabs collected over 30 days in the month of May 2008. Each of the Yellow Cab's captured in the dataset were equipped with GPS tracking device which is used by dispatchers to efficiently reach customers. The data is transmitted from

each cab to a central receiving station, and then delivered in real-time to the dispatch computers via a central server. This server system broadcasts the cab call number, location and whether the cab had a passenger or not.

The format of each mobility trace file is as follows with each line containing latitude, longitude, occupancy and time. The latitude and longitude were expressed in decimal degrees. To show if the cab was occupied, it was represented with 1 and 0 if it was not occupied. The time stamp for each coordinate was expressed in a UNIX epoch format.

3.3 Determining Distance using Haversine Formula

Navigators used logs to get round the difficulties they always had doing long-multiplication and long-division. The problem was that logarithms couldn't be used with negative numbers, considering that applying log to a negative number is a meaningless concept. Ordinary trig functions ranged over positive and negative values. The Haversine formula is an equation important in navigation, giving great-circle distances between two points on a sphere from their longitudes and latitudes.

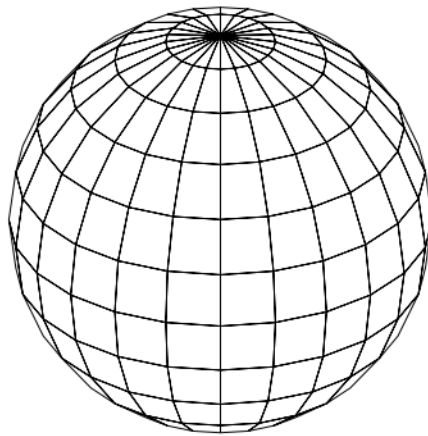


Figure 10. A typical representation of the earth

The great-circle distance is a reference between two points which is the shortest distance over the earth's surface –giving an ‘as-the-crow-flies’ distance between those points. The distance d between two points along the latitude and longitude of the earth can be calculated by;

$$\text{haversine}\left(\frac{d}{r}\right) = \text{haversin}(\vartheta_2 - \vartheta_1) + \cos(\vartheta_1) \cos(\vartheta_2) \text{haversin}(\lambda_2 - \lambda_1)$$

Where haversin is the Haversine function and h is haversin (d/r)

$$\text{haversine}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$$

$$d = r \text{haversine}^{-1}(h)$$

$$d = 2r \arcsin(\sqrt{h})$$

$$d = 2r \arcsin\left(\sqrt{\text{haversin}(\vartheta_2 - \vartheta_1) + \cos(\vartheta_1) \cos(\vartheta_2) \text{haversin}(\lambda_2 - \lambda_1)}\right)$$

$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\vartheta_2 - \vartheta_1}{2}\right) + \cos(\vartheta_1) \cos(\vartheta_2) \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Equation 1. Haversine distance formula

Where;

- r is the radius of the sphere,
- ϑ_1, ϑ_2 : latitude of point 1 and latitude of point 2
- λ_1, λ_2 : longitude of point 1 and longitude of point 2

3.4 Approach Using Probability Model to Predict Destination

Coordinates captured from GPS devices may not provide accurate results and as solution we used gridded representation of a geographical location in our approach to predicting the destination of moving bodies. One of such errors in coordinates recorded by the GPS devices for example can be observe in the reading captured within a 3000ft² area; there will be different readings generate within area even though movement within this area is not significant. We used

the Grid based subdivision of space to smooth over this GPS recording errors and also to help in the prediction of destination of route based on past recorded data of how vehicles move from cell to cell when following routes. The grid subdivision can be customized to meet the needs of the geographical location and traffic. An example will be in a densely populated city, a grid with many subdivisions will be generated; whereas a grid with fewer subdivisions will be generated for a country side area which is sparsely populated.

3.4.1 Definitions of terms. Grid – $N \times N$ cells in which geographical space is divided, each GPS point belongs to one cell. Each cell is labeled with a number (ID) in the left to right, top to bottom order as shown in example below. **Route** – trajectory prototype of which there may exist many trajectory examples. A route is simply a sequence of cell IDs. Each unique route has its own ID (again a number). **Trajectory** - a trajectory is a sequence of specific GPS coordinates time stamped. Each GPS coordinate pair falls inside a cell, so trajectory on the grid becomes a sequence of cell IDs and thus is an example of a route.

3.4.2 Data Description. The data generated was simulated using calico. A GPS location is created by clicking in a cell followed by another click for next the GPS location. Once done with a trajectory, save the data in a file by pressing the 's' key on the keyboard. The program will then request for the route ID of which this trajectory is an example.

At this point a number or sequence of digits are entered and finalize by pressing 'd' for done on the keyboard. All the trajectories are repeated in this manner and key 'Q' on the keyboard is pressed for quit. This creates a data file which is similar to the San Francisco cab data format – column 1 is X, column 2 is Y, column 3 is occupancy (always 1 for simulated data), column 4 is timestamp and column 5 is cell ID. Column 5 which represents the cell ID's is the only difference from the San Francisco cab data format.

3.4.3 Conditional Probability. Here we are asking the following types of questions – given that the vehicle was observed to have driven through path (1,2) (which is our evidence E), what is the probability that it is following route # 1 (which is our hypothesis H)? The base term (numerator) is expanded over all possible routes / route hypotheses.

Bayes' theorem is a relationship between the conditional probabilities of two events. A conditional probability, often written $P(A|B)$ is the probability that Event A will occur given that we know that Event B has occurred. Bayes' theorem states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem is often interpreted as a statement about how a body of evidence E , affects the probability of a hypothesis, H :

$$P(H|E) = P(H) \frac{P(E|H)}{P(E)}$$

This can be expanded into –

$$P(H|E) = P(H) \frac{P(E|H)}{\sum_i P(E|H_i)P(H_i)}$$

Where H_i represents all possible different hypotheses.

Conditional probabilities arise naturally in the investigation of experiments run repeatedly where an outcome of a trial may affect the outcomes of the subsequent trials.

Conditional probability is explained with the following example.

- **Example:** In a group of 160 players completing in an under 18 World Cup tournament, 7.9% of them are 17 years old, while 4.7% can expect to live to age 18. Given that a player is 17, what is the probability that the player live to age 18 before the tournament starts?

This is an example of a conditional probability. In this case, the original sample space can be thought of as a set of 160 soccer players. The events E and F are the subsets of the sample space consisting of all players who live at least 17 years, and at least 18 years, respectively. We consider E to be the new sample space, and note that F is a subset of E . Thus, the size of E is 79, and the size of F is 47. So, the probability in question equals $47/79 = 0.595$. Thus, a player who is 17 has a 0.0059 chance of reaching age 18 before the tournament is over.

CHAPTER 4

Experiments and Results

4.1 Ranking Tourist Attractions Using Time Series GPS Data of Cabs

This chapter presents a method to visualize landmarks within an urban area and also to rank them according to popularity. We used working information extracted from the cab dataset within the San Francisco Bay area for the experiments performed.

4.1.1 All GPS Coordinates (Experiment 1). In the first part of the experiments, we used the values of the GPS coordinates only during the period when the cab is occupied. By using a program, we calculated the Haversine distance of each landmark coordinate against the GPS coordinates in a particular file. The selected GPS coordinates were determined using a 0.1 mile radius distance of the GPS coordinates close to the landmark.

The same procedure used to determine the Haversine distance was applied to the remaining 536 Yellow cab files to determine their Haversine distance. A frequency count for each file was generated to find the number of visits to a particular landmark within the 0.1 mile radius from the Haversine distance calculation.

4.1.2 Pickup and Drop-off coordinate of GPS (Experiment 2). In the second set of experiment, we only used the starting and ending values of the GPS coordinate during the time period when the cab is occupied. The pickup point and its corresponding GPS coordinate were characterized with '1' to show it was occupied by a passenger. GPS coordinates assigned with '0' symbolized that the Yellow cab was not occupied by a passenger even though it might be moving.

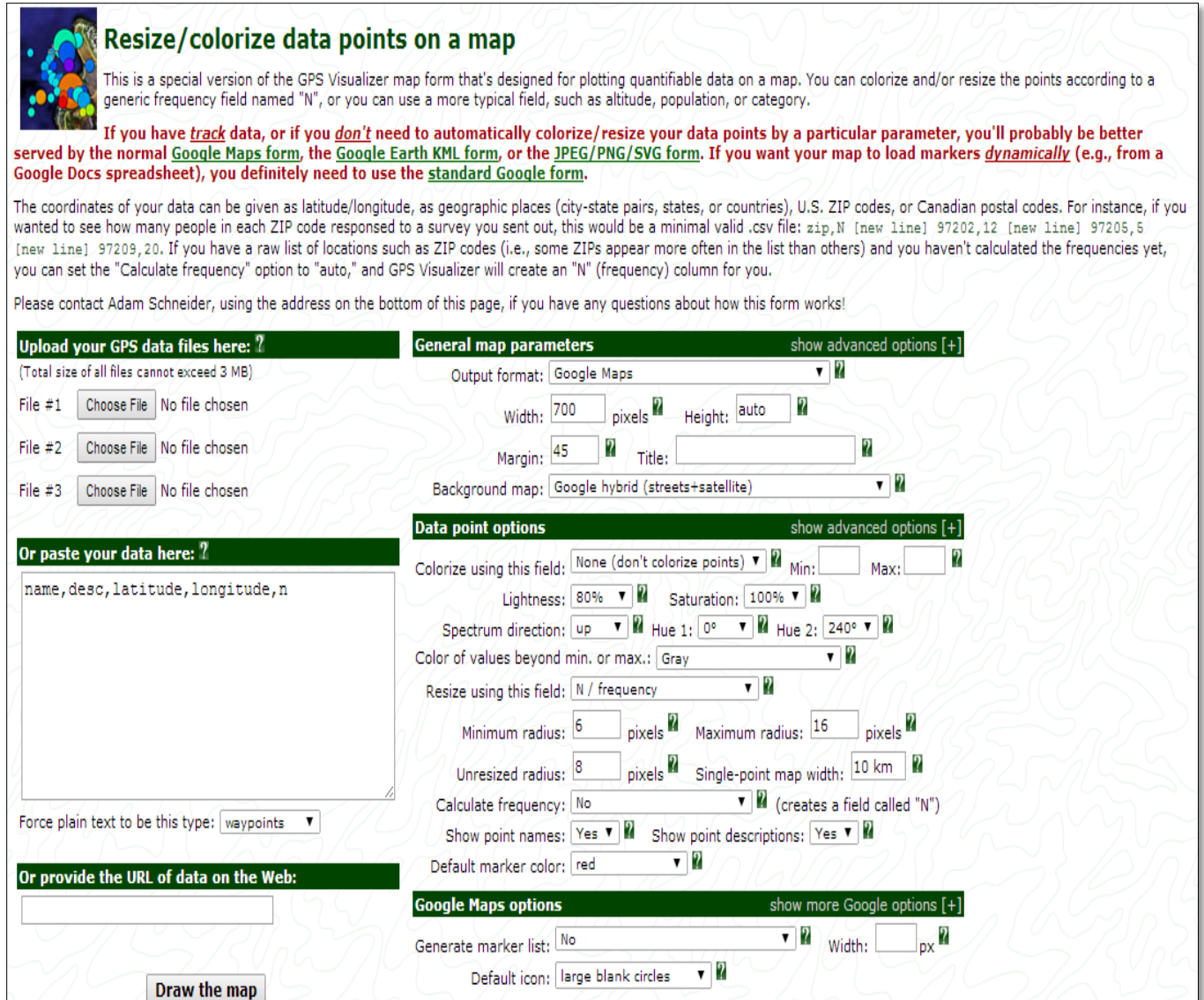
Using the same approach in experiment 1, the Haversine distance between each of the landmarks and GPS coordinates of the pickup and drop-off point was calculated for each cab file

within 0.1 mile radial. A frequency count for each file was generated to find the number of visits to a particular landmark within that 0.1 mile radius.

4.1.3 Plotting of landmarks on a map. By means of the geographical coordinates of the 25 landmarks, we plotted on a map each landmark. This was possible using a GPS visualization website that allowed files of the coordinates to be uploaded to the website (<http://www.gpsvisualizer.com/>) as shown in Figure 11. The GPS visualization website plotted each landmark assigned with a default red circular shape.

When the cursor of a mouse is placed on the landmark icon, it displays a name description of it. Figure 12 below shows a Google Ariel view of the landmarks in the San Francisco Bay area as plotted using GPS visualizer. The visualization websites after uploading file containing the coordinates also generated an optional source code which can be uploaded into Google Earth to generate an animated version of the map.

Figure 11 shows the input form used in generating the aerial map shown in Figure 12. The form allows the flexibility to change the points generated on the map and also change field parameters of the map. The data to be used can be copied and pasted in a field provided or attached as an excel file or text file. For our research work, we inputted our data as excel file attachments.



Resize/colorize data points on a map

This is a special version of the GPS Visualizer map form that's designed for plotting quantifiable data on a map. You can colorize and/or resize the points according to a generic frequency field named "N", or you can use a more typical field, such as altitude, population, or category.

If you have track data, or if you don't need to automatically colorize/resize your data points by a particular parameter, you'll probably be better served by the normal [Google Maps form](#), the [Google Earth KML form](#), or the [JPEG/PNG/SVG form](#). If you want your map to load markers dynamically (e.g., from a [Google Docs spreadsheet](#)), you definitely need to use the [standard Google form](#).

The coordinates of your data can be given as latitude/longitude, as geographic places (city-state pairs, states, or countries), U.S. ZIP codes, or Canadian postal codes. For instance, if you wanted to see how many people in each ZIP code responded to a survey you sent out, this would be a minimal valid .csv file: zip,N [new line] 97202,12 [new line] 97205,5 [new line] 97209,20. If you have a raw list of locations such as ZIP codes (i.e., some ZIPs appear more often in the list than others) and you haven't calculated the frequencies yet, you can set the "Calculate frequency" option to "auto," and GPS Visualizer will create an "N" (frequency) column for you.

Please contact Adam Schneider, using the address on the bottom of this page, if you have any questions about how this form works!

Upload your GPS data files here: ?

(Total size of all files cannot exceed 3 MB)

File #1 No file chosen

File #2 No file chosen

File #3 No file chosen

General map parameters show advanced options [+]

Output format:

Width: pixels Height:

Margin: Title:

Background map:

Or paste your data here: ?

```
name, desc, latitude, longitude, n
```

Force plain text to be this type:

Data point options show advanced options [+]

Colorize using this field: Min: Max:

Lightness: Saturation:

Spectrum direction: Hue 1: Hue 2:

Color of values beyond min. or max.:

Resize using this field:

Minimum radius: pixels Maximum radius: pixels

Unresized radius: pixels Single-point map width:

Calculate frequency: (creates a field called "N")

Show point names: Show point descriptions:

Default marker color:

Or provide the URL of data on the Web:

Google Maps options show more Google options [+]

Generate marker list: Width: px

Default icon:

Figure 11. GPS Visualizer input form



Figure 12. A visual of the landmarks in the San Francisco Bay area differentiated by colors

Figure 12 shows a Google map aerial view of a section of San Francisco Bay area. The points located on the map represent the individual landmarks and are differentiated by colors for easy identification.

4.1.4 Results of Experiment 1. Figure 13 and Table 3 below show the summation of the frequency count by all the individual cabs which visited a particular landmark within the 30 days period using experiment 1.

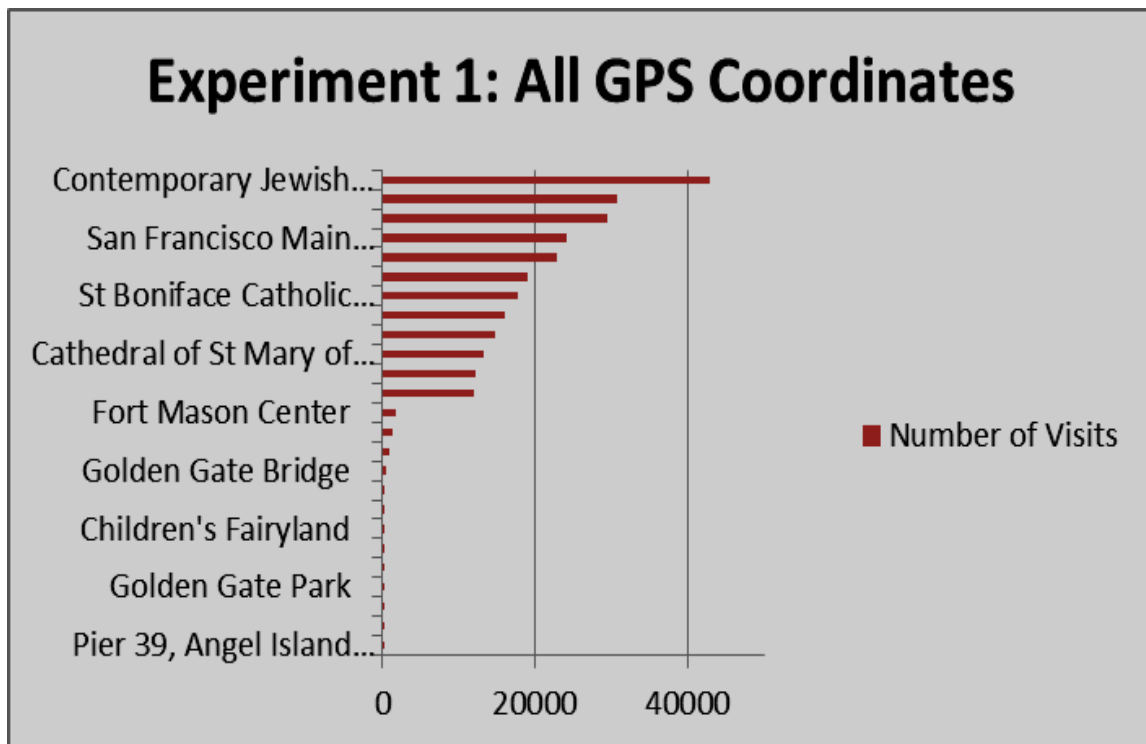


Figure 13. Horizontal bar graph representation of results from experiment 1

Figure 13 displays a horizontal bar graph of the frequency of visits versus the 25 landmark names. The graph was generated using the values provided in Table 3. These values were recorded generated from performing experiment 1. The values provided in Table 3 were arranged in descending order for easy ranking of the landmarks. With the highest table value being 42,852 and the lowest being 3, the ordered values in the table aided with arrangement of the bars in the graph.

Analyzing Figure 13 and Table 3 which are the results of experiment 1, the three most visit places were found be the Contemporary Jewish Museum, Old Saint Mary's Cathedral and Grace Cathedral respectively in descending order. The bottom three of the least visited were San Francisco Zoo, Pier 33 and Pier 39 respectively.

Table 3

Results of ranking landmarks using all GPS coordinates

Name	Number of Visits
Contemporary Jewish Museum	42851
Old St Mary's Cathedral	30571
Grace Cathedral	29467
San Francisco Main Library	24020
San Francisco Symphony at Davies Symphony Hall	22803
The Cable Car Museum	18913
St Boniface Catholic Church	17719
Lombard Street	15936
Aquarium of the Bay	14649
Cathedral of St Mary of the Assumption	13223
City Hall	12277
The San Francisco Giants at AT&T Park	12026
Fort Mason Center	1716
The de Young Museum, San Francisco Museum of Modern Art	1348
California Academy of Sciences	1004
Golden Gate Bridge	433
Beach Chalet & Park Chalet	309
Bay Area Discovery Museum	33
Children's Fairyland	30
Treasure Island	26
The Exploratorium	19
Golden Gate Park	18
San Francisco Zoo	12
Pier 33, Also pier for Alcatraz Island	9
Pier 39, Angel Island State Park	3

4.1.5 Results of Experiment 2. Figure 14 and Table 4 below show the summation of the frequency count by all the individual cabs using their pickup and drop-off coordinates within the defined proximity of the landmarks within the 30 days period using experiment 2.

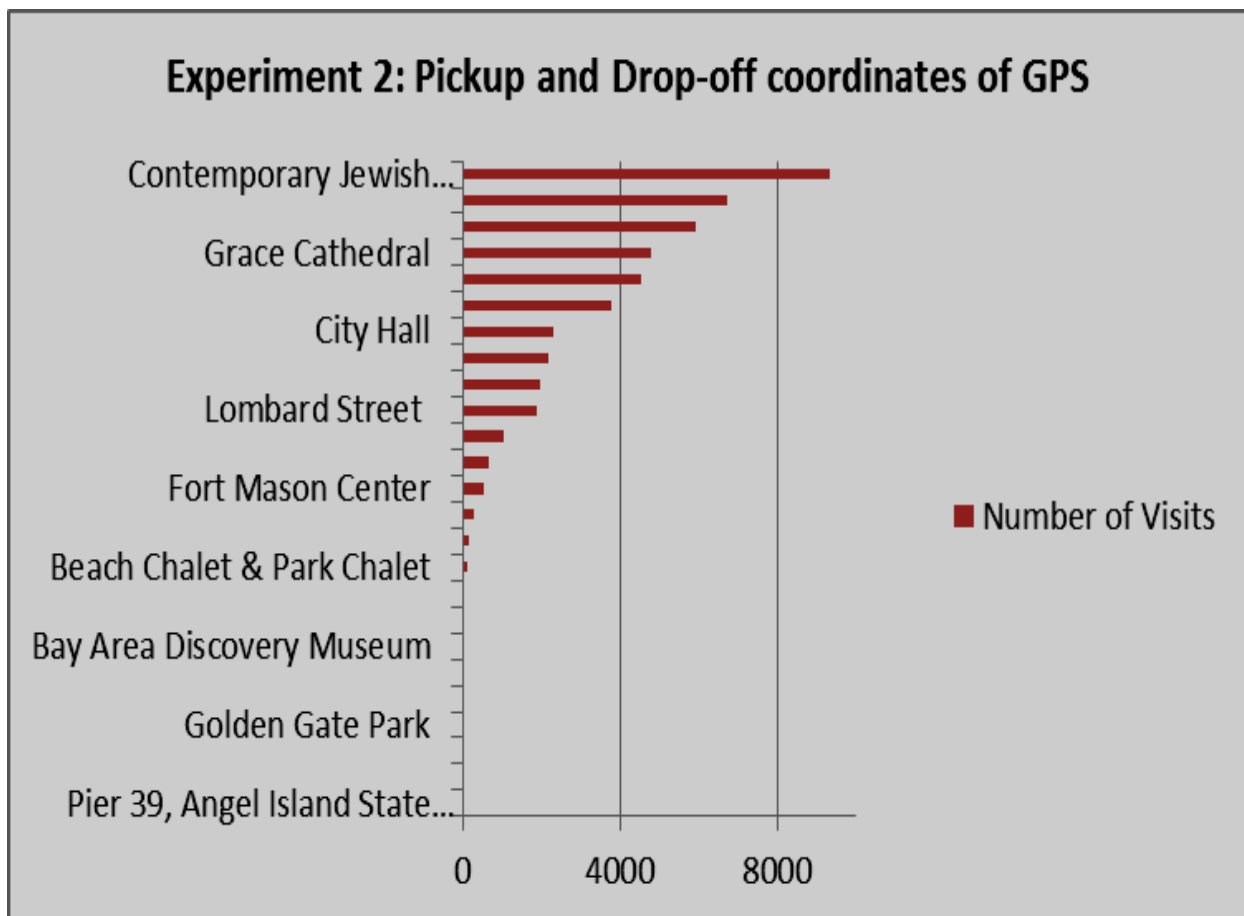


Figure 14. Horizontal bar graph representation of results from experiment 2

From experiment 2, the values of Table 4 and the graph displayed in Figure 23 were treated in the same way in terms of content format as those of the graph and table generated in experiment 1. The highest value of the number of visits recorded in Table 4 was 9,335 and its lowest value recorded was 3.

From Figure 14 and Table 4 the results of experiment 2, the three most visit places were found be the Contemporary Jewish Museum, Aquarium of the Bay and San Francisco Main Library respectively in descending order. The bottom three of the least visited were Golden Gate Bridge, Pier 33 and Pier 39 respectively.

Table 4

Results of ranking landmarks using start and end coordinates

Name	Number of Visits
Contemporary Jewish Museum	9335
Aquarium of the Bay	6719
San Francisco Main Library	5934
Grace Cathedral	4773
Old St Mary's Cathedral	4541
The San Francisco Giants at AT&T Park	3763
City Hall	2306
St Boniface Catholic Church	2160
San Francisco Symphony at Davies Symphony Hall	1982
Lombard Street	1864
The Cable Car Museum	1056
Cathedral of St Mary of the Assumption	684
Fort Mason Center	536
California Academy of Sciences	273
The de Young Museum, San Francisco Museum of Modern Art	174
Beach Chalet & Park Chalet	133
Children's Fairyland	24
Treasure Island	23
Bay Area Discovery Museum	16
San Francisco Zoo	12
The Exploratorium	9
Golden Gate Park	7
Golden Gate Bridge	4
Pier 33, Also pier for Alcatraz Island	4
Pier 39, Angel Island State Park	1

4.1.6 Summary of findings for ranking tourist attractions using time series GPS data of cabs. Using the frequency count of the number of visits to the landmarks, the output of the generated map when plotted by the GPS visualizer website is displayed in Figure 15. The size of each plotted coordinate corresponds to the number of visits by the various cabs to that particular landmark. Figure 16 shows the landmarks represent by their numerical description as listed under Table I.



Figure 15. A map showing the varying sizes of point representing the landmarks

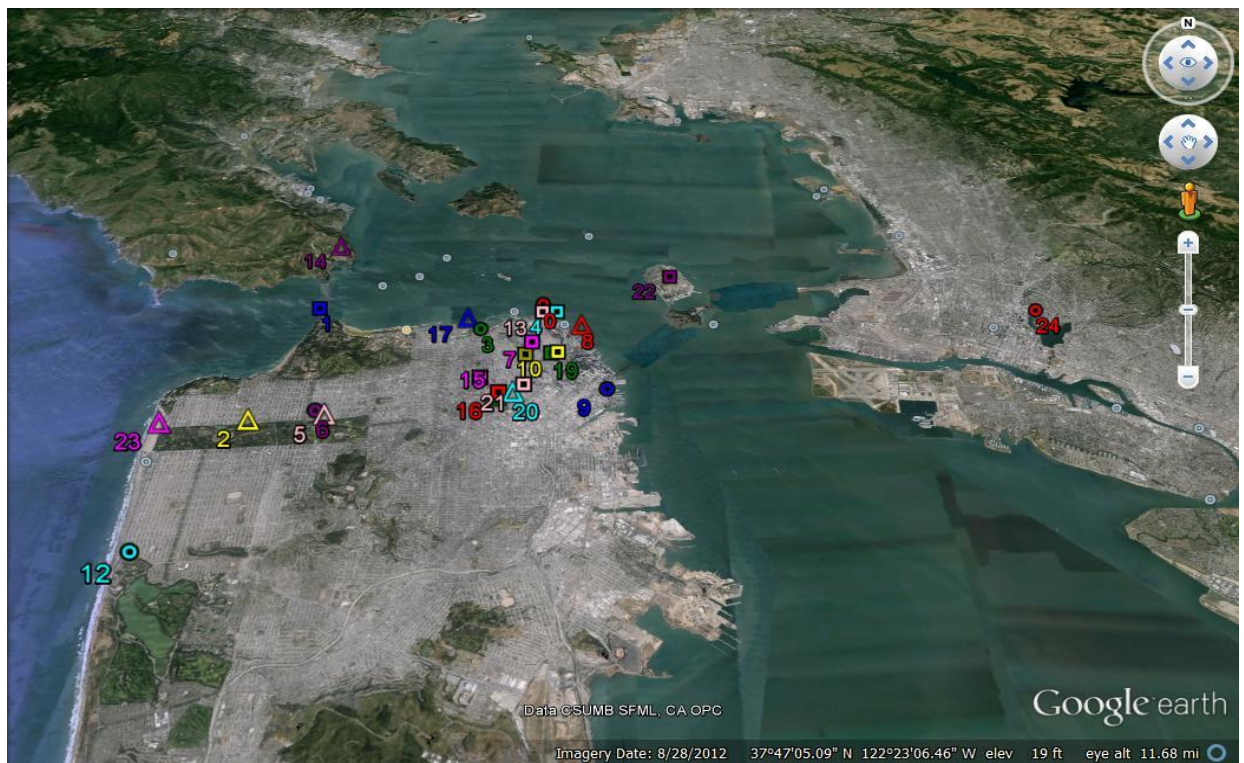


Figure 16. A map showing the numerical representation of the landmarks

From our experiments , the three most visited landmarks were all edifices of religious bodies. More children oriented landmarks listed in the area happened to be at the bottom list of our result. This can be as the result of the population in the area being Christians hence the top three attractions being churches. Half of the landmarks were located within close proximity after plotting them on the map.

4.2 Probability Model to Predict Destination

In this section of our research, we use simulated data built on the same data format as that of the San Francisco cab data to predict the possible routes to be taken by a vehicle. All the probability calculations performed in this section of the experiment were done using Bayes theorem of conditional probability formula.

4.2.1 Experiment 1 probability calculation and results. In experiment 1, a grid of 10 x 10 is created to represent a map. The sequence of cell numbers used is shown in below.

Table 5

Observation table of route number, cell sequence and frequency experiment 1

Commonly observed route number	Cell sequence	Frequency
1	54,64,74,84,85,86,87,88,89,79,69,59,49,39	10
2	27,37,47,57,67,66,65,64,63	14
3	1,2,3,4,5,6,7,8,18,28,38,48	12
4	21,22,23,24,34,44,54,55,56,57	18
5	87,86,76,66,65,64,63,53,43,33,23	8
6	82,83,73,74,64,65,66,67,68	6
7	54,64,74,84,85,86,87,88,89,79,69,59,49,48	6
8	27,37,47,57,67,66,65,75	6
9	1,2,3,4,5,6,7,8,18,28,38,39	4
10	21,22,23,24,34,44,54,55,56,46	4
11	87,86,76,66,65,64,63,62	10
12	82,83,73,74,64,65,66,56	8

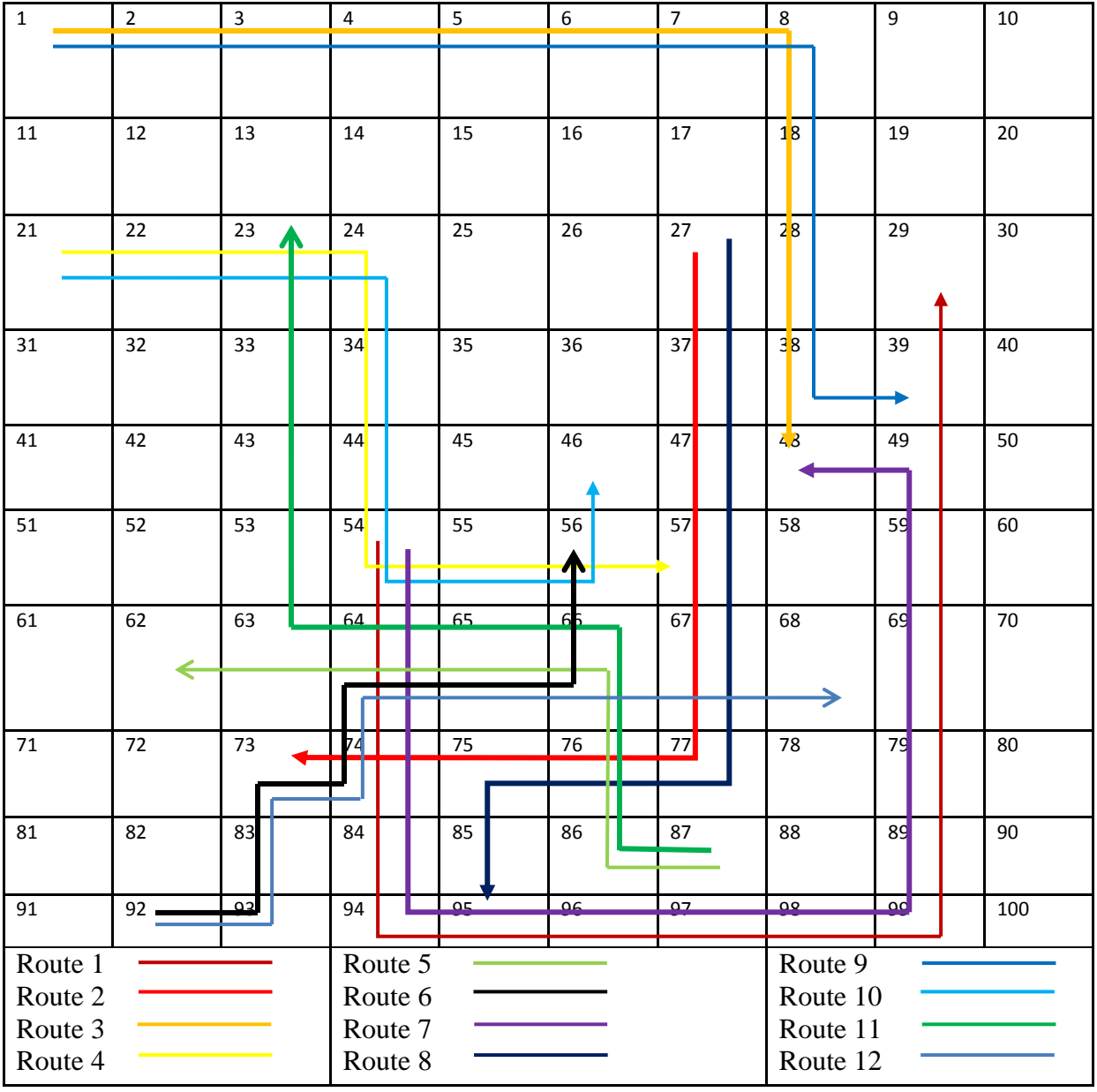


Figure 17. A representation of routes generated with colored route for experiment 1

Figure 17 shows all the various routes that were generated during experiment 1. The routes are color coded for easy tracing and tracking. An example is route ID 3 which is color coded orange. It has a starting cell number of 1 and runs through the grid following the sequence 2,3,4,5,6,7,8,18,28,38 with cell number 48 as its ending destination cell number.

Table 6 below shows the probability values recorded as a result of calculating the likelihood of a route ID being used while following a particular cell sequence. From the table a repetition of probability values can be observed. These repeated values arise as a result of the route ID's having the same initial cell sequence.

Table 6

A grid table showing the outcome of possible events in experiment 1

		Cell sequence											
		S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
Route ID	1	0.625	0	0	0	0	0	0.375	0	0	0	0	0
	2	0	0.70	0	0	0	0	0	0.30	0	0	0	0
	3	0	0	0.75	0	0	0	0	0	0.25	0	0	0
	4	0	0	0	0.447	0	0	0	0	0	0.105	0	0
	5	0	0	0	0	0.444	0	0	0	0	0	0.556	0
	6	0	0	0	0	0	0.429	0	0	0	0	0	0.571
	7	0.625	0	0	0	0	0	0.375	0	0	0	0	0
	8	0	0.70	0	0	0	0	0	0.30	0	0	0	0
	9	0	0	0.75	0	0	0	0	0	0.25	0	0	0
	10	0	0	0	0.447	0	0	0	0	0	0.105	0	0
	11	0	0	0	0	0.444	0	0	0	0	0	0.556	0
	12	0	0	0	0	0	0.429	0	0	0	0	0	0.571

Listed in table 7 are the cell sequence used in the calculation of the conditional probability for experiment 1. The twelve route IDs generated were paired up based on their cell sequence. The cell sequence S# is part of the cell numbers common to the route IDs. Using route ID 1 and 7 as an example, they have cell sequence numbers of 54,64,74,84,85,86,87,88, 89,79,69,59,49,39 and of 54,64,74,84,85,86,87,88, 89,79,69,59,49,48 but common to those cell sequence are the numbers 54,64,74,84,85,86,87 which is noted as S1.

Table 7

Probability table showing outcome of 12 events

Cell Sequence (S1 – S6)	Route ID	S#	Probability
54,64,74,84,85,86,87	1	S1	0.625
27,37,47,57,67,66,65	2	S2	0.70
1,2,3,4,5,6,7	3	S3	0.75
21,22,23,24,34,44,54	4	S4	0.447
87,86,76,66,65,64,63	5	S5	0.444
82,83,73,74,64,65,66	6	S6	0.429
54,64,74,84,85,86,87	7	S1	0.375
27,37,47,57,67,66,65	8	S2	0.30
1,2,3,4,5,6,7	9	S3	0.25
21,22,23,24,34,44,54	10	S4	0.105
87,86,76,66,65,64,63	11	S5	0.556
82,83,73,74,64,65,66	12	S6	0.571

- **Route ID 1 and 7.** For the calculation of their probability, both route ID's used the cell sequence 54, 64, 74, 84, 85, 86, 87. Route ID 1 was used a total of 10 times and route ID

7 was used 6 times. This generated a probability of 0.625 and 0.375 chances of these routes being used respectively.

- **Route ID 2 and 8.** The cell sequence used in the calculation of the probability for these two routes ID's were 27,37,47,57,67,66,65. The number of times route ID's 2 and 8 were used summed to up 14 and 6 times respectively. These generated probabilities of 0.70 and 0.30 respectively for both route ID's.
- **Route ID 3 and 9.** During the calculation of the likelihood of using route ID 3 and 9 in the future, the following cell sequences 1,2,3,4,5,6,7 were used. The chance of route ID 3 and 9 being used were calculated to be 0.75 and 0.25 respectively. These outcomes were obtained after route 3 was observed to have been used 12 times and route 9 observed to have been used 4 times.
- **Route ID 4 and 10.** For the calculation of their probability, route ID's 4 and 10 used the cell sequence 21, 22, 23, 24, 34, 44, 54. Route ID 4 was observed to have been used a total of 18 times while's route ID 10 was used 4 times. This generated a probability of 0.447 and 0.105 chances of those routes being used respectively.
- **Route ID 5 and 11.** The cell sequence used in the calculation of the chance of using routes ID's 5 and 11 were 87,86,76,66,65,64,63. The number of times route ID's 5 and 11 were used summed to up 8 and 10 times respectively. This generated a probability of 0.444 and 0.556 respectively for both route ID's.
- **Route ID 6 and 12.** Calculations for the possibility of using route ID 6 and 12 were obtained to be 0.429 and 0.571. The number of times route ID's 6 and 12 were used summed to up 6 and 8 times respectively using the cell sequence 82,83,73,74,64,65,66.

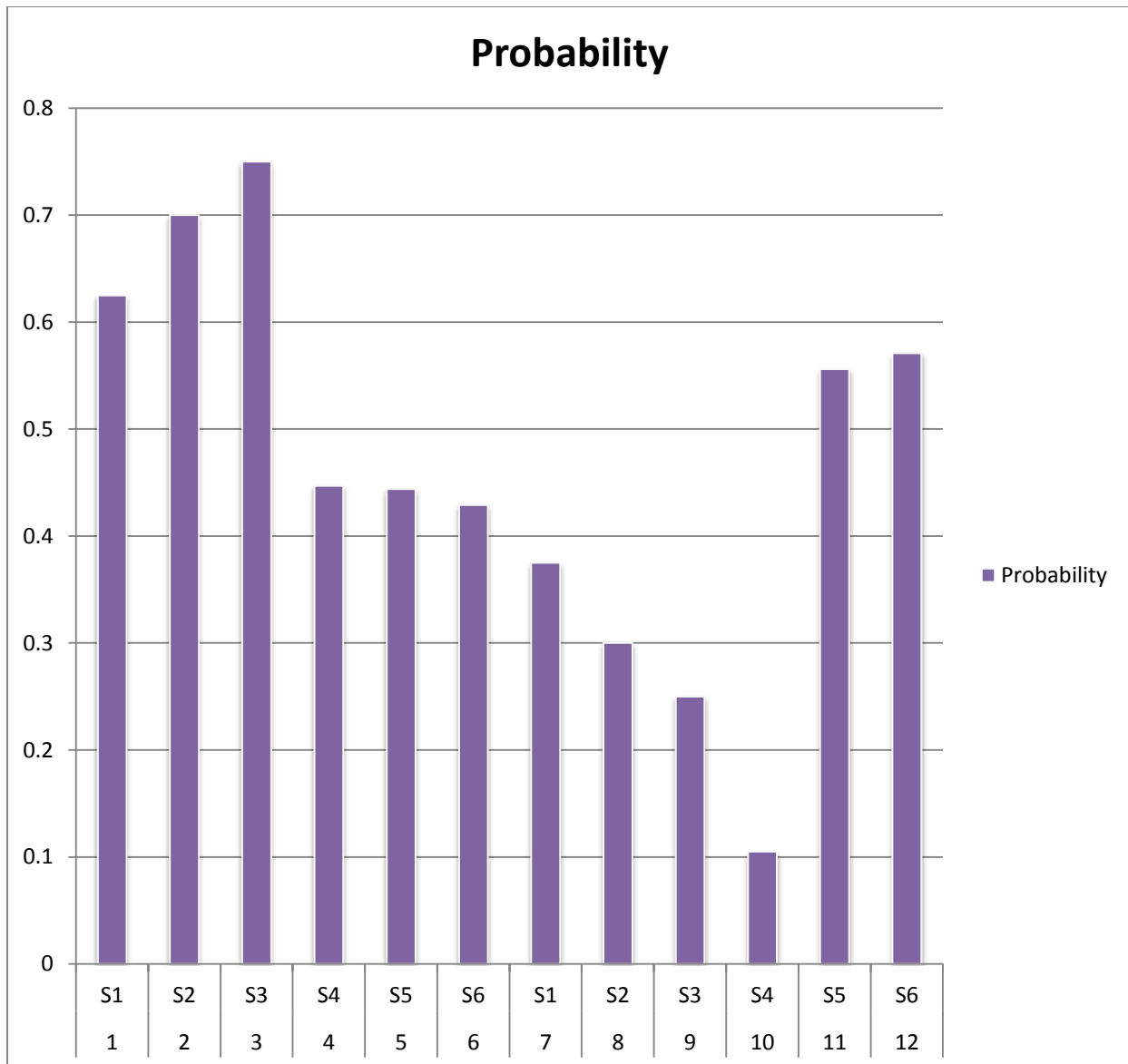


Figure 18. A graph showing probability verses route ID and cell sequence

Displayed in Figure 18 above is a graph of the route IDs versus the probability. From the calculations it was observed that, the higher the number of time a particular route was used the higher the chances of it being used in the future.

4.2.2 Experiment 2 probability calculation and results. In experiment 2 four more routes were added to those created in experiment 1. These new routes were created with a little more variance with respect to the cell sequence they followed. This was done to see how the probability outcome was affected. Experiment 2 follows the same principles as those of experiment 1 but in this situation there are three route IDs sharing a common cell sequence. As an example, a sequence of cell number 1 and 2 is common to route ID's 3, 9 and 15.

Table 8

Observation table showing route number, cell sequence and frequency experiment 2

Commonly observed route number	Cell sequence	Number of times observed
1	54,64,74,84,85,86,87,88,89,79,69,59,49,39	10
2	27,37,47,57,67,66,65,64,63	14
3	1,2,3,4,5,6,7,8,18,28,38,48	12
4	21,22,23,24,34,44,54,55,56,57	18
5	87,86,76,66,65,64,63,53,43,33,23	8
6	82,83,73,74,64,65,66,67,68	6
7	54,64,74,84,85,86,87,88,89,79,69,59,49,48	6
8	27,37,47,57,67,66,65,75	6
9	1,2,3,4,5,6,7,8,18,28,38,39	4
10	21,22,23,24,34,44,54,55,56,46	4
11	87,86,76,66,65,64,63,62	10
12	82,83,73,74,64,65,66,56	8
13	54,64,74,84,94,95,96,97,98,99,89,79,69,59,60	4
14	27,37,47,57,67,77,87,97,,96,95,94,93,83,73,63,62	3
15	1,2,12,13,3,4,5,6,7,17,18,19	6
16	21,22,23,24,14,15,16	7
17	82,83,93,94,84,74	7

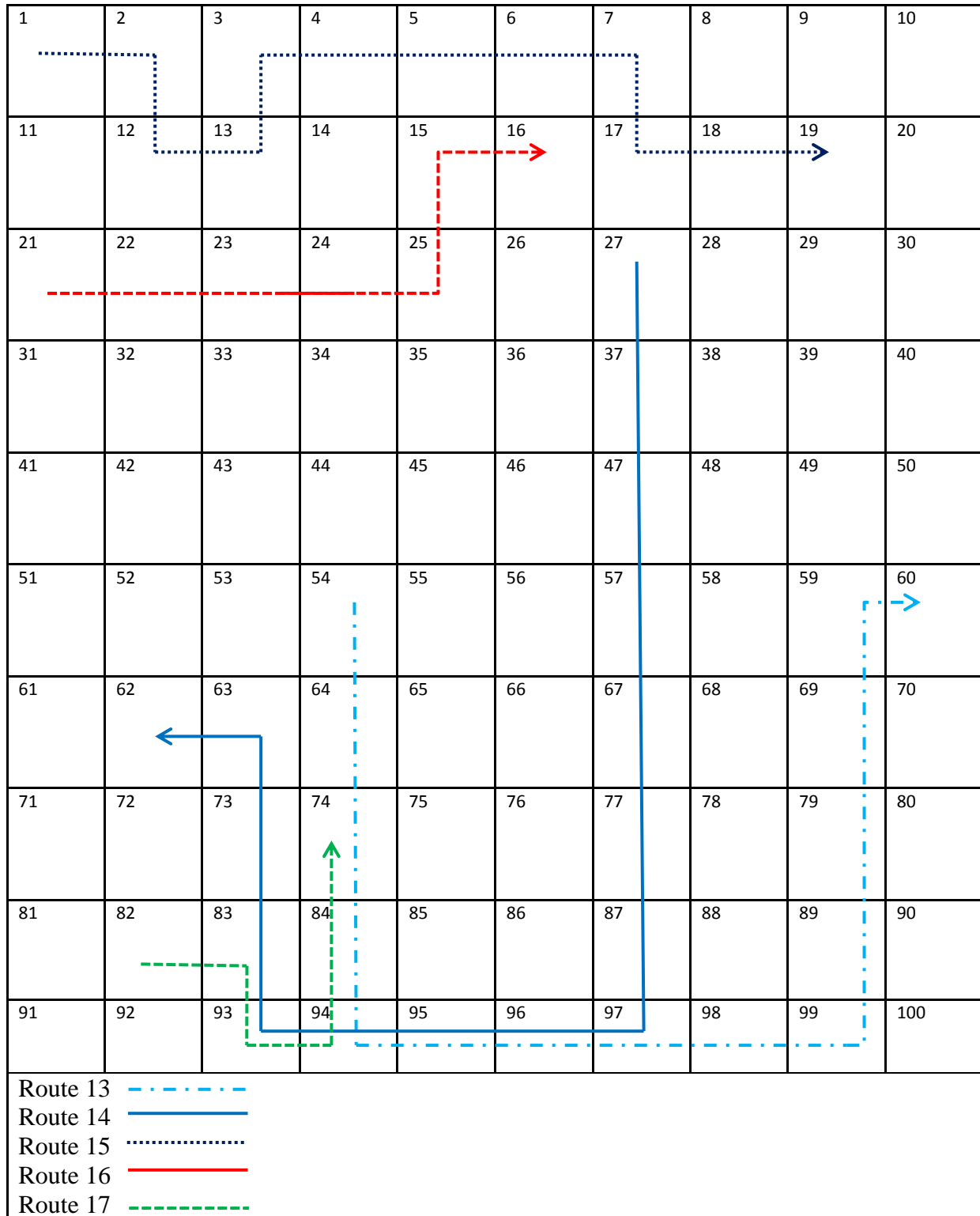


Figure 19. Representation of routes generated with colored route for experiment 2

Showing below in Table 9 are the probability values recorded from performing experiment 2. From the table a repetition of probability values can be observed times. This is as a result of part of the same cell sequences being used by three different route IDs. An example will route ID 1, 7, and 13 using part of the same cell sequences.

Table 9

A grid table showing the outcome of possible events in experiment 2

		Cell sequence																
		S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S6
Route ID	1	0.454	0	0	0	0	0	0.272	0	0	0	0	0	0.136	0	0	0	0
	2	0	0.609	0	0	0	0	0	0.261	0	0	0	0	0	0.13	0	0	0
	3	0	0	0.545	0	0	0	0	0	0.182	0	0	0	0	0	0.273	0	0
	4	0	0	0	0.38	0	0	0	0	0	0.089	0	0	0	0	0	0.156	0
	5	0	0	0	0	0.44	0	0	0	0	0	0.56	0	0	0	0	0	0
	6	0	0	0	0	0	0.286	0	0	0	0	0	0.381	0	0	0	0	0.333
	7	0.454	0	0	0	0	0	0.272	0	0	0	0	0	0.136	0	0	0	0
	8	0	0.609	0	0	0	0	0	0.261	0	0	0	0	0	0.13	0	0	0
	9	0	0	0.545	0	0	0	0	0	0.182	0	0	0	0	0	0.273	0	0
	10	0	0	0	0.38	0	0	0	0	0	0.089	0	0	0	0	0	0.156	0
	11	0	0	0	0	0.44	0	0	0	0	0	0.56	0	0	0	0	0	0
	12	0	0	0	0	0	0.286	0	0	0	0	0	0.381	0	0	0	0	0.333
	13	0.454	0	0	0	0	0	0.272	0	0	0	0	0	0.136	0	0	0	0
	14	0	0.609	0	0	0	0	0	0.261	0	0	0	0	0	0.13	0	0	0
	15	0	0	0.545	0	0	0	0	0	0.182	0	0	0	0	0	0.273	0	0
	16	0	0	0	0.38	0	0	0	0	0	0.089	0	0	0	0	0	0.156	0
	17	0	0	0	0	0	0.286	0	0	0	0	0	0.381	0	0	0	0	0.333

Listed in table 10 are the cell sequence used in the calculation of the conditional probability for experiment 2. The seventeen route IDs generated were paired up based on their cell sequences.

Table 10

Probability table showing outcome of 17 events

Cell Sequence (S1 – S12)	Route ID	S#	Probability
54,64,74,84	1	S1	0.454
27,37,47,57	2	S2	0.609
1,2	3	S3	0.545
21,22,23,24	4	S4	0.38
87,86,76,66	5	S5	0.44
82,83,73,74	6	S6	0.286
54,64,74,84	7	S1	0.272
27,37,47,57	8	S2	0.261
1,2,3,4	9	S3	0.182
21,22,23,24	10	S4	0.089
87,86,76,66	11	S5	0.56
82,83,73,74	12	S6	0.381
54,64,74,84	13	S1	0.136
27,37,47,57	14	S2	0.13
1,2,3,4	15	S3	0.273
21,22,23,24	16	S4	0.156
82,83,73,74	17	S6	0.333

- **Route ID 1, 7 and 13.** These three route IDs share a common cell sequence of 54, 64, 74, 84. Route ID 1 was observed to have been used a total of 10 times, route ID 7 was used 6 times and route ID 13 was used 4 times. This generated a probability of 0.454, 0.272 and 0.136 chances of these routes being used respectively in the future.
- **Route ID 2, 8 and 14.** The cell sequence used in the calculation of the probability for these routes ID's were 27,37,47,57. The number of times route ID's 2, 8 and 14 were used summed to up 14, 6 and 3 times respectively. This generated a probability of 0.609, 0.261 and 0.130 respectively for both route ID's.
- **Route ID 3, 9 and 15.** During the calculation of the probability of route ID 3, 9 and 15 being used in the future, the following cell sequence 1, 2 was used. The chance of route ID 3, 9 and 15 being used was calculated to be 0.545, 0.182 and 0.273 respectively. The outcomes were obtained after route ID 3 was observed to have been used 12 times, route ID 9 observed to be used 4 times and route ID 15 used 6 times.
- **Route ID 4, 10 and 16.** For the calculation of their probability, route ID's 4, 10 and 16 used the cell sequence '21, 22, 23, 24'. Route ID 4 was observed to have been used a total of 18 times, route ID 10 was used 4 times and that of route ID 16 was 7 times. This generated a probability of 0.38, 0.089 and 0.156 chances of these routes being used respectively in the future.
- **Route ID 5 and 11.** The cell sequence used in the calculation of the probability of routes ID's 5 and 11 were 87,86,76,66,65,64,63. The number of times route ID's 5 and 11 were used summed to up 8 and 10 times respectively. This generated a probability of 0.444 and 0.556 respectively for both route ID's.

- Route ID 6, 12 and 17.** Calculations for the possibility of using route ID 6, 12 and 17 were obtained to be 0.286, 0.381 and 0.333. The number of times route ID's 6, 12 and 17 were used summed to up 6, 8 and 7 times respectively using the cell sequence 82,83,73,74.

From the probability calculations and graph displayed in Figure 20, it was observed that the more route IDs which used part of the same cell sequence, the less chance are created for each route ID to be used in the future.

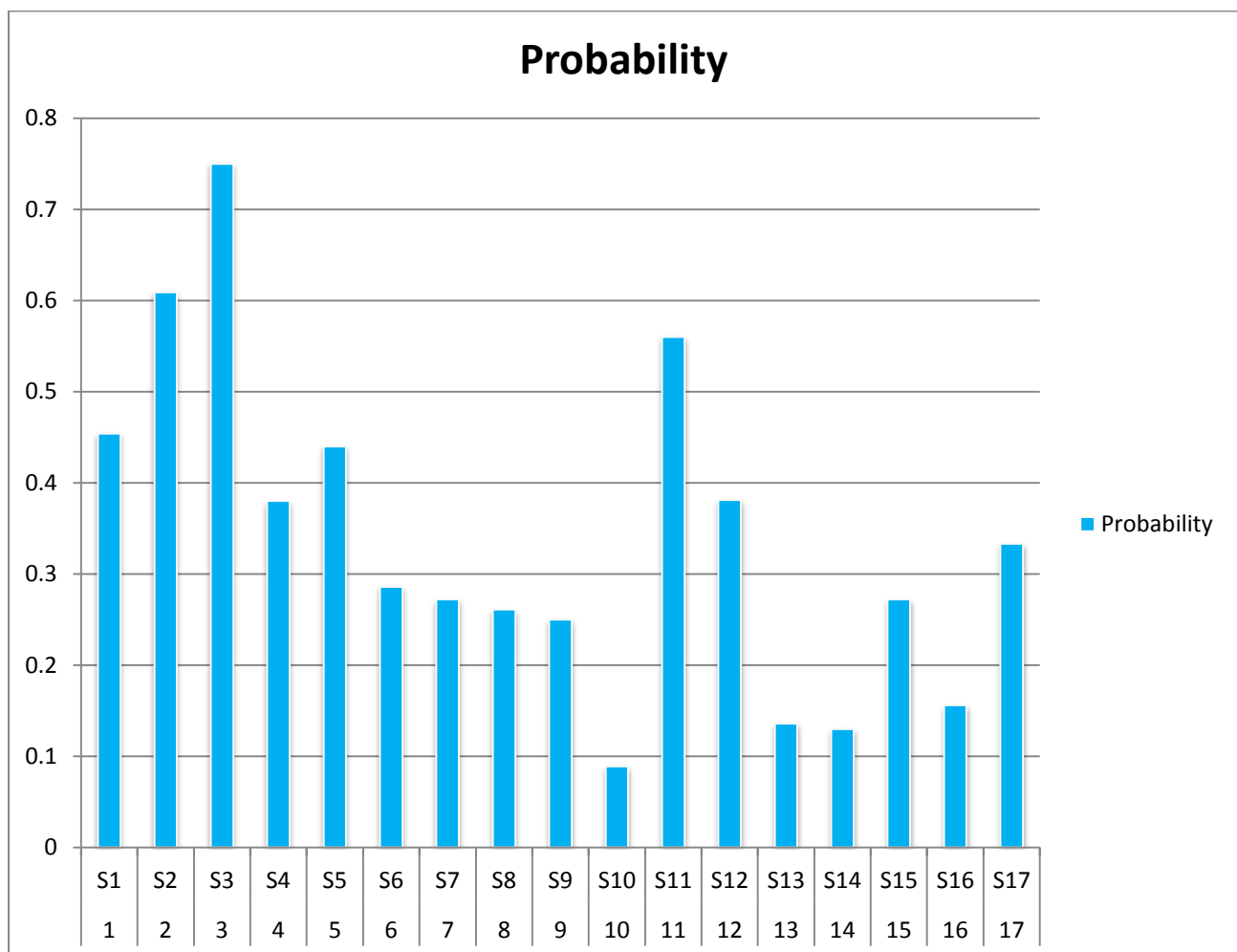


Figure 20. A graph showing probability verses route ID and cell sequence in experiment 2

4.2.3 Experiment 3 probability calculation and results. Under this section of our research we use a real map of a section of the city of Greensboro North Carolina. We took this decision to help us achieve prediction results from data generated by using real routes and destinations as shown in Figure 21.

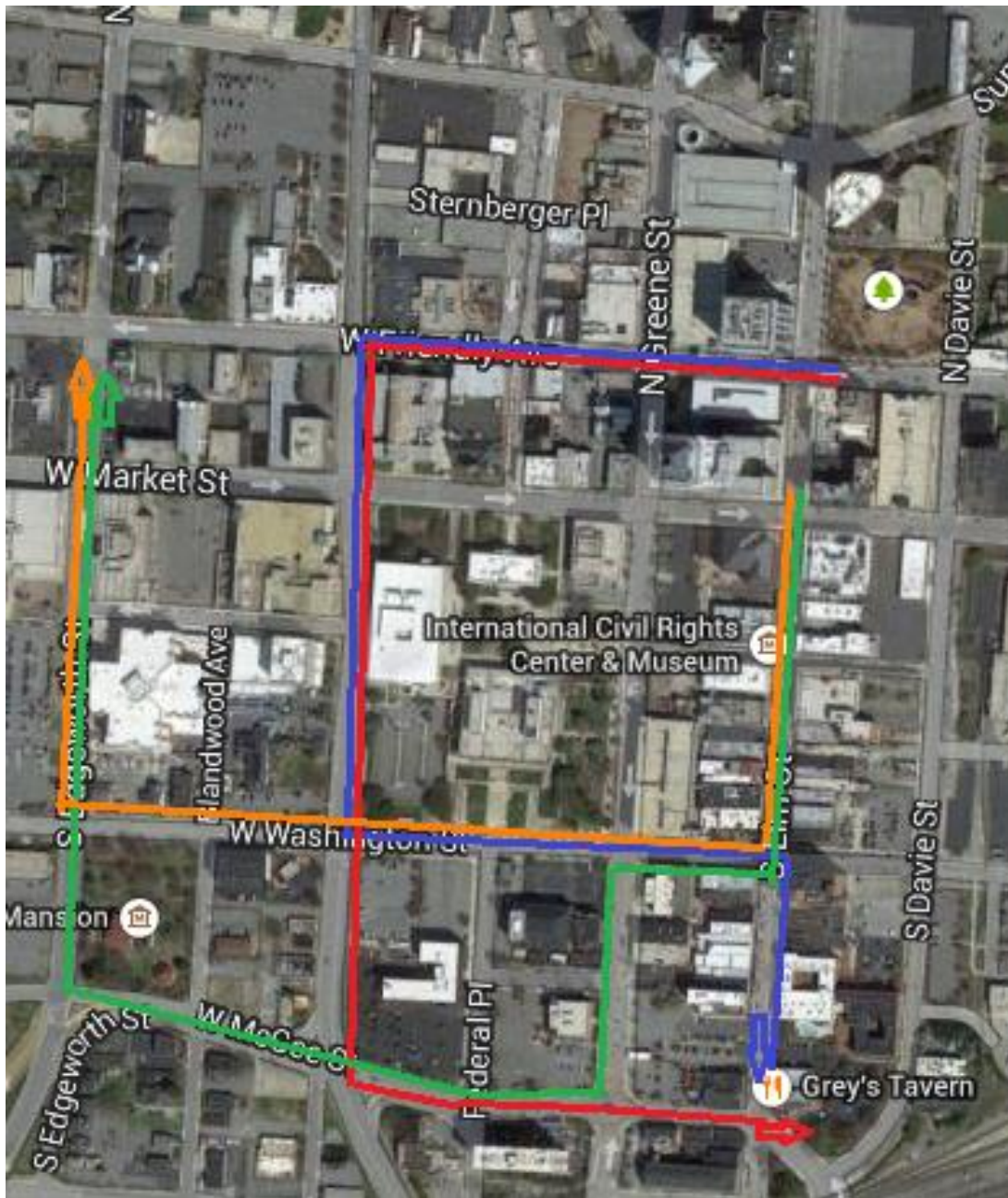


Figure 21. Satellite image showing a section of downtown Greensboro with four routes

Table 11 shows a record of the number of times a particular route was used and the sequence followed. The cell sequences which represent actual street names are described as follows:

- Cell numbers 4,3,2,1 – West Friendly Avenue. This route is a one way direction running from east to west. Cell numbers 5,6,7,8 – West Market Street. This is also a one way direction running west to east. Cell number 9,10,11,12 – West Washington Street. This is bi-directional Street. Cell numbers 13,14,15,16 – West McGee Street. This is also a bi-directional street.
- Running vertically is South Edgeworth Street with cell numbers 13,9,5,1 south to north geographically. Cell numbers 2, 6, 10, 14 represent North Eugene Street and is a bi-directional street. Cell number 3, 7, 11, 15 represents North Greene Street and it is a one way running from north to south. Cell number 4, 8, 12, 16 represents North Elm Street and it is a bi-directional Street.

Table 11

Observation table of route numbers, cell sequence and frequency experiment 3

Commonly observed route number	Cell sequence	Number of times observed
1	4,3,2,6,10,14,15,16	10
2	4,3,2,6,10,11,12,16	14
3	8,12,11,10,9,5	12
4	8,12,11,15,14,13,9,5	18

Table 12 and table 13 shows the probability calculated for the chances that a particular route ID would be used in the future. To determine the probability, we used part of the cell sequence assigned to a route ID versus the number of times that cell was. Using one of the cell

sequences as an example, S1 which has a cell sequence of 4, 3, and 2 is common to route ID 1 and route ID 2 which is made of up cell sequence 4, 3, 2, 6, 10, 14, 15, 16 and 4, 3, 2, 6, 10, 11, 12, 16.

Table 12

A grid table showing the outcome of possible events in experiment 3

		Cell Sequence			
		S1	S2	S1	S2
Route ID	1	0.409	0.591	0	0
	2	0.409	0.591	0	0
	3	0	0	0.593	0.407
	4	0	0	0.593	0.407

Table 13

Probability table showing outcome of 4 events

Cell Sequence (S1 – S12)	Route ID	S#	Probability
4,3,2	1	S1	0.409
4,3,2	2	S2	0.591
8,12,11	3	S1	0.593
8,12,11	4	S2	0.407

Displayed in Figure 23 is a graphical representation of results of experiment 3. The route IDs and the probability corresponding to those route IDs are displayed in the graph below.

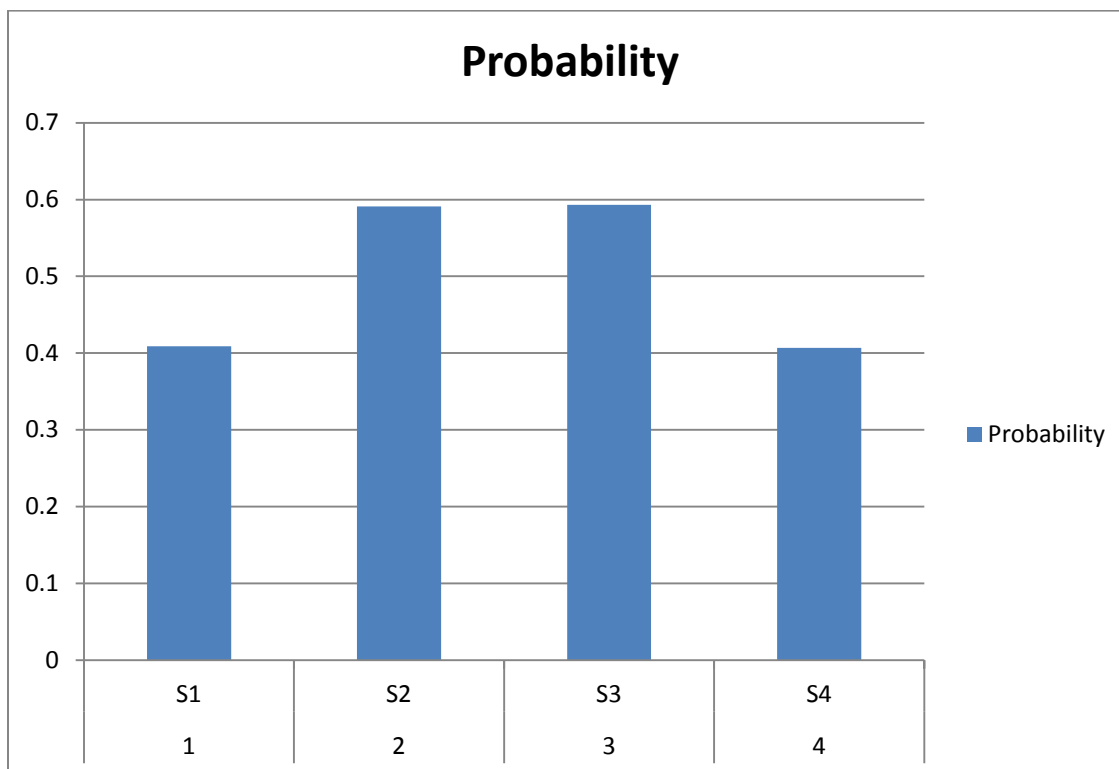


Figure 23. A graph showing probability verses route ID and cell sequence in experiment 3

The probability values generated from the simulated and real map shows that the frequency at which a route directly influences the chance of predicting the likelihood of it usage in the future.

CHAPTER 5

Discussion and Future Research

For the experiments performed in this thesis work, we used real GPS datasets and simulated. Using programs and algorithms, we conducted two experiments to help determine the ranking of tourist attractions. We were also able to predict, using conditional probability calculations, the route that a vehicle might use in the future based past routes used. This part of our research was possible after carefully generating gridded subdivisions of geographical locations and identifying the starting cell a vehicle follows until it reaches its destination cell.

For our proposed future work, we would like to create a predictive model for physical security by analyzing data from the tracking of GPS coordinates of suspects by taking a closer look at pickup points and the number of times they visited particular drop-off points. We are also interested in using it in the retail domain by capturing the movement/buying pattern of a customer to a shopping mall stall. This can be done using a mobile application with the customer's permission to allow access by the mall stalls. Offers can be sent right at the time of shopping based on the customer's previous movement. From examining the results of the experiments, we were able to identify the most visited places within San Francisco. Prior to performing this research, we initially thought the most visited places would be centered on family fun or entertainment grounds; the results proved otherwise.

References

- Andersson, M., Gudmundsson, J., Laube, P., & Wolle, T. (2007). *Reporting leadership patterns among trajectories*. Paper presented at the Proceedings of the 2007 ACM symposium on Applied computing.
- Benkert, M., Gudmundsson, J., Hübner, F., & Wolle, T. (2008). Reporting flock patterns. *Computational Geometry*, 41(3), 111-125.
- Birds, A. a. t. (2007). All about the birds. from <http://www.birds.cornell.edu/AllAboutBirds/studying/migration/>
- Birmingham, B. P. a. W. (2005). Modeling Form for On-line Following of Musical Performances. *Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania,*.
- Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*: Springer.
- Brandtzæg, P. B. (2013). Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years. *Science Daily*.
- Dodge, S., Weibel, R., & Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. *Information Visualization*, 7(3-4), 240-252.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.
- Erwig, M. (2004). Toward Spatio-Temporal Patterns *Spatio-Temporal Databases* (pp. 29-53): Springer.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.

- Gray, L. (2013). *How Does GPS Work?* : The Rosen Publishing Group.
- Laube, P., & Purves, R. S. (2006). An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, 30(3), 347-374. doi: 10.1016/j.compenvurbsys.2005.09.001
- Lee, O. H. P., Keun Ho Ryu. (2004). Temporal moving pattern mining for location-based service. *Journal of Systems and Software*, 73(3), 481-490. doi: 10.1016/j.jss.2003.09.021
- Mazzoni, D. (2005). *LibFeature: A software library for quickly generating feature vectors on the fly from structured data*. Paper presented at the Eighth Workshop on Mining Scientific.
- McDonald, D. B. (2013). Population Ecology. from <http://www.uwyo.edu/dbmcd/pepecol/feblects/lect06.html>
- Noyon, V., Devogele, T., & Claramunt, C. (2005). A formal model for representing point trajectories in two-dimensional spaces *Perspectives in Conceptual Modeling* (pp. 208-217): Springer.
- Piorowski, M., Sarafijanovic-Djukic, N., & Grossglauser, M. (2009). CRAWDAD data set epfl/mobility (v. 2009-02-24).
- Roger. (2008). Uses associated with HMMs. *Hidden Markov Models*. from http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/hmms/s2_pg1.html
- Rosenbaum, D. (2008). Top Attractions in San Francisco. *San Francisco Travel*. from <http://www.sanfrancisco.travel/todo/Top-Attractions-in-San-Francisco.html>
- Satish, L., & Gururaj, B. (1993). Use of hidden Markov models for partial discharge pattern classification. *Electrical Insulation, IEEE Transactions on*, 28(2), 172-182.
- Zheng, Y., Xie, X., & Ma, W.-Y. (2010). GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.*, 33(2), 32-39.

Zheni, D., Frihida, A., Ghezala, H. B., & Claramunt, C. (2009). A semantic approach for the modeling of trajectories in space and time *Advances in Conceptual Modeling-Challenging Perspectives* (pp. 347-356): Springer.

Zweig, G., & Burges, C. J. (2011). The Microsoft Research sentence completion challenge: Technical Report MSR-TR-2011-129, Microsoft.