ANALYZING DESTINATION CHOICES OF TOURISTS AND RESIDENTS FROM LOCATION BASED SOCIAL MEDIA DATA

by

MD MEHEDI HASNAT
B.Sc. Bangladesh University of Engineering and Technology, 2014

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2018

Major Professor: Samiul Hasan

ABSTRACT

Ubiquitous uses of social media platforms in smartphones have created an opportunity to gather digital traces of individual activities at a large scale. Traditional travel surveys fall short in collecting longitudinal travel behavior data for a large number of people in a cost effective way, especially for the transient population such as tourists. This study presents an innovating methodological framework, using machine learning and econometric approaches, to gather and analyze location-based social media (LBSM) data to understand individual destination choices. *First*, using Twitter's search interface, we have collected Twitter posts of nearly 156,000 users for the state of Florida. We have adopted several filtering techniques to create a reliable sample from noisy Twitter data. An ensemble classification technique is proposed to classify tourists and residents from user coordinates. The performance of the proposed classifier has been validated using manually labeled data and compared against the state-of-the-art classification methods. *Second*, using different clustering methods, we have analyzed the spatial distributions of destination choices of tourists and residents. The clusters from tourist destinations revealed most popular tourist spots including emerging tourist attractions in Florida. *Third*, to predict a tourist's next destination type, we have estimated a Conditional Random Field (CRF) model with reasonable accuracy. *Fourth*, to analyze resident destination choice behavior, this study proposes an extensive data merging operation among the collected Twitter data and different geographic database from state level data libraries. We have estimated a Panel Latent Segmentation Multinomial Logit (PLSMNL) model to find the characteristics affecting individual destination choices. The proposed PLSMNL model is found to better explain the effects of variables on destination choices compared to trip-specific Multinomial Logit Models. The findings of this study show the potential of LBSM data in future transportation and planning studies where collecting individual activity data is expensive.

ACKNOWLEDGMENTS

I would like to convey my heartiest gratitude to my honourable supervisor Dr. Samiul Hasan for his excellent supervision and constant support in this thesis.

I would also like to acknowledge the support from my research group mates, the expedient input from Dr. Naveen Eluru, and great encouragement from my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE 3. 1: Heuristic Algorithm for Tourist Identification 16

FIGURE 3. 2 Comparison of Performances of the Classifiers with proposed heuristic method. 19

FIGURE 3. 3: Daily Activity Plots. (a) Total Activity Plots, (b) Geo-tagged Activity Plots (inside Florida) and (c) Horizontal Shift in Hourly Activity. 20

FIGURE 3.4: Hourly Activity Plots. (a) Tourists, (b) Residents. 21

FIGURE 3. 5: Weekday and Weekend Geo-located Activities. (a) Tourists' Weekly Activity Plot and (b) Residents' Weekly Activity Plot. 22

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1 Introduction

Travel demand models are crucial to transport planners and policy makers to develop, assess, and select suitable long term plans (Rashidi et al., 2017). Surveys complemented by additional sources of information such as travelers' feedbacks (by phone, mail or online) have been used as established sources of information for inputs to such models. However, implementing these surveys are costly and time consuming (Flyvbjerg et al., 2005). Moreover, tour-based schemes such as activity-based modeling approaches need individual level travel information (Abbasi et al., 2015). The shift towards activity based modeling has made individuals and households more significant contributors as decision making units (Rasouli and Timmermans, 2014). The evolution of travel demand modeling techniques brought about the need for high resolution databases in which individual socio-economic attributes are used to model their daily travel behavior. A complete household survey with all the required travel information costs about $200 per household (Zhang and Mohammadian, 2008). Therefore, although access to such individual level travel information is crucial for developing advanced travel behavior models, it is infeasible in terms of cost and time. Nowadays, technologies are being used to collect this information in a cost effective way. For example, web-based surveys (trip planning apps), social networking sites or applications, smart phones (accelerometers), and personal health sensors have been explored to collect individual travel information. However, researchers are yet to explore the full potential as well as limitations of these emerging technology-based methods (Abbasi et al., 2015).

Collecting individual travel behavior data becomes more difficult for cities with a large number of tourists who are the most dynamic population group whose size and travel choices

change rapidly compared to residents. Tourism activities in a city can be unevenly distributed as they are superimpose on a spatial system and infrastructure network that may not have been designed specifically to cater for it (Gladstone and Fainstein, 2001). Locating tourists points of interests within a city and how they travel from one point of interest to the next is not something discovered through subjective observation (Edwards et al., 2008). For major tourism dependent cities, it is essential to understand tourist travel behavior since tourism related traffic cause huge pressure on their transportation systems (Cho et al., 2011; Gursoy et al., 2002). Although census statistics reveal total inflow and outflow of tourists, it only presents as a macro level data considering over large regions. However, it is difficult to collect individual level travel information which includes trip purpose, activity type, activity location, departure time, traffic condition, mode of transport etc. from tourists.

To collect travel data, researchers are looking for complementary data sources. With the transformation of Web into a true collaborative and social platform (Chi, 2008), we can access a large volume of user generated contents shared in various social media platforms (Kuflik et al., 2017). Social media can be defined as a collection of internet-based applications which allows users to generate and exchange their contents (Kaplan and Haenlein, 2010). Social media platforms such as Twitter is now considered as a useful source of travel behavior information in various studies (Cao et al., 2014; Chang et al., 2012; Gal-Tzur et al., 2014; Maghrebi et al., 2015). Cost of acquiring such data is minimal compared to other traditional travel survey methods. The easy availability and wide range of applications have made the data valuable for researchers in multiple fields including social science, marketing, public health, computer science, and transportation science. Social media data have been used in activity recognition (Lian and Xie, 2011), finding mobility and activity choices (Chen et al., 2017; Hasan and Ukkusuri, 2014),

classification of activity choice patterns (Cheng et al., 2011), role of friendship on mobility (Hasan et al., 2016; Sadri et al., 2017), and modeling activity sequence (Hasan and Ukkusuri, 2017). In transportation planning, researchers have used this data to estimate urban travel demand (Lee et al., 2017; Liu et al., 2014) and traffic flow (Liu et al., 2014; Wu et al., 2014). Thus, social media data has a significant potential for travel demand models, traffic operations and management and long term transportation planning purposes (Rashidi et al., 2017).

The main challenge in using such data sources is the significant noises that have to be filtered before any meaningful information can be accessed. To extract information such as trip purpose, travel mode etc. advanced text mining, linguistic techniques and data mining techniques are required (Cramer et al., 2011; Maghrebi et al., 2015). In this regard, it is relatively easier to work with check-in and geo-tagged data as they are already associated with a location. This study presents a data mining framework for understanding tourist and resident travel behavior of Florida from geo-tagged posts of a popular social media platform, Twitter.

1.2 Background

Florida has a number of famous tourist spots attracting millions of tourists from home and abroad every year. In 2016, Florida hosted more than 113 million visitors from outside of USA, which supported 1.4 million jobs and making a spending of 109 billion USD (Eye et al., 2017). Central Florida region had 68 million visitors in 2016 with Orlando being one of the top destinations among the global tourist cities and second in annual tourist spending among the domestic cities (Eye et al., 2017). A study conducted by Florida Department of Transportation (FDOT) also found that in 2010, nearly 8% of Florida's vehicle miles travelled were comprised of tourism related travel (*Florida Transportation Trends and Condition 2012*, 2012). Individual movement, route

choice, origin and destination of this large number of seasonal population have a significant impact on transportation infrastructure. Such information can provide vital insights for transportation and city planners.

With millions of active users, social media platforms such as Facebook, Twitter, Instagram, Flickr etc. have become potential big data sources of individual behavior. Nearly 80 percent of Americans use social media while two third of the global internet population visits social networks (Perrin, 2005). Thus, ubiquitous uses of social media platforms have created a tremendous opportunity to gather digital traces. Analyzing millions of user footprints, it is possible to extract travel behavior at a scale unimaginable before (Hendrik and Perdana, 2014). However, not all social network data are available and have rich information. Twitter is a potential data source as it is available through simple web scraping and has a wide range of information within each post (tweet). Twitter has become a popular communication platform with 317 million monthly active users (67 million users from the USA) sending 500 million tweets per day ("Twitter by the Numbers: Stats, Demographics & Fun Facts," 2017). Despite being unstructured, tweets provide important clues about latent user attributes and activities- absent in GPS logs and mobile phone records (Cao et al., 2014). From Twitter we can extract spatial (geo-tagged) and temporal (time-stamped) information for a longer period and large number of users without invading user privacy (Frias-Martinez et al., 2012; Hasan and Ukkusuri, 2015).

Traditional travel surveys are limited in terms of sample size, area of coverage and updating frequencies. For instance after the National Household Travel Survey (NHTS) household survey of 2009, the database is recently being updated based on the most recent data collected in April 2017. The data set contains travel information of slightly over 129,000 households. Few organizations are trying to collect updated travel information through some innovative ways.

North Florida Transportation Planning Organization has initiated an online travel survey through a third party named Resource Systems Group, Inc. 2017 ("North Florida Travel Survey," 2017) for six-county of North Florida region (Baker, Clay, Duval, Nassau, Putnam, and St. Johns counties). The surveys were open from July 2017 to January 2018 and the responses are yet to be explored. On the other hand, with big data sources it is possible to record the movements of millions of individuals at unprecedented spatial and temporal accuracy (Beyer and Laney, 2012). However, it is vital to note that, this types of high resolution spatial data comes with its own trade-offs as often the social demographic attributes are not available, making it extremely difficult to correctly weigh the sample (Beyer and Laney, 2012) and use this in contrast of transportation planning purposes. This publicly available data are limited or highly aggregated and the collection and sampling methodologies are normally not available for validation (Morstatter et al., 2013). With the advantages and limitations of traditional survey and location based social media data, this study focuses on harnessing the goods from both the sources by developing a framework to combine their attributes.

1.3 Objectives and Scopes of the Study

In this study, we develop a framework to collect most recent travel information in a cost effective way to be used in various transportation and planning studies. We present a data mining framework for understanding tourists' and residents' travel behavior using social media data. We have gathered data using Twitter's search interface and followed several filtering steps to create a reliable sample. With the sampled database we propose stepwise procedures to achieve some specific objectives.

- **Data gathering**

We have utilized different streaming and search interface to gather real time and historical Twitter data. However, the collected data cannot be readily used for transportation related studies. Therefore, we present several filtering steps to create a reliable sample from noisy data.

- **User classification**

We propose a classification method to identify the users who are non-native to a particular area. The proposed method is validated with manually labeled data and compared with state of the art classification techniques. With reliable features extracted from the data set, we further propose an advanced ensemble classifier to improve prediction results.

- **Location clustering**

After identification and validation of the tourist accounts, we find the spatial patterns of tourists and residents visited destinations. With application of state of art clustering techniques, we find the most visited locations of tourists and compare them with the most recent tourist database.

- **Tourist travel sequence and next destination type prediction**

With a larger volume of sample dataset, we analyze the tourist's destination patterns using Markov chain and Conditional Random Field (CRF) approaches. We have analyzed travel sequence of tourists and find out their most probable destination using their transition probabilities between different types of destinations. We explored effects of different features of particular visits to predict the next destination types through the application of CRF models.

- **Resident destination choice modeling**

We propose a framework to develop resident destination choice models using residents' geo-tagged Twitter posts. This step includes extensive data merging techniques among social media

data and different geographic database preserved in state level data libraries. The data preparation sub-section in chapter six describes the challenges faced and overcame in identifying resident profiles and extracting their home locations and destination locations. We frame the problem into a Panel Latent Segmentation Multinomial Logit (PLSMNL) model and explain the outcome qualitatively as well as quantitatively.

1.4 Thesis Contribution

This study has several contributions in the field of data analytics in transportation. It shows the potential of social media data for understanding travel behavior of different groups of users. It presents several filtering steps to create a reliable sample from noisy social media data. Using a classification method, this study separates residents and visitors within a study area. Using available spatial clustering methods, this study determines the most common attraction tourism spots in the study area. To understand tourists travel patterns this study utilizes undirected graphical models which predicts the next possible location to be visited by a tourist. We have developed a destination choice model by integrating the census tract database with the extracted location information which incorporates individual level characteristics of the resident users. From the outcome of this study, we will have a better understanding about the tourist's as well as the residents' choice of destinations inside the study area. Thus, this study shows the potential of collecting travel behavior data from social media in a cost-effective way to be used in future transportation studies.

1.5 Thesis Organization

This thesis is divided into several chapters. Chapter one introduces the topic with background and main objectives of the study. The information provided in chapter one justifies the selection of this topic as an important and timely research matter.

Chapter two presents the data collection efforts in detail. Important discussions on data filtration are included in this chapter.

Chapter three to chapter six present the methods developed in this study. Chapter three presents the classification of users into two different groups: residents and tourists. The results of the classification techniques are reinforced with time series analysis of tourists and residents Twitter activities. Chapter four shows various clustering methods applied to find recent tourists attractions as well as the residents' point of interests from their visited locations. Chapter five analyses tourists' destination patterns using advance modeling framework including Markov chain and CRF. And lastly, chapter six presents the destination choice model of resident users.

In the final chapter the findings and the limitations of the study are summarized. Based on the conclusions some recommendations are stated in this chapter with some future research scopes.

CHAPTER TWO: DATA COLLECTION AND FILTERING

2.1 Data Sources

In this study, we used Twitter as our major data source. The major advantages of using Twitter data include easy and free availability, and large sample size. These also come with the intrinsic disadvantages of large volume of unnecessary information making the data collection and cleaning a crucial step. Twitter provides free APIs to collect real time Twitter streams and historical tweets. We have collected Twitter data using its Streaming API and REST API in several steps. In the first step we collected data for about 4 weeks and applied various filtrations. This segment of the data is used for user classification and clustering. In the second step we utilized a large and more extensive data source for advanced modeling of purposes. We have also utilized census tract based demographic, infrastructure and economic data base for resident destination choice model. Apart from Twitter, the data sets used for the various segments of this study includes:

1. 2016 census tract of Florida (in ArcGIS shape file)
2. 2015 landuse data base of Florida (in ArcGIS shape file)
3. 2015 economic database of Florida
4. Florida point of interests (POI) database from Florida's geographic data library ("Florida Geographic File Database," 2008)

This chapter mainly focuses on the data collection and filtration parts of the study. Specific steps undertaken for data preparation are discussed in chapter four and chapter six. In this way we have ensured the flow of this report and also, tried to make sure to put the right context in right place.

2.2 Data Collection

Real time Twitter contents are downloaded using its Streaming API from March 29, 2017 to April 24, 2017 within a geographic boundary. The primary search focused on Central Florida region, defined by the coordinates -82.059860, 27.034087 (lower left corner of De-soto County) and -81.153310, 29.266654 (a corner of Volusia County). However, not all the geo-tagged tweets extracted from the search process are within this boundary. Collected data also included tweets from the users who did not have any tweets tagged with a latitude and longitude values but their profile information stated that they were from Florida; this is not unusual as explained in Twitter Developer Documentation ("Twitter Developer Documentation: Streaming API," 2006). Locations of the geo-tagged tweets are plotted in Figure 1. We find that geo-tagged tweets are spread across the whole state of Florida instead remaining within the defined boundary of Central Florida region only. This motivated us to run our analyses for the whole state of Florida.

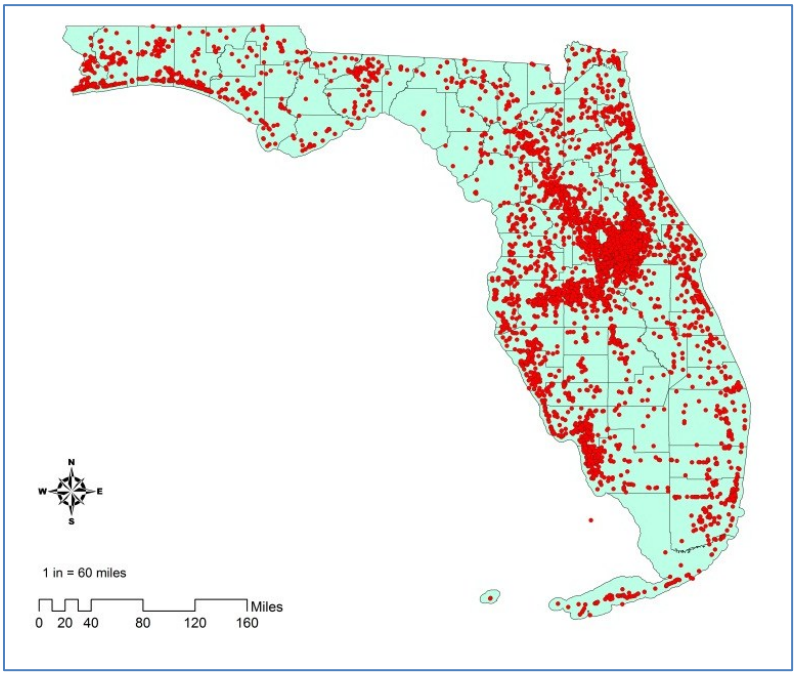
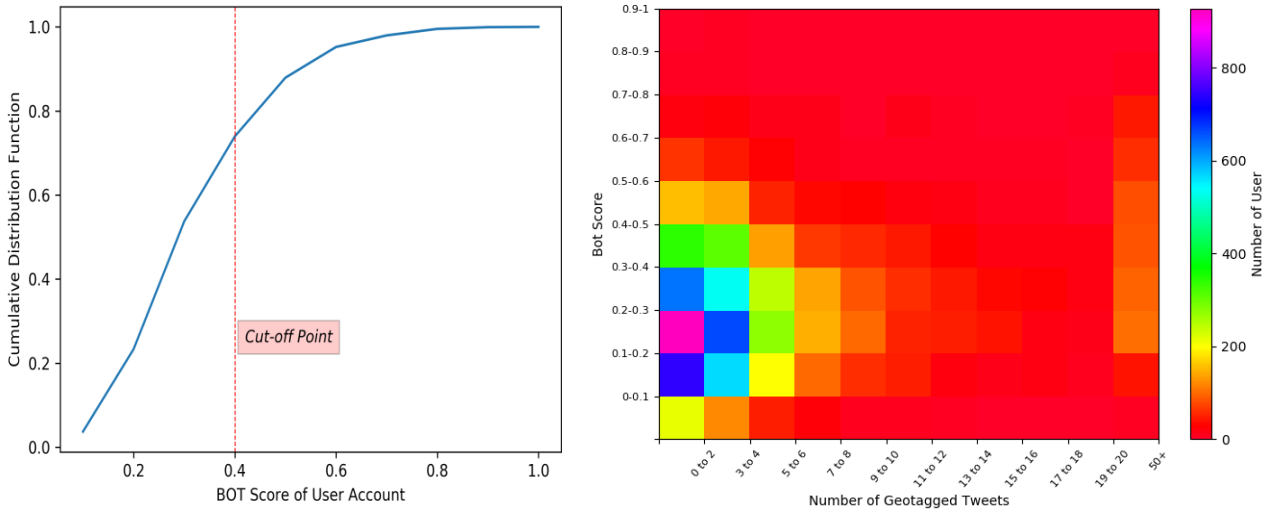


FIGURE 2.1: Geo-tagged Tweets Collected from March 29, 2017 to April 24, 2017.

2.3 Data Filtration

Since we are interested to analyze geographically active users, as a first step of the filtering process, users with at least two unique geo-tagged tweets are selected for further analysis. This yielded 8,707 users out of 66,919 users. In the second step, we filtered out organizational or any promotional/advertising accounts. For that purpose, we collected the BOT score which can be interpreted as the probability that the user is a bot ("Botometer," 2014). A social BOT can be defined as a sophisticated software program designed to interact like any human user on a social media platform (Woolley, 2016). Botometer provides the bot-likelihood score of a user by retrieving the recent activities of the user and analyzing various features such as content, sentiments, friends, networks etc. (Davis et al., 2016). Figure 2.2 (a) shows the cumulative density function (CDF) of user BOT scores and Figure 2.2(b) presents the number of users under different range of BOT scores and number of geo-tagged tweets.



(a) **(b)**

FIGURE 2.2: (a) Cumulative Distribution Function of user BOT scores and (b) Heat-map of the number users in a specific range of BOT score and number of geo-tagged tweets.

A higher BOT score indicates that a user is more likely to be a social bot. However, there is no

defined threshold of this value to classify a user as a bot or not a bot. In this study, we manually reviewed a randomly selected sample of user profiles and determined whether they are bot/organizational accounts or humans. Based on our observations, we decided to select a cut off value of 0.4 and thereby keep the users with BOT scores equal to or less than 0.4 in a different set and omit the users with BOT score greater than 0.4 from further analyses. This reduced the sample to 6615 user accounts. To collect the tweets of these 6615 users, we used Twitter REST API ("Twitter Developer Documentation: REST API," 2006) which gives the most recent 3200 tweets of a user. We were able to collect the tweets from 6519 users as rests of the user profiles are not public. Finally, the data set contained 676,864 tweets from 6519 users' one month time line (data collection period) of which 108,560 are geo-tagged tweets with 36,157 unique coordinates (see Table 2.1).

TABLE 2.1: Dataset Description (Phase-1).

Dataset: Phase 1	
Data Collection Period	March 29, 2017 to April 24, 2017
Total Users	66,919
Total Tweets	635,787
Total Geo-tagged Tweets	94,333
Sample for User Classification (Users with at least 2 geo-tagged tweets and BOT score ≤ 0.4)	
Total Users	6615
Total Tweets (Most Recent Tweets of 6519 users)	676,864
Total Geo-tagged Tweets	108,560
Sample for Location Clustering (Users identified and validated with ground truth)	
Total Users	3,088
Tourists	1,600
Residents	1,488

Sample for Location Clustering (Users identified and validated with ground truth)	
Coordinates inside Florida, from March 29 to April 24, 2017	
Tourists	12,470
Residents	24,116
Dataset: Phase 2	
Sample for CRF and PLSMNL (Users with BOT score ≤ 0.4 and at least 2 geo-tagged tweets within March 29, 2017 to October 10, 2018)	
Total Users	11,122
Users with posted place in their Twitter profiles	7039
Number of Tourists	2438
Number of Residents	4601
Total coordinates inside Florida (within March 29, 2017 to October 10, 2018)	
Tourists	35,680
Residents	77,751

As research progressed, we continued to collect Tweets using the Streaming API within the specified boundary. We separated a second data set starting from March 29, 2017 upto October 10, 2017. This provided about 1.6 million Tweets from nearly 156,000 users. We ran the same filtration procedure on the second data set and found 11,122 users with BOT score less than or equal to 0.4 and with at least two geo-tagged posts from March 29, 2017 till October 10, 2017. From user posted places we labeled nearly 7039 users as resident and tourists. We found 2438 tourists and 4601 residents in the phase-2 data set. Using REST API we then collected the latest 3200 Tweets of these 7039 users. Within March 29, 2017 till October 10, 2017 these users have posted 732,590 tweets among which 113,431 are geo-tagged. Table 2.1 summarizes the data collected for the phase-2 analyses. In the subsequent two chapters we have utilized the first data (phase-1) set for user classification and location clustering. The second data (phase-2) set is used for tourists' next destination type prediction model (chapter five) and in residents' destination choice model (chapter six).

CHAPTER THREE: USER CLASSIFICATION

3.1 Introduction

To extract behavioral information from different demographic groups of social media users, it is necessary to classify users based on some specific criteria. Previously, user profiles have been classified depending on the type of application. McNeill et al. (McNeill et al., 2016) used a simple heuristic approach counting certain locations in a user's geo-tagged tweets to identify home and work locations. Abbasi et al. (Abbasi et al., 2015) used geo-tagged tweets to identify the most active tourists inside Sydney who visited the place within the data collection period (four weeks data in four phases). The users present in only one (or two) phases of the data collection period with at least 9 unique geo-tagged tweets were considered as active tourists(Abbasi et al., 2015). To classify the locals and tourists in Barcelona, Manca et al.(Manca et al., 2017) proposed a heuristic algorithm which considered the values in 'user location' of the tweets and the duration of users inside the studied region. Manca et al. (Manca et al., 2017) limited the duration to 20 days, while as per the state of the art practices the users who publishes all his/her objects online (photos, personal views, check-ins etc.) within 30 days are considered as tourists(Girardin et al., 2007; Theobald, 2005). In another study(Andrienko et al., 2013), users who were at least 10 days inside and at most 8 days outside the greater Seattle area within 2 months of data collection period were considered as local residents.

As found from the literature, the time and location stamps of tweets have been widely used, although majority of these studies used only a single feature to separate tourists and resident users. In this study, we propose a comprehensive classification analysis starting from a simple heuristic approach with a single feature to an ensemble classification method with multiple

geographical features extracted from user profiles. Starting from a simple heuristic classifier, we propose several classification techniques to improve prediction results. To validate our results, we have used the self-reported place in a user's Twitter account profile as a ground truth. Out of the 6519 users, about 5123 users have their 'Place' field filled and for the rest of the users that field is empty. One important aspect of our approach is that, it does not have to use the content of the tweets. However, we are able to extract at least state level locations (for places inside USA) or country level location (for places outside USA) for 4696 users. Out of 4696 users, 2331 users are residents as they have stated Florida in their place field and the rest 2365 are labelled as tourists.

3.2 Heuristic Classifier

We have used user's location information extracted from the coordinates to classify whether he/she is from Florida or not. A simple heuristic approach is proposed based on the assumption that during the night users are more likely to tweet from their homes. In this method, we denote the users with most of their geo-tagged tweets during certain hours of the day inside the Florida's geographical bounding box as 'Residents' and the rest as 'Tourists'. We propose that during night the user normally post more Tweets from their home location. Therefore, we have selected a window of six hours from 12 am to 6 am in the morning and calculated the number of tweets posted during these hours and during the other hours (from 6 am morning till 12 am at night). The heuristic is presented in Algorithm 1 (Figure 3.1). The results from this algorithm are validated using the ground truth data extracted from users' posted paces in their Twitter profiles.

Algorithm 1: Tourist Identification from Tweet Coordinates

Input: Set of users (U) with coordinates of their geo-tagged tweets (C) posted at any time of the day (T)

Output: User Sets identified as Tourists (U_T) and Residents (U_R).

for user i in user set U :

extract the set of all coordinates throughout the day: C_iT_i

extract the set of coordinates associated with time frame between 12 am and 6 am: $C_iT_{i(12-6)}$

for the coordinates in set $C_iT_{i(12-6)}$:

extract the coordinates in set $(C_iT_{i(12-6)})$ which were within Florida boundary:

$C_{i(FL)}T_{i(12-6)}$

extract the coordinates in set $(C_iT_{i(12-6)})$ which are outside Florida boundary:

$C_{i(Others)}T_{i(12-6)}$

if the Number of element in 1st set $N[C_{i(FL)}T_{i(12-6)}]$ is **greater** than Number of elements in 2nd set $N[C_{i(Others)}T_{i(12-6)}]$:

append user i in the Resident user list U_R

else:

append user i in the Tourist user list U_T

FIGURE 3.1: Heuristic Algorithm for Tourist Identification (Continued)

The accuracy of the proposed heuristic classifier has been found as 79.09%, i.e. out of 100 instances it was able to label 79 of them correctly either as resident or tourist.

3.3 Supervised Machine Learning Techniques

To test the performance of the proposed heuristic classifier, we have applied three supervised classification techniques: Decision Tree (Safavian and Landgrebe, 1990), K-Nearest Neighbors (KNN) (Manning et al., 2008), and Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000). For each user, five features are extracted:

- **Feature 1:** The ratio between the number of geo-tagged tweets inside Florida and the number of geo-tagged tweets outside of Florida during nighttime (for entire time period).
- **Feature 2:** Mean distance of the successive coordinates of users' geo-tagged tweets.

- **Feature 3:** Standard deviation of distance between successive coordinates of users' geo-tagged tweets.
- **Feature 4:** Radius of gyration
- **Feature 5:** 100 mile distance between successive coordinate

Here radius of gyration is used as an indicator of how far and how often a user moves and is defined as (Bolivar, 2014):

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2} \tag{3.1}$$

Where, n is the number of geo tagged posts of the user, and $(r_i - r_{cm})$ is the distance between the geo-tagged location of post i and the center of all the locations of that user r_{cm}. Feature 5 is a binary feature, i.e. 1 if there exists a 100 mile jump among the distance of successive coordinates and 0 otherwise.

Based on these features, we have applied a *k*-fold cross validation approach for training and prediction using the three classifiers. In a *k*-fold cross validation, the sample is divided into *k* groups and prediction function is trained using the data from (*k*-1) groups. The remaining group is used for testing the predictions made by classifier. In this method, the training and validations are iterated *k* times where in each iteration a different set of data (fold) is left out for test (Refaeilzadeh et al., 2009). We have used 10-fold cross validation as it has been successfully used in previous studies (Kim, 2009; McLachlan et al., 2005).

To evaluate model performance, we have computed accuracy, precision, recall and f-score for each classifier. Accuracy is defined as the proportion of users identified as tourists among the users who are actually tourists. Precision is the proportion of users correctly identified as tourists

among all the users who are identified as tourists and recall is the proportion of users who are correctly identified as tourists among the users who are actually tourists. F1-score combines precision and recall by calculating their harmonic mean. Figure 3.2 shows the results of the supervised classifiers along with the heuristic. The heuristic performed better among all the methods along accuracy, precision, recall and f1-score of 0.7909, 0.7911, 0.7910 and 0.7908, respectively.

3.4 Ensemble Classifiers

Since the classifiers adopted in the previous section failed to produce results better than the proposed heuristic, we explored several ensemble techniques. Ensemble is a technique where multiple classifiers are combined (Dietterich, 2010; Rokach, 2010) and which are found to work better than a single classifier. The ensemble methods applied in this study includes bagging, adaptive boosting, random forest and majority voting. Voting or majority voting accounts the output of individual classifier and reports the label that is predicted by majority of the classifiers (Rokach, 2010). Bagging uses re-sampling the training dataset in order to learn individual classifiers and then uses majority vote to report the combined classifier label (Breiman, 1996). Random Forest uses decision tree as its base classifier which also uses bagging technique to create new training sets (Chan and Paelinckx, 2008). Adaptive boosting or AdaBoost is a more complex method where in each step the models selects the training data set based on the performance of the previous step (Freund and Schapire, 1996).

We have applied these four ensemble classifiers whose results are reported in Figure 3.2. The sample was split in 70% training set and 30% test set. For AdaBoost, random forest and bagging ensembles, we used decision tree classifier as the base classifier. Random forest and bagging both

are trained using bootstrap aggregation. In adapting boosting, we used the real boosting algorithm, i.e. considering the output of decision trees as a class probability. In voting classifier, we used the input from the three supervised classifiers discussed in section 3.2 as it provided better result than using any two of the classifiers from section 3.2.

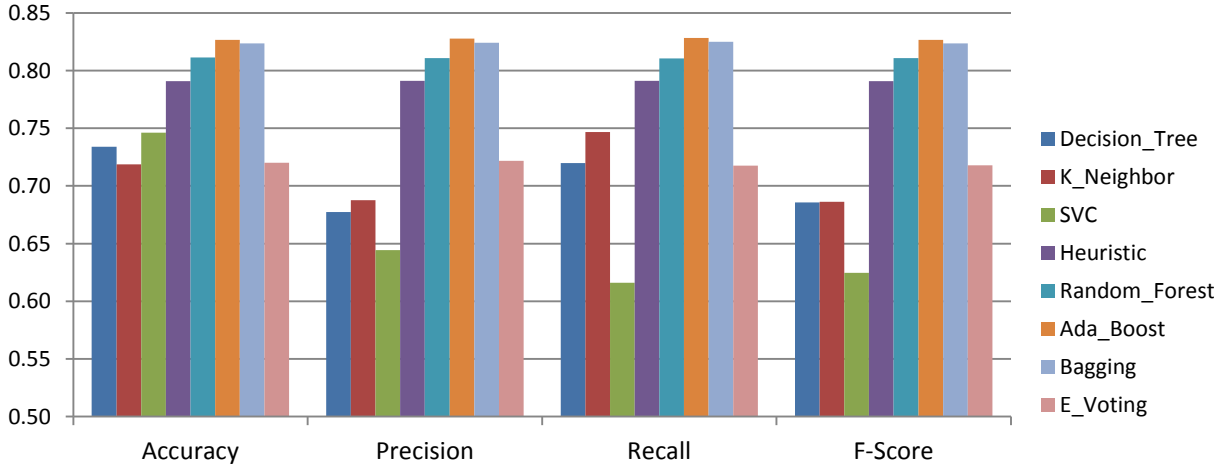
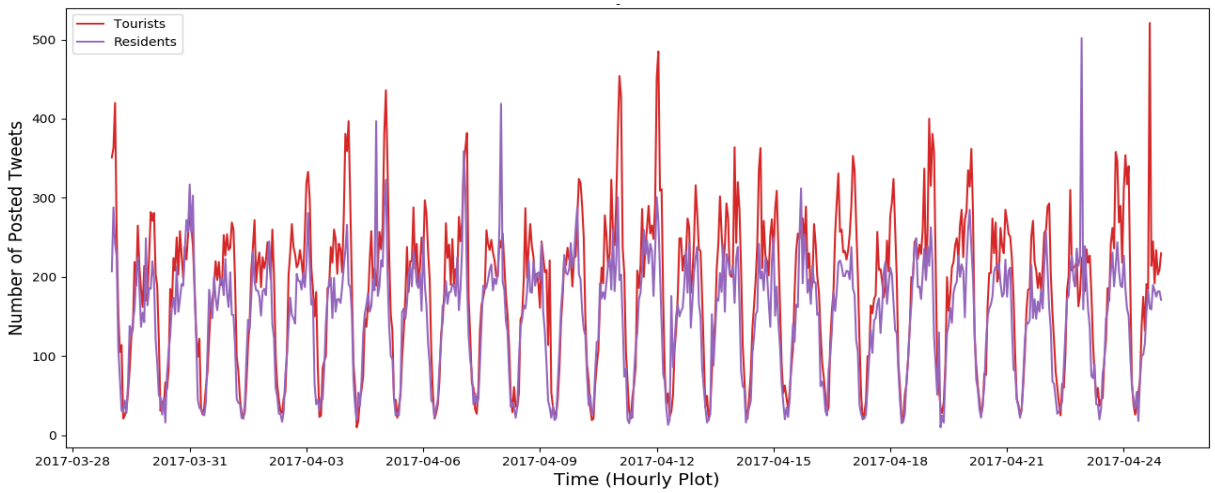


FIGURE 3.2 Comparison of Performances of the Classifiers with proposed heuristic method.

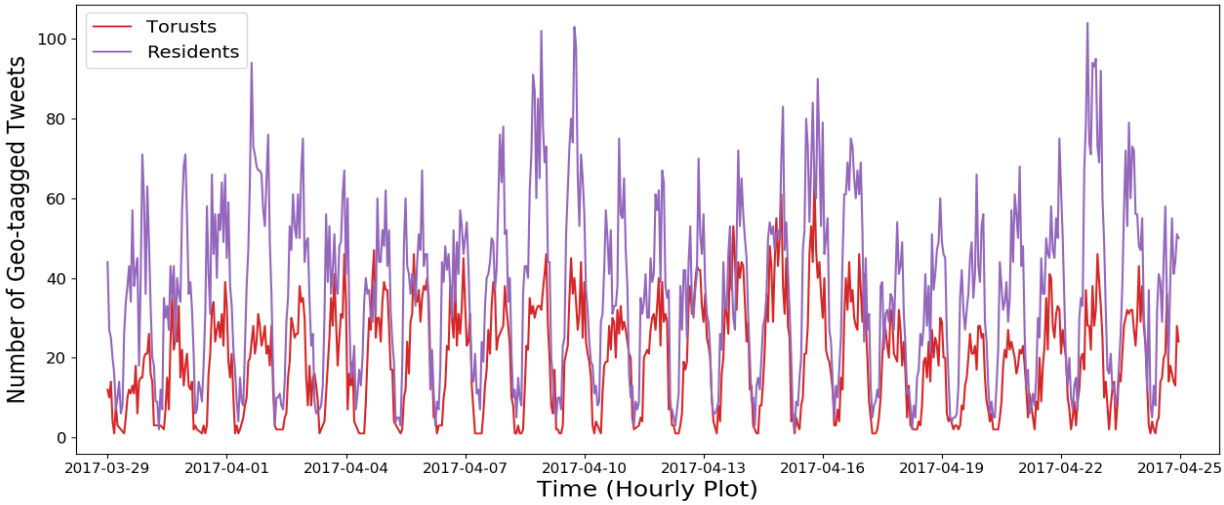
Figure 3.2 presents all the performance indices of all the classification techniques adopted in this study. Random forest, AdaBoost and bagging approach performed better than the proposed heuristic. Among the ensemble methods AdaBoost performed best with accuracy, precision, recall and f1-score of 0.8277, 0.8276, 0.8267 and 0.8267 respectively. We have also measured the performances of the ensemble classifiers using the heuristic as the base classifier and found that AdaBoost has the best performance with accuracy, precision, recall and f1-score of 0.7740, 0.7773, 0.7766 and 0.7740 respectively.

3.5 Time Series Analysis

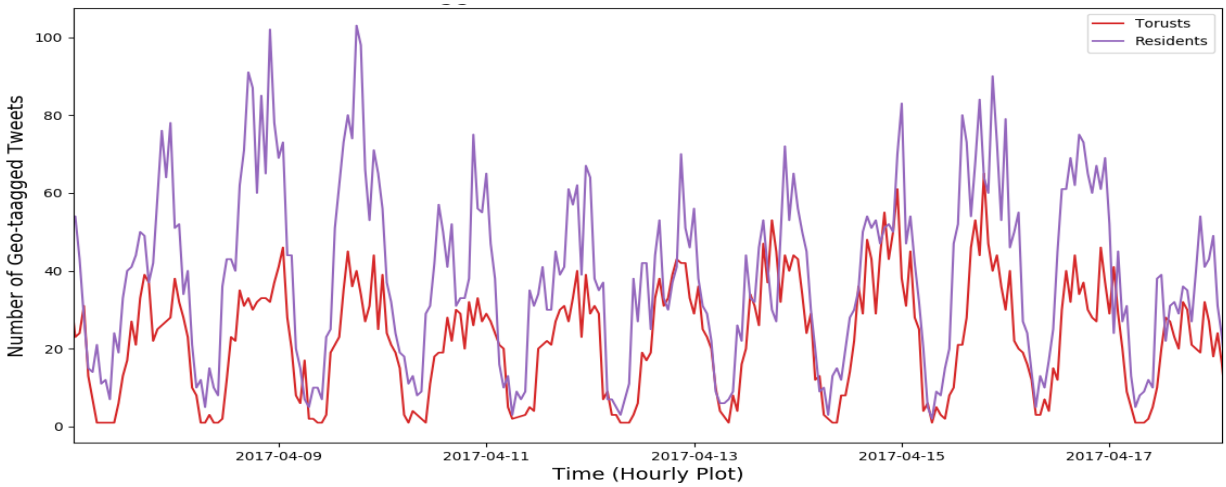
From the classified and validated dataset, we separately plotted the activity time-series for both tourists and residents. Figure 3.3 shows the activities (in terms of number of tweet posts) for both resident and tourist.



(a) Total Activity Plots



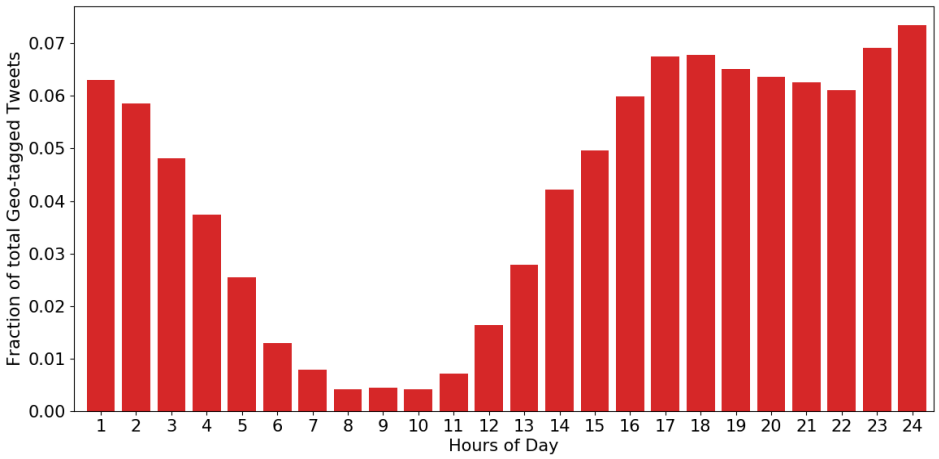
(b) Geo-tagged Activity Plots (inside Florida)



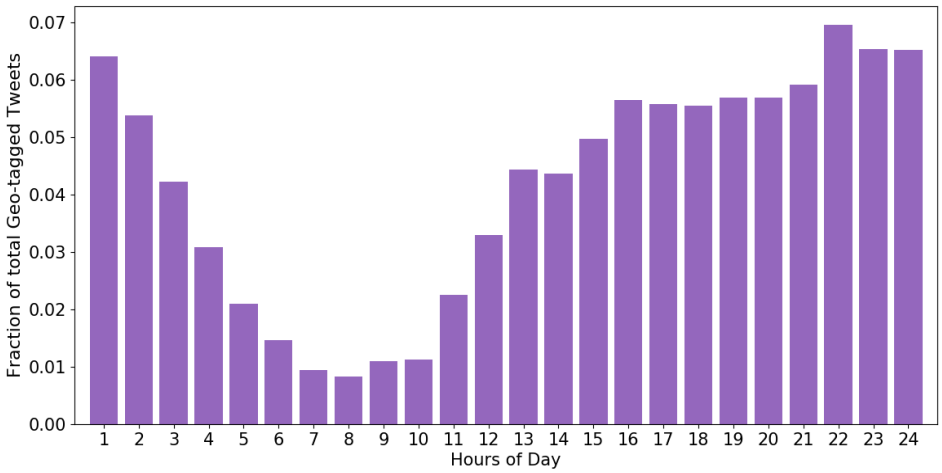
(c) Horizontal Shift in Hourly Activity

FIGURE 3.3: Daily Activity Plots of Tourists and Residents.

For the entire data collection period we plotted the activity considering all the tweet posts and also, considering only the geo-tagged posts (inside Florida State boundary). There is a repeating trend for both user groups in daily activities which reaches to maximum at the end of the day. From Figure 3.3(a) we find that the number of total posts for both resident and tourist are close to each other and in Figure 3.3 (b) we see that the number of resident geo tagged posts is greater than the number of tourist geotagged posts. In Figure 3.3 (c) we have focused a portion of Figure 3.3 (b) to better explain the activities.



(a) Hourly Activity of Tourists

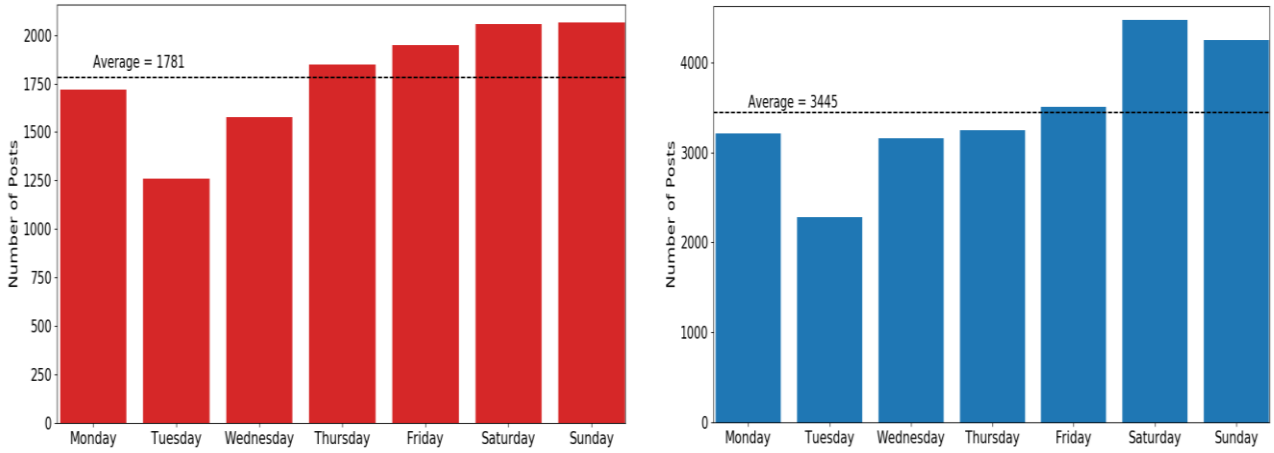


(b) Hourly Activity of Residents

FIGURE 3.4: Hourly Activity Plots. (a) Tourists, (b) Residents.

Figure 3.4 shows the fractions of geo-tagged tweets posted in different hours of the day. Figure 3.4(a) and Figure 3.4(b) shows the hourly plots for tourists and residents, respectively. Figure 3.4(a) shows that tourists remain less active from 6 am to 11 am and after that there is gradual increase in activity from 12 pm. For residents, the less active hours are from 7 am to 10 am and activities start increasing a little early in morning, from 11 am. This shift is clearly viewed in Figure 3.3(c) in the continuous geo-tagged activity posts. The highest activity for residents is around 10 pm and for tourists is around midnight.

The weekly activity of the resident and tourists are shown in Fig.9.



(a) Tourists Weekly Activity **(b)** Resident Weekly Activity

FIGURE 3.5: Weekday and Weekend Geo-located Activities. (a) Tourists' Weekly Activity Plot and, (b) Residents' Weekly Activity Plot.

From Figure 3.5(a) and Figure 3.5 (b) we see that on average Tourists posted 1781 number of geo-tagged tweets and residents posted 3445 geo-tagged tweets. Interestingly, for tourists the daily posted tweets are greater than the average on some weekdays (Thursday and Friday), whereas for residents the average number is exceeded only on weekends (Saturday and Sunday). This behavior is normal as tourists do come to visit places on days other than the weekends.

These time series analyses further validate the classification techniques of this study. The two

groups of users have distinguishable temporal patterns. Combining the spatial clusters on temporal basis, i.e. clustering for the locations posted during a range of periods of day or on different days of a week it is possible to find the spatio-temporal densities of tourists around different parts of study area.

3.6 Summary

This chapter described the classification of users into resident and tourist. Starting from a simple heuristic classifier, we propose several single and ensemble classification techniques to improve prediction results. The self-reported place in a user's Twitter account profile has been used as a ground truth to validate the results. One important aspect of the approach is that, it did not use the tweet contents (i.e. texts, hash-tags, mentions etc.), rather the features extracted from the geo-locations to achieve nearly 80% efficiency in supervised ensemble classification method. Using the resident and tourist (identified and validated in heuristic method) Twitter posts we have demonstrated the activity patterns in time series plots. The results showed distinguished behavior for resident and tourist which further validate hour assumptions on the ground truth.

CHAPTER FOUR: TRAVEL PATTERN FROM LOACTION CLUSTERING

4.1 Introduction

We propose different clustering techniques to find out the most visited places by tourists and residents. In this study we have applied three clustering methods: K-Means (Kanungo et al., 2002), DBSCAN (Ester et al., 1995), and Mean-Shift (Comaniciu and Meer, 2002) in order to find the spatial patterns of destination choices made by tourists and residents. The methods have been chosen based on their efficient applications in similar types of researches found in the literatures. For the clustering purpose we used the geo-tagged posts collected in phase-1 of data collection period (Table 2.1). The landmarks close to the cluster centers represent the popular destinations visited by tourist and most commonly visited locations of the resident. We put on detail discussion about parameter selection of the clustering methods and finally measure their performances based on some internal validation measures.

4.2 Spatial Clustering

Despite its wide adoption, few studies have investigated tourist behavior using Twitter data. Abbasi et al.(Abbasi et al., 2015) considered tourists who are traveling both into and outside of Sydney within four weeks. Analyzing geo-tagged tweets, they could identify the most visited places by local residents and tourists. Using geo-tagged tweets, Lee et al.(Lee et al., 2016) demonstrated the growth of activity space of 116 Twitter users over 17 weeks and determined their major activity locations. In literatures different types of spatial clustering techniques have been used as a popular tool to find the groups of closely related destinations. Most of the clustering methods like the partitioning clustering methods such as K-Means (Kanungo et al.,

2002), hierarchical clustering such as Ward's method(Ward Jr, 1963), and density-based clustering methods such as density-based spatial lustering of applications with noise or DBSCAN (Ester et al., 1995) uses distance measure (i.e. Euclidean distance) to group the similar (nearby) objects together. Zheng et al. (Zheng et al., 2012) used DBSCAN on geo-tagged photos to identify tourist regions of attractions. Majid et al. (Majid et al., 2013) used DBSCAN to find tourists locations from geo-tagged photos. In similar kind of data sets other studies applied K-Means(Kennedy and Naaman, 2008) and Mean-Shift(Yin et al., 2011) clustering methods for location identification.

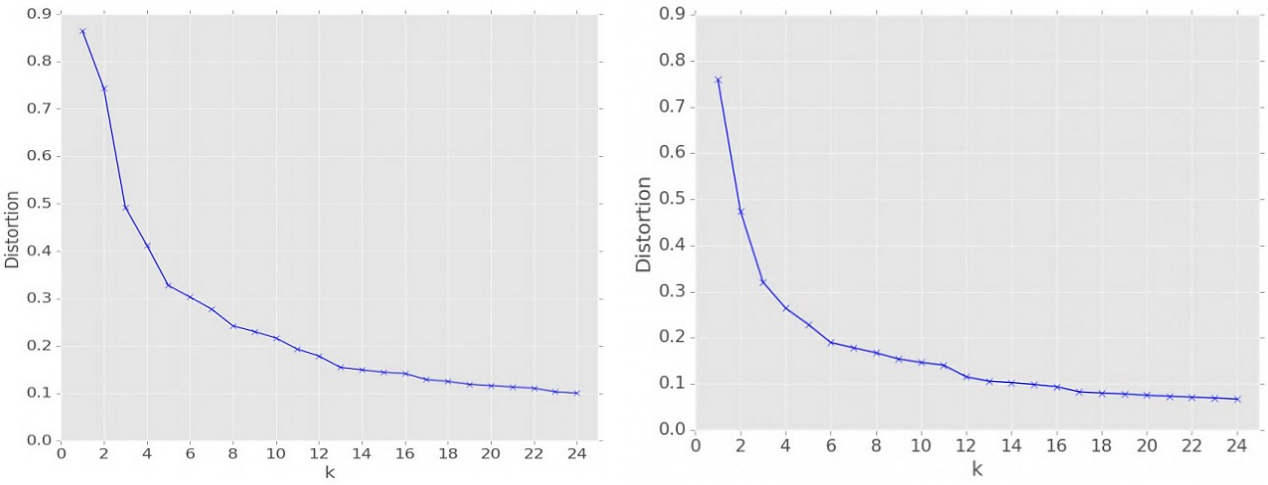
K-Means clustering algorithm divides a set of n observations in a d-dimensional space into k number of sets ($k \leq n$) in a way that the within-cluster sum of squares or mean squared distance is minimized. With input parameter k (number of expected clusters), the algorithm uses Euclidean distance as a metric and variance as a measure of cluster scatter.

Mean-Shift clustering, popular as the mode seeking algorithm (Cheng, 1995), locates the maxima of a density function. The iterative process starts with initial estimation and typically uses Gaussian Kernel Density function to re-estimation of the mean from the weight of nearby points. It requires a parameter bandwidth that determines the shape of kernel density distribution (Comaniciu and Meer, 2002).

DBSCAN is a density-based clustering algorithm which forms a set of points and groups together the points that are packed closely within a given threshold distance in space and marks points as outliers that lie alone in low density regions. DBSCAN requires two parameters, i.e. epsilon which is the maximum distance between two samples to be considered in the same neighborhood and the minimum number of points required to form a dense region (Ester et al., 1996).

4.3 Parameter Selection

The selected clustering methods require different boundary parameters. We determined the optimum parameters to start the clustering process. It should be noted that the perfect values of the parameters cannot be known beforehand. Based on some preliminary analyses on the data sets we select the most likely values to start the clustering process. The parameters selected are therefore, liable to changes based on the outputs as we try to find better results.

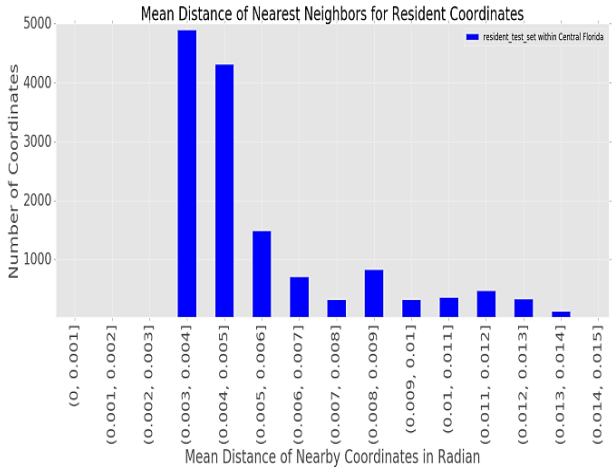


(a) Elbow plot for Residents **(b)** Elbow plot for Tourists

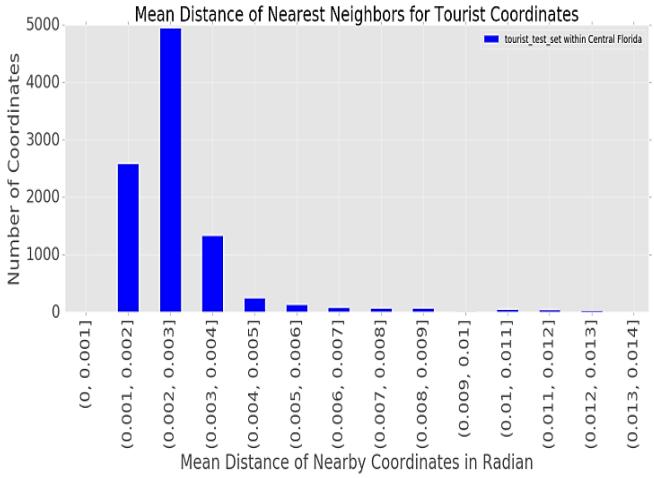
FIGURE 4. 1 Parameter Selection for K-Means clustering method.

In *K*-Means clustering, each observation is assigned to one of the *k* number of clusters, where *k* is decided by the analyst. To select *k*, we have used an elbow plot which is a 2-dimensional plot of the distortion (percentage of variance) vs. the number of clusters (*k*) (Bholowalia and Kumar, 2014). The optimum number of clusters should be chosen in a way that adding another cluster does not significantly reduce the variance of the data (Bholowalia and Kumar, 2014). From Figure 4.1(b), the optimum number of cluster for tourists is 6. As there is also another bend in the region with $k = 12$, hence we also tried with 12 clusters for tourists. Similarly, from Figure 4.1 (a) for the

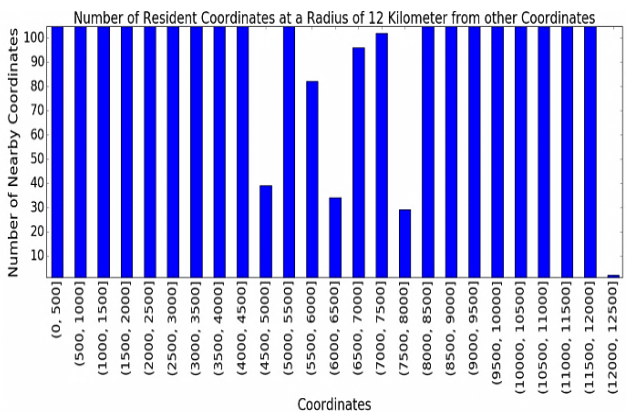
residents there is no clear elbow and therefore we have run the clustering model with two different *k* values (k=8 and k=13) and considered the better result.



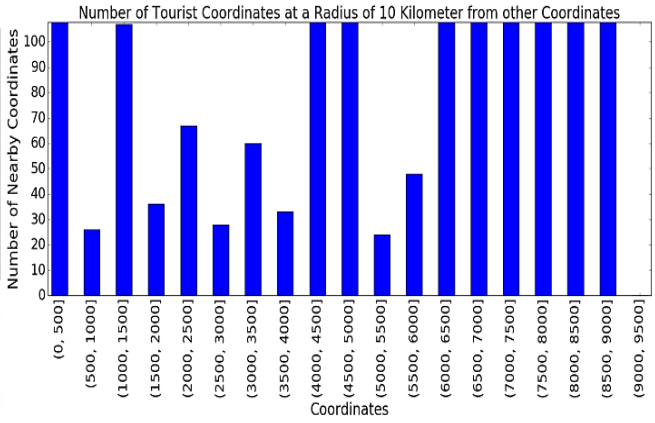
(a) Mean Distance of each points from the other points for Resident Coordinate



(b) Mean Distance of each points from the other points for Tourist Coordinate set



(c) Minimum number of coordinates within 12 km radius for Resident Coordinate set



(d) Minimum number of coordinates within 10 km radius for Tourist Coordinate set.

FIGURE 4.2: Parameter Selection for Mean-Shift and DBSCAN clustering methods.

In order to find a realistic value of the bandwidth of Mean-Shift algorithm, we determine the mean distance of each point to all of its nearby neighbor points and plotted a histogram of the number of coordinates vs. the ranges of mean distance. For residents, about 19,000 coordinates are within a radius of 0.025 radian (159.275 kilometer) from all other resident coordinates; and

for tourists, more than 9,500 coordinates are within a radius of 0.015 radian (95.565 kilometer) from all other tourist coordinates. Since the data was collected within the Central Florida region, we find the radius for the coordinates inside the data collection boundary. For the coordinates inside Central Florida, we find that most of the resident coordinates are within 0.006 radian or 38.226 kilometer (Figure 4.2(a)) from the other resident coordinates and most of the tourist coordinates are within 0.004 radian or 25.484 kilometer (Figure 4.2(b)) from the other tourist coordinates. Thus, in the Mean-Shift algorithm, we have used the bandwidth parameter of values 0.006 radian and 0.004 radian for clustering resident and tourist locations, respectively.

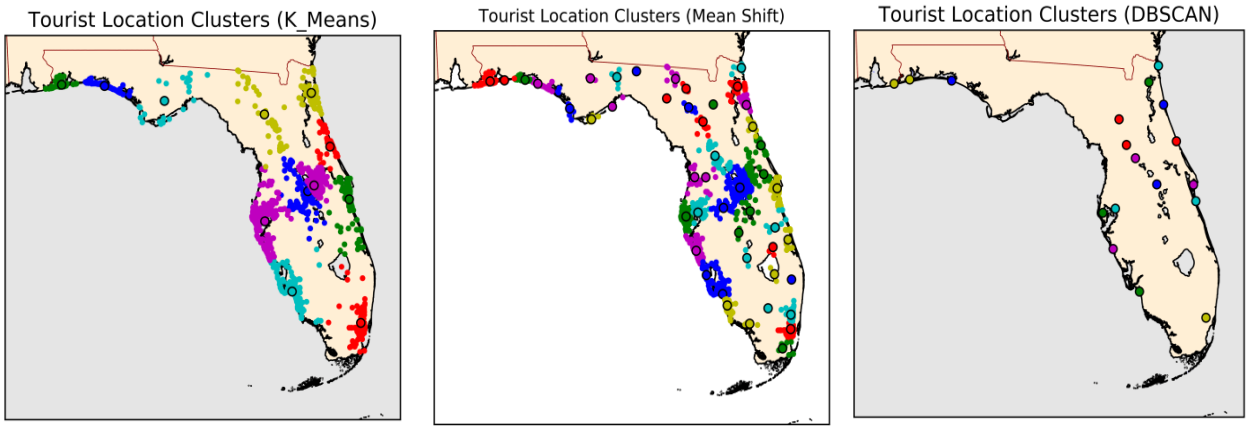
Although the parameter 'epsilon' in DBSCAN is similar to 'bandwidth' of Mean-Shift, using the same value in both methods leads to a misrepresentation of the sample data. With larger epsilon value, DBSCAN will reduce the number of clusters as with each iteration it's core points will reach more neighbors within specified epsilon. Whereas, Mean-Shift will go for the densest region with radius set equal to the bandwidth. In DBSCAN, we have the freedom to choose the distance in kilometer of earth surface distance between two coordinates (as DBSCAN uses 'haversine' distance metric instead of 'Euclidian' in K-Means and Mean-Shift).

In our study, we have selected epsilon value equal to12 kilometer for resident coordinates and 10 kilometers for tourist coordinates. To find the minimum number of points we plotted the number of coordinates within a minimum distance from each coordinates, the minimum distance being equal to 12 kilometers for resident coordinates (Figure 4.2(c)) and 10 kilometers for tourist coordinates (Figure 4.2 (d)). The minimum number of samples for DBSCAN is selected in a way that too many points do not fall as noisy points or outliers. From Figure 4.2(c-d) the minimum sample is selected to be 25 for tourist location clustering and 50 for resident location clustering.

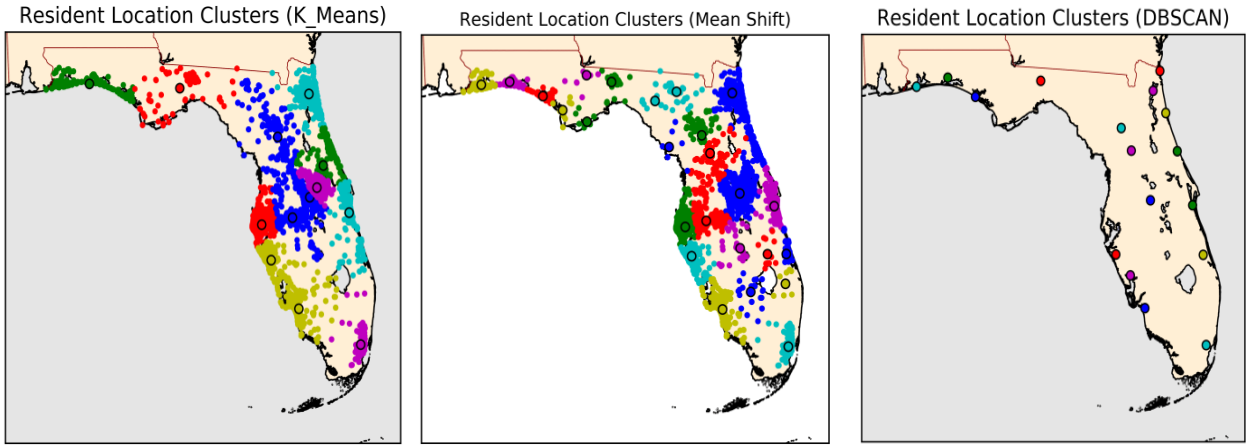
4.4 Clustering Results

Our main goal of clustering is to find the most visited regions/areas by tourists and residents inside the state of Florida. The only similarity of the points inside a cluster is that they are nearer to each other than other points in other clusters and/or from points without clusters (outliers in DBSCAN). From the output clusters, we have found the centers of each cluster in all the three methods. Google Places API ("The Google Places API Web Service," 2017) is used to extract the street level address information of the places associated with the cluster centers' coordinates. Also, the number of unique users and number of sample coordinates forming each cluster are also determined. The rationale of using three different approaches is to find the method which best serves our goal, i.e. grouping the coordinates into distinctive clusters in a real time fashion. The centers of clusters in all the three different methods are shown in Figure 4.3.

In K-Means clustering, k=13 provides good results for resident's location clustering. For Mean-Shift there are 25 clusters and for DBSCAN we have found 18 clusters. The numbers of clusters for tourists were found to be 12, 47 and 17 for K-Means, Mean-Shift and DBSCAN clustering techniques, respectively (Figure 4.3). It should be noted that the number of outliers (coordinates without any cluster) was 543 (of 207 users) for tourist location clustering and 879 (of 212 users) for resident location clustering.



(a) K-Means **(b) Mean-Shift** **(c) DBSCAN***



(d) K-Means **(e) Mean-Shift** **(f) DBSCAN***

**shown only the centers for DBSCAN as there are too many cluster in DBSCAN*

FIGURE 4.3: Clustering of Tourist Coordinates: (a)- (c); and Resident Coordinates: (d) - (f).

The dots represent the coordinates within specific clusters separated by different colors and the comparatively larger dot with black border represents the centers of the clusters.

Table 4.1 presents the detail information of the top clusters (based on the total number of unique users and total number of coordinates in each cluster) in all the three methods. We did not report the clusters with too few users and sample coordinates.

TABLE 4.1: Clustering Results for Tourist Coordinates.

Sl. No	Street Address	City/Area	Nearby Landmarks	Number of Unique Users	Number of Points
	K-Means				
1	4050 Kingsport Dr	Orlando	Universal Studios, Island of Adventure	941	4788
2	Coronado Springs	Kissimmee	Walt Disney World Resort, Animal Kingdom and Theme Park, Epcot	908	4801
3	2425 N Rocky Point Dr	Tampa	Cypress Point Park, Tampa Bay, Tampa International Airport	117	571
4	9974 NW 87th Terrace	Doral	Francis S. Taylor Wildlife Management Area	96	460
5	38 Bramble Grove Pl	Santa Rosa Beach	Deer Lake State Park, Grayton Beach State Park	80	394
6	8084 Estero Blvd	Fort Myers Beach	Lovers Key State Recreation Area	71	335
7	Pineda Causeway	Satellite Beach	Banana River Aquatic Preserve, Manatee Cove Golf Course	107	305
8	1662 Century Acres Ln	Jacksonville	Julington Creek Golf Club	70	217
9	4731-4735 White Tail Ln	Sarasota	Stoneybrook Golf and Country Club, TPC Prestancia	44	199
10	522 Fairpoint Dr	Gulf Breeze	Shoreline Park, Pensacola bay bridge	45	182
11	732 Iowa St	Daytona Beach	Daytona Rising-Daytona International Speedway, Daytona Beach International Airport	59	147
Mean Shift					
1	8519-8527 Sand Lake Shores Dr	Orlando	Orange County Convention Center, Rosen Inn At Pointe Orlando	1251	9230
2	1201 NW 89th Ct	Miami	Mall of the Americas, Doral Central Park	69	286
3	6727 126th Ave N	Largo	Largo Golf Course, Travel World RV Park,St. Pete–Clearwater International	59	276

Sl. No	Street Address	City/Area	Nearby Landmarks	Number of Unique Users	Number of Points
			Airport		
4	1514-1598 N Florida Ave	Tampa	Water Works Park, The Florida Aquarium, Amalie Arena, Tampa General Hospital	65	243
5	120-130 Cullman Ave	Santa Rosa Beach	Grayton Beach State Park, Grayton beach, Deer Lake State Park	57	242
6	Martin Andersen Beachline Expy	Merritt Island	Cape Canaveral, Cocoa Beach, Kennedy Space Center	80	238
7	4325 E Memorial Blvd	Auburndale	Schalamar Creek Golf & Country Club Community, Sadle Crek Park	68	181
8	7219 Antigua Pl	Sarasota	TPC Prestancia-Golf Club	40	166
9	732 Peake's Point Dr	Gulf Breeze	Pensacola NAS(Naval Air Station) DRMO, Blue Wahoos Stadium, Pensacola Bay Bridge	41	165
10	21381 Widgeon Terrace	Fort Myers Beach	Estero Bay Preserve State Park, Fort Myers Beach	43	158
11	6781-6785 Southern Oak Ct	Naples	Clam Pass Park, Kensington Glof Course	29	130
12	5200 Hancock Rd	Southwest Ranches	Sunshine Ranches Equestrian Park, Flamingo Gardens, Everglades Wildlife Management Area	26	126
13	341-349 Regatta Bay Blvd	Destin	Henderson Beach State Park, Emerald Bay Golf Course, Mid Bay Bridge	26	113
14	724 S Palmetto Ave	Daytona Beach	Daytona Beach, Samuel L. Butts Archeological Park	42	111
15	2598 Pit Bull Ln	Mims	Buck Lake Conservation Area, Seminole Ranch Conservation Area	50	102
16	12 Grouper Hole Dr	Cape Haze	The Boca Grande Resort & Hotel,Cape Haze Aquatic Preserve	26	71
17	678-944 Woodlawn Rd	St. Augustine	Northeast Florida Regional Airport, Twelve Mile Swamp Conservation Area	22	58

Sl. No	Street Address	City/Area	Nearby Landmarks	Number of Unique Users	Number of Points
DBSCAN					
1	10232 Turkey Lake Rd	Orlando	Orange County Convention Center, Aquatica, SeaWorld's Waterpark Orlando, Seaworld Orlando	1280	9490
2	2650-2660 W 76th St	Hialeah	Carl F Slade Park	85	412
3	9155 Charles M Rowland Dr	Port Canaveral	Disney Cruise Line, Carnival Cruise Line-Port Canaveral	73	198
4	45 Town Center Loop	Santa Rosa Beach	Gulf Place Getaway-vacation spot	70	350
5	1900-1998 E 13th Ave	Tampa	Centro Ybor Complex, Historic Ybor city	66	238
6	Unnamed Road	Fort Myers Beach	Lover's Key State Park - beach, Recreation Area	62	302
7	12547 66th St N	Largo	Vacation Village RV Resort	59	275
8	720 Peake's Point Dr	Gulf Breeze	Shoreline Park, pensacola bay bridge	38	153
9	337 N Tamiami Trail	Osprey	Bay Preserve at Osprey-reception venue, Historic Spanish Point	35	165
10	721 Ballough Rd	Daytona Beach	Daytona Beach Brodwalk, Daytona Lagoon	35	89

Some city/areas such as Orlando, Tampa, Daytona Beach, Fort Myers, St. Augustine, Gulf Breeze, and Santa Rosa Beach are common as a center in all three clustering methods. The nearby landmarks column in Table 4.1 reports the famous visiting places and tourist spots within 3 kilometer of the centers. To qualitatively validate the clustering output, we have considered some of the most recent statistics about tourism spots in Florida. According to an FDOT report (Eye et al., 2017), top most annual visitor attendance in Central Florida for the year 2014 were found in Magic Kingdom, Epcot, Animal Kingdom, Hollywood Studios, Universal Studios, Island of

Adventure, Sea World etc. All of these places are within 3 km radius of cluster centers found with K-Means method (colored sections of Table 4.1). The latest tourist attractions in Florida are Daytona Rising and Expansion of Port Canaveral (Eye et al., 2017), which are also found from clustering outputs. The output clusters also reveal popular state parks and reserved forests and wetlands. Along with the existing facilities, clustering techniques are able to identify some emerging attractions, accommodation facilities such as Orange County Convention Center, Rosen Inn at Pointe Orlando, Centro Ybor Complex, Mall of the Americas etc.

In Table 4.2 we find the nearby locations of the residents' cluster centers.

TABLE 4.2: Clustering Result for Resident Coordinates.

Sl. No	Street Address	City/Area	Nearby Landmarks	Number of Unique Users	Number of Points
	K-Means				
1	520 S Lake Formosa Dr	Orlando	Florida Hospital, Residential Housing complexes, Menello Museum, Orlando Science Center,	792	6647
2	3301 Bonnet Creek Rd	Orlando	Grand cypress golf resort, Disney World Cast Softball Field, Disney's Port Orleans Resort	715	5133
3	I-275	Feather Sound	St. Pete-Clearwater International Airport, Golf clubs.	234	4082
4	1106 Bartow Rd	Lakeland	Lake Bonny, Philip O'Brien Elementary School, Lakeland Senior High School, Residential Area and Apartment complexes	185	1565
5	3050 Aberdeen Stables	Deltona	Sand Lakes	174	1355
6	21100-21298 NW 86th Ave	Micanopy	Paynes Prairie Preserve State Park, near I-75	106	767
7	874-898 Spiller St	Melbourne	Residential area at the bank of Indian river	105	942
8	8700-3 Western Lake Ap	Jacksonville	Residential Area, Lake Crest Condos	101	864
9	Three Oaks Pkwy	Bonita Springs	Residential Apartment, Estero High School, Golf clubs	74	820
10	7625 Kapok Dr	Sarasota	Residential Area, Lakeview Elementary School, Golf Clubs	65	993
11	111 SW 107th Ave	Miami	Residential area, Town Shopping	52	326

Sl. No	Street Address	City/Area	Nearby Landmarks	Number of Unique Users	Number of Points
			Center, Florida International University,		
	Mean Shift				
1	FL-400	Orlando	The Holy Land Experience, Millenia Plaza, near I-4	1156	13010
2	4906 E Dr M.L.K. Jr Blvd	Plant City	Industrial establishment, Residential areas, Plant City Airport,	228	1929
3	Ulmerton Rd & FL-93 & I-275	St. Petersburg	St. Pete-Clearwater International Airport, Weedon Island Preserve	188	3595
4	6621 Southpoint Pkwy	Jacksonville	Autobahn Indoor Speedway, St. Vincent's Medical Center Southeast, Residential Area	109	896
5	265 Stewart Dr	Merritt Island	Banana River Aquatic Preserve, Indian River bank	94	796
6	22132-22198 Cinnamon Ln	Estero	Residential Area, Golf clubs	76	822
7	2100 NE 30th Ave	Ocala	Suntran Station	71	504
8	4725 Hamlets Grove Dr	Sarasota	Residential Area	64	960
9	14409 Co Rd 234	Micanopy	Paynes Prairie Preserve State Park	52	312
10	10033 SW 33rd St	Miami	Residential Area, Tamiami Park	50	320
DBSCAN					
1	16326 Macon St	Clermont	Residential area, Lake Louisa State Park, Golf Club	1292	18370
2	Robles Ln	Rockledge	Residential area, Indian River bank, near US Highway Route 1.	86	767
3	4537 Emerson St	Jacksonville	Cuba Hunter Park, Church, UF Health Endocrinology	73	587
4	901 6th St	Holly Hill	Shopping Mall, Residential area.	70	224
5	22050 US-41	Estero	Residential area	64	731
6	8425 Country Park Way	Sarasota	Residential area, Shopping Mall	52	730

Most of the residents' location clusters are centered around residential areas with some major schools, shopping centers/malls, small golf courses in the radius of 3 kilometer. Some clusters are found on the recreational establishments, near the down town area and near some of the state parks.

4.5 Clustering Performance Measure

Clustering performance can be measured based on an external or internal validation technique. We utilized internal validation methods as we applied unsupervised clustering methods. Among the various validation indices, we applied Calinski-Harabasz (Caliński and Harabasz, 1974), Dunn(Dunn, 1974), Davies-Bouldin (Davies and Bouldin, 1979) and Silhouette (Rousseeuw, 1987) Index. We adopted these measures based on the accuracy and popularity of these measures in the literature, and simplicity/efficiency of implementation. Calinski-Harabasz uses the average between- and within cluster sum of squares to evaluate the cluster validity. Higher values of Calinski-Harabasz are expected for better clusters.

TABLE 4.3: Cluster Performance by Internal Validation Indices

Validation Index	Clustering Method	Tourist_Cluster	Resident_Cluster	Optimum Criteria
Calinski-Harabasz	DB_SCAN	4831	2704	Maximum
	K_MEANS	**61619**	**3.39290**	
	MEAN Shift	37103.3	27237	
Silhouette Index	DB_SCAN	0.7063	0.2090	Maximum
	K_MEANS	0.3.3438	0.3.33.300	
	MEAN Shift	**0.7661**	**0.7023.3**	
Dunn Index	DB_SCAN	0.0484	0.0833.3	Maximum
	K_MEANS	**0.6782**	**0.8444**	
	MEAN Shift	0.3738	0.3.3603.3	
Davies Bouldin Index	DB_SCAN	223.2449	23.39.663.36	Minimum
	K_MEANS	**217.0733.3**	**188.9890**	
	MEAN Shift	1486.7238	400.2237	

Dunn index is the ratio of weighted value of inter-cluster separation to weighted values of intra-cluster compactness, where separation is the minimum pairwise distance between objects in

different clusters and compactness is the maximum diameter among all clusters. A higher Dunn index indicates better clusters. Silhoeutte index gives an idea about the samples similarity with other samples within the same cluster (cohesion) and dissimilarity with other samples in other clusters (separation). It ranges from −1 to +1, where the higher the value the greater the within cluster similarity and the lower the intra-cluster similarity. For a well-separated cluster, the Davies-Bouldin Index is expected to be lower. In Davies-Bouldin, the highest value of similarities (i.e. C^1_s) between a single cluster and all other clusters is computed and this value for all the clusters (i.e. C^1_s to C^n_s) are then averaged to report the index.

From Table 4.3 it is found that according to Calinski-Harabasz, Davies-Bouldin and Dunn Index K-Means clustering has performed best for both tourist and resident location clustering. From Silhouette Index Mean Shift has found to be the best among the three methods for tourist and resident location clustering. From the clustering outputs, we find that K-Means clustering gives satisfactory results when the input parameter (number of cluster) is carefully selected. With selected epsilon and minimum number of samples, DBSCAN provides clusters with low number of unique users and points; about 4.33.3% tourist coordinates and 3.64% resident coordinates have been marked as noisy data in DBSCAN. On the other hand, Mean-Shift algorithm provides satisfactory results with the selected bandwidth and it provides most number of clusters with least average number of coordinates in each cluster. Parameter for K-Means are estimated rather easily, whereas for Mean-Shift and DBSCAN more detail procedures were adopted.

We could have selected the best clustering technique (i.e. selecting proper k for K_Means) depending on the optimum values of these indices. However, some of these index measurement methods are associated with high computational costs. For instance, measuring Dunn index becomes difficult as the number of clusters and dimensionality of the data increases. Therefore,

we started from basic data visualization to assume the initial starting parameters for the clustering techniques rather than running these time and computational-intensive methods in an iterative way. As we found out appropriate clustering results, we applied internal validation indices to comment on the best type of clustering method for the data set.

4.6 Summary

In this chapter we have utilized K-Means, Mena-Shift and DBSCN clustering techniques to visualize the spatial patterns of tourist and resident destinations. From the nearby landmarks of the top cluster centers we found that the tourist mainly cluster around the popular tourist attraction places such as theme parks, beaches, famous state parks and reserves; while the majority of the resident geo-tagged posts are found to be clustered around some of the dense residential areas with schools, shopping centers/malls and small golf courses in the neighborhood. We have also found some resident clusters around the famous tourist spots, beaches and state parks in Florida. From preliminary analyses we found the parameters of the three clustering methods and finally evaluated their performances based on some common internal validation indices. K-Means is found to perform better than DBSCAN and Mean-Shift clustering methods.

CHAPTER FIVE: TOURISTS NEXT DESTINATION TYPE PREDICTION

5.1 Introduction

Knowing where the tourist will visit next can help to build a proactive method to enhance the traffic operation of certain region. Predicting the type of next visited location is considered to have a sequential structure as an individual tourist's future activity location depends on his/her current location. Generative models such as hidden Markov models (HMM), Gaussian mixture models (GMM) etc. and discriminative models such as maximum entropy Markov model (MEMM), conditional random field (CRF) etc. can find out statistical patterns from sequential relationships between the visited places by individual tourists. These models are also probabilistic in nature as they provide a probability distribution as solution rather than a single valued. As we have longitudinal data for tourist visited locations it is therefore possible to draw meaningful relationships from their travel sequence. In this study, we have applied a probabilistic model to predict the next destination type of tourists. From each of the geo-located tweets we have extracted several features such as the day of week, the hour of the day and day of the trip.

5.2 Data Preparation

As described in section 2.3, for this study Twitter data is prepared for two different time window. For the analyses described in this chapter we used the data set collected from March 29, 2017 to October 10, 2017. We then filtered the data based on BOT score and number of geo-tagged tweets in the same way as the first set of data used in chapter 3 and chapter 4. Then we extracted the user posted locations for the filtered users and separated 2438 tourists (posted location in Twitter profile is describe places outside of Florida). Using Twitter REST API we then collected the latest

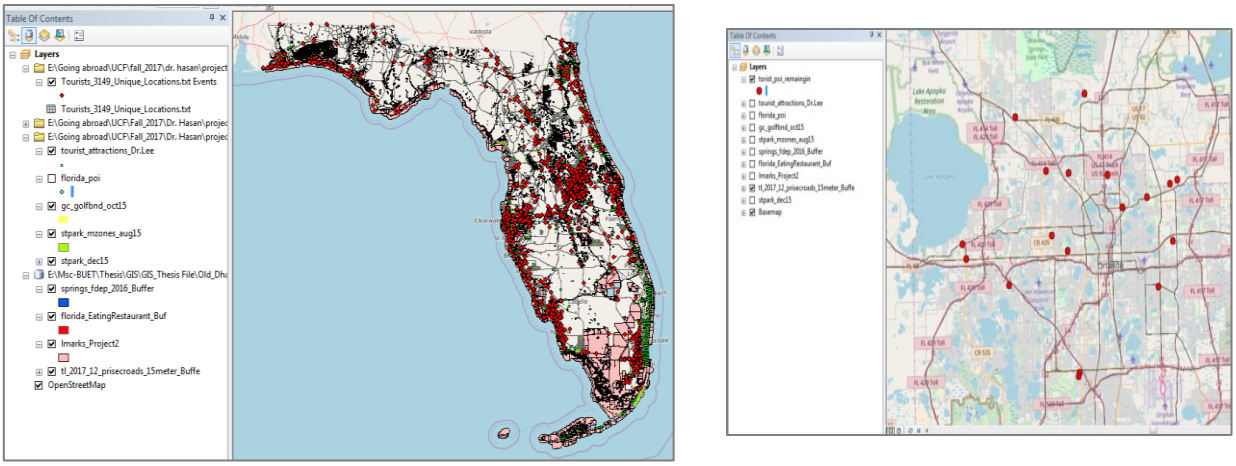
3200 tweets of 2438 user accounts. From March 29 to October 10, these users have posted 35,680 geo-tagged tweets. From them we selected the tweets posted with at least 1 hour time difference. As we are considering the tweets as sequence of activities, keeping same location several times within small time frame for the same user might affect the model performance. These reduced the sample size from 35,680 to 26,187 geo-tagged posts.

Using ArcGIS we found the POIs (point of interests) of the locations given by the geo-tagged posts. For the ease of analyses we divided the POIs into eight different classes (Table 5.1).

TABLE 5. 1: Location Types visited by Tourists.

Location type	Description	Number of geo-tagged posts	Percentage of total geo-tagged posts
1	Airport, Amtrak, bus Stations (Entry/Exit)	828	3.2
2	Beach and Bay areas, beach side restaurant	2838	10.8
3	Theme Parks, Sea World etc.	12546	47.9
4	Restaurants, Fast Food	2129	8.1
5	Other Entertainments (Stadium, Arena, Amway Center, Convention Center, Lake, Shopping Mall, cemetery, university, hospitals, ZOO)	3742	14.3
6	Hotel, Motel, Small resorts (Residential Areas)	891	3.4
7	National/State Park, Reserved Forests, Golf courses	874	3.4
8	On the road, Gas Stations, Garage	2339	8.9

We have spatially joined destinations with Florida geographic shapefiles including polygon shapefiles of golf courses, national/state parks boundary, wildlife reserves, lakes etc. We have used buffer and intersection for available point and line shapefiles which includes the hotels, civic centers, tourists attraction points, springs, highways, trails and scenic byways etc.



(a) **(b)**

FIGURE 5. 1: Spatial Join of Tourists Location Coordinates: (a) with available geographic POI files and (b) points labeled manually.

All the files were gathered from different sources including the tigerline shape files (United States Census Bureau), Florida geographic files database("Florida Geographic File Database,") etc. (Figure 5.1(a)).

The points those did not fall within any of the joins were classified manually by using the latest street map in ArcGIS basefile.

5.3 Model Selection

Semantic labels of locations can be predicted using a hidden-Markov model (HMM) which represents the joint probability distribution *p(y, x)*, where y represents the semantic labels of the locations that are to be predicted, and *x* represents the observed features extracted from the geo-located tweets. Different studies used different types of Markov models for location prediction and/or inference. Alvarez et al. (Alvarez-Lozano et al., 2013) used HMM to predict the next point of interests or POI from mobile phone data. A hybrid model based on HMM was proposed in (Mathew et al., 2012), where the HMM is trained using the clusters made earlier based on the

users visited locations. Hierarchical HMM was applied in (Liao et al., 2007b) to identify users transportation routines.

However, for a large feature set, modeling the joint distribution becomes difficult as one has to account for the complex dependencies among the features in general HMM. Also, as described in Lafferty et al. (Lafferty et al., 2001) probabilistic models such as maximum entropy Markov model (MEMM), HMM, etc. have the problem of label bias. These models will prefer the output label that has been more common in the training data set and thereby will affect the predictive capability of the model. Sequence modeling problems can be framed into a conditional random field (CRF) (Lafferty et al., 2001) model which directly models the conditional distribution p(y|x) instead of modeling the joint distribution *p(y,x)*. Unlike discrete classifiers a CRF can take context into account; e.g., the linear chain CRF predicts sequences of labels for sequences of input samples. In case of sequential data different types of CRF have also been applied in different transportation problems. Liao et al. (Liao et al., 2007a) applied hierarchical CRF to extract location and activity types from users GPS data. In (Liao et al., 2006) the authors applied Relational Markov Networks which is an extension of CRF to label individual's activities performed in significant places using their GPS data.

Given the longitudinal data, the extractable features relating to the location choices and the successful application of CRF models in the literature we decided to apply linear chain CRF in our study.

5.4 Model Formulation

In our problem formulation we have a list of observation, each containing a list of features (i.e. trip day, day of week, hour of day etc.) and the label for the visited location type (1,2,…8). Our

inputs in the model are the feature set, arranged in a list and the output will be the location types. Following figure shows a graphical presentation of the CRF model structure for this study.

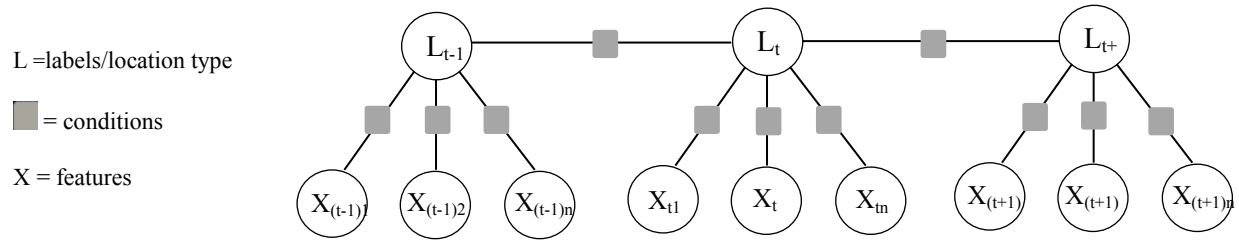


FIGURE 5. 2: Graphical model representation of linear chain CRF.

As linear chain CRF are closely related to HMM we discuss the model formulation by comparing it with HMM structure. HMM makes two independence assumptions while modeling the joint distribution *p(y,x)* (Sutton and McCallum, 2011). First, it assumes that each current state (y_t) depends only on its immediate predecessor (y_{t-1}). Second, it also assumes that each observation variable x_t depends only on the current state y_t. Following the discussion in (Sutton and McCallum, 2011), we can specify an HMM using three probability distributions: the distribution $p(y_1)$ over initial states, the transition distribution $p(y_t|y_{t-1})$ and the observation distribution $p(x_t|y_t)$. Thereby, the joint probability of a state sequence y and an observation sequence x factorizes as:

$$p(y,x) = \prod_{t=1}^{T} p(y_t \mid y_{t-1}) p(x_t \mid y_t) \tag{5.1}$$

In order to describe linear chain CRF first equation 1 is re-written in the following form:

$$p(y,x) = \frac{1}{Z} \prod_{t=1}^{T} \exp\{ \sum_{i,j \in S} \theta_{ij} 1_{\{y_t=i\}} 1_{\{y_{t-1}=j\}} + \sum_{i \in S} \sum_{o \in O} \mu_{oi} 1_{\{y_t=i\}} 1_{\{x_t=0\}} \tag{5.2}$$

Where, $\theta = \{\theta_{ij}, \mu_{oi}\}$ represent the real valued parameters of the distribution and Z represents normalization constant selected in a way so the sum of distribution becomes 1. Equation 2 can be presented using feature functions. Here, each feature function has the form $f_k(y_t, y_{t-1}, x_t)$. Each feature f_k describes the sequence *x* at position *t* with label y_t observed along a transition from label states y_{t-1} to y_t in the finite state machine. The feature function f_k ranges over both all of the f_{ij} and all of the f_{io}. The probability distribution can be written as:

$$p(y,x) = \frac{1}{Z}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1,}x_t)\} \tag{5.3}$$

Then the conditional distribution will be:

$$p(y \mid x) = \frac{p(y,x)}{\sum_{y'} p(y',x)} = \frac{\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, x_t)\}}{\sum_{y'}\prod_{t=1}^{T}\exp\{\sum_{i,j\in S}\theta_k f_k(y_t, y_{t-1}, x_t)\}} \tag{5.4}$$

In general the linear chain CRF describes the conditional probability for a state sequence $y = y_1, y_2, \ldots, y_T$ given an input sequence of feature $x = x_1, x_2, \ldots, x_T$ to be:

$$p(y \mid x) = \frac{1}{Z(x)}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1,}x_t)\} \tag{5.5}$$

Where, *Y, X* are random vectors, Θ is parameter vector and $\{f_k(y_t, y_{t-1}, x_t)\}_{k=1}^{K}$ are set of real valued feature functions. And, *Z(x)* is defined as:

$$Z(x) = \sum_{y}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1,}x_t)\} \tag{5.6}$$

To estimate the parameter Θ of CRF the training data set containing state sequence $y = y_1, y_2,$

....., y_T given an input sequence of feature $x = x_1, x_2, ..., x_T$ is given. In this study we assume all the tourists behave is a similar fashion, thereby creating a single sequence for the whole observation. Assuming an arbitrary prior $p(y; \Theta`)$, the joint likelihood of $p(y,x)$ can be written as:

$$p(y,x) = p(y \mid x;\theta)p(y;\theta') \tag{5.7}$$

The logarithm on both sides of equation 5.7 provides:

$$\log p(y,x) = \log p(y \mid x;\theta) + \log p(y;\theta') \tag{5.8}$$

As the choice of Θ` does not affect optimization over Θ, we can write:

$$LL(\theta) = \sum_{i=1}^{N} \log p(y^{(i)} \mid x^{(i)};\theta) \tag{5.9}$$

The term (i) denotes the sequence for individual users. As we consider all users behaving in same manner, we do not have use of this notation. Substituting the CRF model in equation 5.3 we have the following log-likelihood equation:

$$LL(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, x_t^{(i)}) - \sum_{i=1}^{N}\log Z(x^{(i)}) \tag{5.10}$$

Optimization of $LL(\Theta)$ yields the model parameters. In this study we have used LBFGS or limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm for the optimization. More details of the CRF model formulation and parameter estimations can be found in Lafferty et al., (2001) and Sutton and McCallum, (2011).

5.5 Results

We have developed CRF model to predict the next destination types using tweet posted time (in hour of day), tweet posted day (in day of week), the type of current location visited by the tourists, individual tourist's trip day (i.e. 1st day or 2nd day or nth day of his/her visit in Florida) as

features. The best results were found by using the first three features. We used 70% of the data set for training and 30% for testing.

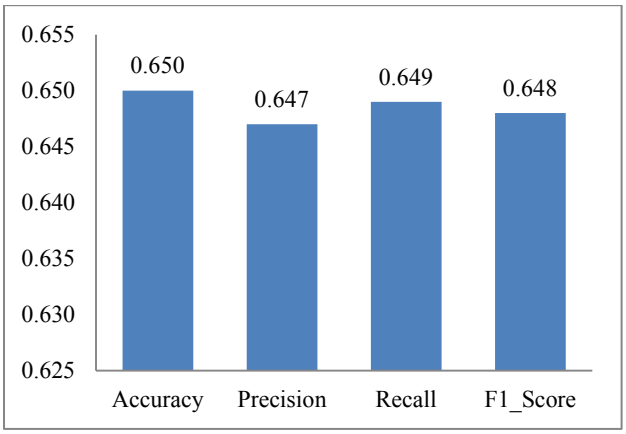


FIGURE 5.3: Performance of linear chain CRF in location type prediction.

The results show accuracy, precision, recall and f-score of 65%, 64.7%, 64.9% and 64.8% respectively while predicting the next location type. In the following table the prediction performances for each type of destination is reported.

TABLE 5.2: Performance with CRF model in Predicting Destination Type.

Location Type	Description	precision	recall	f1-score	support
1	Airport, Amtrak, bus Stations (Entry/Exit)	0.259	0.178	0.211	230
2	Beach and Bay areas, beach side restaurant	0.62	0.603	0.611	839
3	Theme Parks, Sea World etc.	0.818	0.824	0.821	3787
4	Restaurants, Fast Food	0.329	0.338	0.333	622
5	Other Entertainments (Stadium, Arena, Amway Center, Convention Center, Lake, Shopping Mall, cemetery, university, hospitals, ZOO)	0.527	0.546	0.536	1120
6	Hotel, Motel, Small resorts (Residential Areas)	0.466	0.463	0.465	285
7	National/State Park, Reserved Forests, Golf courses	0.562	0.526	0.543	285
8	On the road, Gas Stations, Garage	0.459	0.48	0.47	689
Average		0.647	0.649	0.648	7857

From Table 5.2 we see that CRF has better performance in predicting location type 2 and 3, i.e. theme parks, beach, beach side attractions etc. It has moderate precisions in predicting location type 5 and 7, i.e. the other entertainment centers (Stadium, Arena, Amway Center, Convention Center, Lake, Shopping Mall, cemetery, university, hospitals, Zoo) and state parks, golf courses etc. Theme parks, beaches and national/state reserves and parks are the main attractions of Florida. These locations have the higher percentages in the geo-tagged tweets. We can relate these predicted travel information with the traffic data in spatial and temporal frames to find out the probable traffic impacts around the facilities.

5.6 Summary

Studies (Liao et al., 2007a, 2006) have found 83.3% to 90% accuracy while using CRF and extensions CRF models to predict place and activity types. Using HMM to similar types of problem some studies has found as low as 14% (Mathew et al., 2012) to as high as 69% (Alvarez-Lozano et al., 2013). But, these studies used high resolution GPS data which is difficult to collect for tourists. Therefore, with limited features, our model has shown reasonable performance with average accuracy of 65%. As CRF can use numerous features, future works can include other attributes of the travel and the traveller to enhance model performance. Users' age and gender can be useful features, which are difficult to extract from profile information, especially when the sample size is large. In proposed CRF model we considered all the tourists behave in the same manner as we did not include in traveller attributes.

CHAPTER SIX: DESTINATION CHOICE MODEL FOR RESIDENTS

6.1 Introduction

Destination Choice is an important input in transportation demand modeling. It is vital to know which groups of people are travelling to where and for what purposes. Trip attributes such as distance traveled, transportation mode chosen; individual attributes such as age, gender, income etc.; and origin and destination attributes such as land use types, number of attraction points (offices, schools, civic centers) etc. are the input parameters for a long term planning of any urban area. Updated travel data are necessary to develop more informed models describing recent travel behavior of a population. Up until now all the major planning agencies rely on time consuming and/or costly traditional data collection methods such as household survey, telephone survey etc. This study proposes an extensive data merging technique to overcome the limitations of traditional surveys in collecting latest travel data and inferring the travel behavior through appropriate model frameworks.

Usually destination choice modeling is characterized by a large set of alternatives (Hendrik and Perdana, 2014). However, during many transport related problems (such as model development) data acquisition for each alternative is not a feasible proposition. Such data collection is arduous as found in some researches. Simma et al. (Simma et al., 2002) explored such variables in detail for long distance leisure travel in Switzerland, reporting that the data collection work was indeed particularly arduous. This paper presents an alternative approach, using aggregated data from the location based social network (LBSN) Foursquare to represent destination attractiveness in the utility function of a multinomial logit model.

Big data, such as those collected by Foursquare or Twitter, are a described as the “topic du

jour" in transport modeling in (Molloy and Moeckel, 2017). Rashidi et al.(Rashidi et. al, 2017) presented the first comprehensive literature review exploring the opportunities and challenges inherit to working with such data, with a special focus on travel demand modeling. They examined the recent applications of social media data to both aggregate and disaggregate models, activity behavior, traffic behavior, incidents and natural disasters. Twitter data has been used in (Lin et al., 2013.3) to model the impact of extreme weather on freeway speed for the Buffalo-Niagara, New York, metropolitan area. The study merged three sets of data namely Twitter data, weather station data and traffic data to develop two linear regression models, one with and another without the Twitter data. From their R-square values they found improved model performance by incorporating the Twitter variables.

To the author's best knowledge, only one research has been done using both social media and existing surveys by Molloy and Moeckel (Molloy and Moeckel, 2017). They utilized foursquare check-in data and Transport Survey of Residents of Canada (TSRC) data to model long distance destination choice model for Ontario, Canada.

6.2 Data Preparation

We utilized python's geohash ("Geohash 1.0," 2015) library to locate the users home census tract. Geohash divides the geographical area into pre-defined rectangular boundaries (in our case we selected 152 meter by 152 meter geohash). We have counted the number of coordinates that fall within each geohash and reported the geohash with the largest number of coordinates as the user's home location. Again, we set a minimum threshold of 3 geo-tagged posts within a geohash to consider the location as the user's home. In this method we found home locations of nearly 400 users, but we were only able to manually extract the demographic information (age group and

gender) of 345 users. Therefore, we have worked the subsequent analyses for the destinations of these 345 users. Using ArcGIS we have spatially joined destinations with Florida census tracts shapefile. We have merged different data sources containing the number of offices, schools, entertainment centers, hospitals etc. in Florida and spatially joined them with the census tract shapefile. The files were gathered from different sources including the tigerline shape files (United States Census Bureau), Florida geographic files database("Florida Geographic File Database") etc. The destinations are divided into three major types, i.e. recreational, shopping and others. Based on the destination types the trips are denoted as recreational trips, shopping trips and others. The data set contained 345 users with home in 199 different census tracts and 44,085 destinations in 1651 different census tracts.

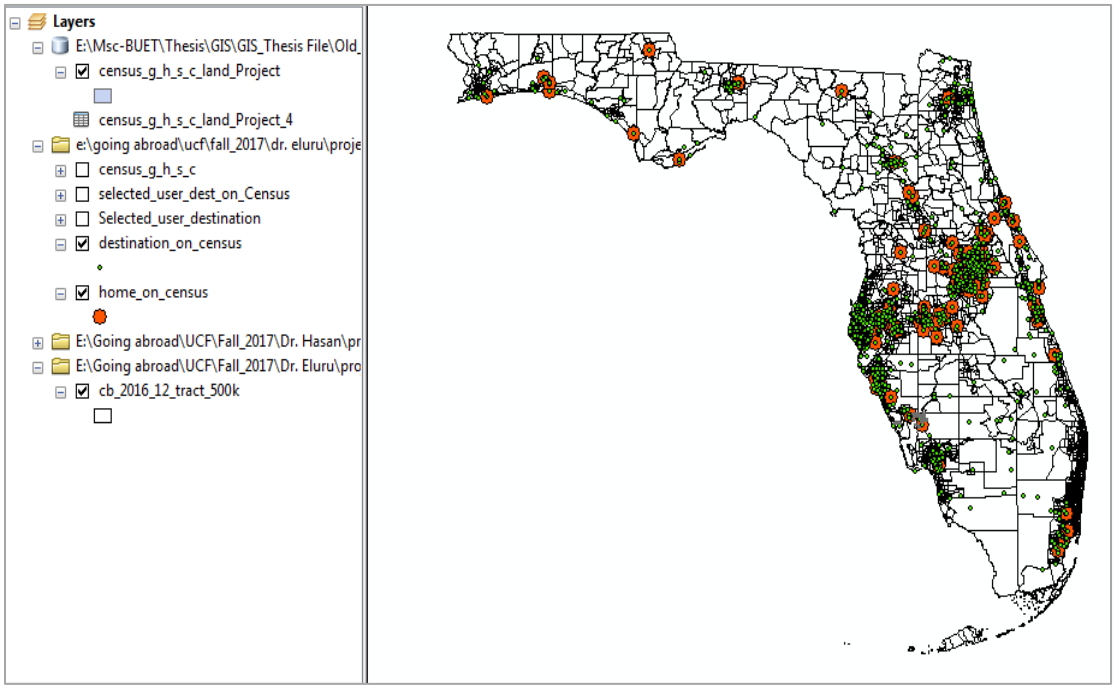


FIGURE 6.1: Merging user home and destination with census tracts.

We kept the destinations those make sense based on timeline analyses. We excluded any destination if a user has posted several times from the same location within very short period of

time. We have randomly drawn 29 alternative census tracts as alternative destination against each trip. The sample data set is given in Table 6.1.

TABLE 6. 1: Sample data for destination choice model.

Case id	Person ID	Trip ID	Gender (male=1)	Age group	Distance (km)	Choice	Trip Purpose*
1307581	5	21006	0	3	7.964441	1	3
1307582	5	21006	0	3	121.1204	0	3
1307583	5	21006	0	3	54.14300	0	3
1307584	5	21006	0	3	332.8362	0	3
1307585	5	21006	0	3	172.9131	0	3

**Trip purpose: 1= recreation, 2=shopping, 3 = others.*

The variables we have extracted include:

- User age (divided into 5 Age groups: up to 15, 16-25, 26-40, 41-55, 56 and above), and user gender from Twitter profile pictures.
- Per-capita income (individual mean, 3.3 year estimate) in 1000 USD.
- Number of civic center, schools, hospitals, government building in point shape files.
- Land use types using the area of residential, industrial, institutional, recreational, office, and landuse mix of the destination and home census tracts.
- Distance from the center of the home census tract to the center of the destination census tract in kilometers.

Table 6.2 lists the variables and their description used in the models.

TABLE 6.2: Description of Variables used in Choice Model

Variable	Description	Variable	Description
HINDUSTR	Industrial area in home	DINDUSTR	Industrial area in destination
HRECREAT	Recreational area in home	DRECREAT	Recreational area in destination
HOFFICE	Office area in home	DOFFICE	Office area in destination
HAGRICUL	Agricultural area in home	DAGRICUL	Agricultural area in destination
HRESIDEN	Residential area in home	DRESIDEN	Residential area in destination
HLANDMIX	Landuse mix in home	DLANDMIX	Landuse mix in destination
HHOSPITA	Number of hospitals in home	DHOSPITA	Number of hospitals in destination
HSCHOOL	Number of schools in home	DSCHOOL	Number of schools in destination
HCIVICCE	Number of civic centers in home	DCIVICCE	Number of civic centers in destination
HINCOME	Per-capita income in home	DINCOME	Per-capita income in destination
HGOVMNTB	Number of government buildings in home	DGOVMNTB	Number of government buildings in destination
DISTKM	Distance in kilometer	Weekend	Dummy variable for Weekend
PShop	Dummy variable for shopping trips	PRec	Dummy variable for recreational trips
Pother	Dummy variable for other trips	Female	Dummy variable for gender (female=1)

Income in home census tract, age, gender etc. are the invariant alternatives (does not change with individual, no matter whatever destination he/she chooses). The dependent variables are two categories: '1' for the selected destination and '0' for the 29 alternatives chosen randomly for each trip. With '0' value all the alternatives are the base categories and significant parameter estimates of the variables signifies the effect of that particular variable on choosing the destination. We have explored various interactive variables which are described in the result section of this chapter.

6.3 Model Selection and Formulation

"Discrete choice models can be used to analyze and predict a decision maker's choice of one alternative from a finite set of mutually exclusive and collectively exhaustive alternatives" (Koppelman and Bhat, 2006). In this study the goal is to capture the destination's characteristics as well as individual's characteristics those affect the choice. Therefore, a number of alternatives have been included in each trip to draw those effects, assuming one individual can select any of the alternatives. This provides a dependent variable with nominal outcome, and therefore, multinomial logistic regression is a better choice for this problem. Again in this study we tried to segment the population into different groups based on the observed variables such as income, land use of the home census tracts, trip purpose etc. Instead of doing it exclusively based on some predefined criteria (male or female, or age group) we used latent or endogenous segmentation (Bhat, 1997) approach which allocated population among different segments in a probabilistic fashion. This helps to better understand the heterogeneity captured in modeling as it allows the influences of exogenous variables to vary across the different segments (Sobhani et al., 2013). Also, in our case we have repeated choice situations for individual which allows us to look into the heterogeneity across the individual as well as panel data. Therefore, we proposed a Panel Latent Segmentation Multinomial Logit (PLSMNL) model. A brief description of PLSMNL model employed in our study is provided below.

Let us consider *S* homogenous segments of trips (the optimal number *S* is to be determined) The utility for assigning a trip *j* (1, 2, … *J*) made by individual *i* (1,2,…, *I*) to segment s is defined as:

$$U_{ijs}^{*} = \beta_s' z_{ij} + \xi_{ijs} \tag{6.1}$$

z_{ij} is a (M x 1) column vector of attributes that influences the propensity of belonging to segment s, β_s' is a corresponding (M x 1) column vector of coefficients and ξ_{ijs} is an idiosyncratic random error term assumed to be identically and independently Gumbel-distributed across trips j and segment s. Then the probability that trip j made by individual i belongs to segment s is given as:

$$P_{ijs} = \frac{\exp(\beta_s' z_{ij})}{\sum_s \exp(\beta_s' z_{ij})} \tag{6.2}$$

Now let us assume k (1,2, … K, in our study K=30) to be an index to represent the destination zone. When a trip is probabilistically assigned to a segment s and zone k is chosen as the destination, the random utility formulation takes the following form:

$$U_{ijk} | \, s = \alpha_s' x_{ij} + \varepsilon_{ijk} \tag{6.3}$$

x_{ij} is a (L x 1) column vector of attributes that influences the utility of destination choice model. α_s' is a corresponding *(L x 1)*-column vector of coefficients and ε_{ijk} is an idiosyncratic random error term assumed to be identically and independently Gumbel distributed across the dataset. Then the probability that trip j chooses zone k as destination within the segment s for individual i is given as:

$$P_{ij}(k) \mid s = \frac{\exp(\alpha_s' x_{ij})}{\sum_k \exp(\alpha_s' x_{ij})} \tag{6.4}$$

Within the latent segmentation framework, the overall probability of trip j by individual i to be destined to zone k is given as:

$$P_{ij}(k) = \sum_{s=1}^{S} (P_{ij}(k) \mid s)(P_{ijs}) \tag{6.5}$$

Therefore, the log-likelihood function for the entire dataset is:

$$LL = \sum_{i=1}^{I} \sum_{j=1}^{J} \log(P_{ij}(k_{ij}^{*})) \tag{6.6}$$

where k_q^* represents the chosen zone for trip j by individual i. By maximizing this log-likelihood function, the model parameters β and α are estimated. GAUSS matrix programming language is used to code the maximum likelihood model estimation.

6.4 Model Results and Interpretation

For PLSMNL we have used 34,000 unique trips of 345 users selected randomly out of 44,085 trips. The first step of PLSMNL is to probabilistically assign each individual into given number of segments based on the exogenous variables. Starting from two segments we have included additional one segment at a time and measured the data fit. Finally, we selected the number of segments in a fashion that adding another segment does not significantly improve the data fit and does not enhance the intuitive interpretations of the variables. We have utilized Bayesian Information Criterion (BIC) to statistically measure the fit as it applies higher penalty on over-fitting and is the most common information criteria used to identify the suitable number of classes for latent segmentation based analysis (Nylund et al., 2007). We have estimated the model with 2, 3 and 4 segments and found the best intuitive results with 3 segments.

The segmentation results are shown in Table 6.3 with the significant variables (at 90% confidence interval) that decide the segment membership.

TABLE 6.3: Segmentation Characteristics of PLSMNL

	Segment 1		Segment 2		Segment 3	
Segment Share	0.2029		0.5359		0.2612	
Variable	**Estimates**	**t-stats**	**Estimates**	**t-stats**	**Estimates**	**t-stats**
Constant	-1.0005	-2.038	0.9274	2.752		
WEEKEND	0.736	3.046	-0.573	-1.933	_	_
FEMALE	-1.0239	-1.917	-1.1573	-2.43	_	_
HAGRICUL	0.5064	2.527	_	_	_	_
HRESIDEN	-2.2669	-2.996	_	_	_	_
HOFFICE	0.219	4.069	_	_	_	_
PShop	_	_	5.1135	20.241	_	_

The estimates of the segment variables reported in Table 6.3 provide the information regarding the segment characteristics. Specifically, destination choices made over the weekend are most likely to be allocated to segment 1 while they are least likely to be allocated to segment 2. In terms of individual gender variable, destination choices of female users are likely to be assigned to segment 3. The segment membership variables are also affected by land use variables. The individuals residing in census tracts with higher agricultural and office area are more likely to be assigned to segment 1 while individuals residing in census tracts with lower residential density are least likely to be allocated to segment 1. Trip purpose variables also influence segment membership. The trips for shopping are most likely to be allocated to segment 2.

In addition to identifying the various factors affecting segment membership, the PLSMNL model allows us to compute the shares of the various segments. In our analysis, the segment shares are as follows: segment 1 – 20.3%, segment 2 – 53.6% and segment 3 – 26.1%. The PLSMNL model can also be employed to generate segment level means for the independent variables (see Table 6.4).

In addition to identifying the various factors affecting segment membership, the PLSMNL model allows us to compute the shares of the various segments. In our analysis, the segment

shares are as follows: segment 1 – 20.3%, segment 2 – 53.6% and segment 3 – 26.1%. The PLSMNL model can also be employed to generate segment level means for the independent variables (see Table 6.4).

TABLE 6.4: Segment shares in PLSMNL

	Segment 1	**Segment 2**	**Segment 3**	**Variable Mena in Overall Sample**
Variables	**Mean of Independent Variables**			
PShop	0.00303	**0.32223**	0.00326	0.17415
PRec	**0.63685**	0.36170	0.55391	0.46774
POther	0.36011	0.31608	**0.44283**	0.35812
WEEKEND	**0.49882**	0.25468	0.34367	0.32747
FEMALE	0.42563	0.26684	0.50669	0.36171
Home Agricultural area	0.09390	-0.01161	0.00780	0.01487
Home Office area	11.65662	1.63968	2.193297	3.81717
Home Residential area	0.45286	0.17878	0.11802	0.21854
Distance in Km	45.83628	**24.39721**	35.74828	31.71260

An examination of the trip purpose variable means indicates that each segment is dominated by one activity purpose: (1) Segment 1 is likely to be recreational destinations, (2) Segment 2 is mostly shopping activity oriented destination and (3) Segment 3 is predominantly other activities. The reader would note that the segment membership allocation is probabilistic (not exclusive) and hence other activity purposes might exist within these segments. Overall, based on segment membership characteristics from Table 4, it is possible to label the various segments in the model. Segment 1 is predominantly a male weekend recreational activity segment. Segment 2 is geared toward shopping destinations on weekdays. Finally Segment 3 mainly represents female other activity destination trips.

All the individuals assigned to particular segment are assumed to have identical preferences while choosing the destination or in other word should have the same utility function (Bhat,

1997). The segment specific multinomial logit models (MNL) are there to capture effects of the variables on destination choice for individuals in each segment (Table 6.5).

TABLE 6.5: Destination Characteristics from Segments specific MNL.

Variable	Segment 1		Segment 2		Segment 3	
	Estimates	***t-stats***	***Estimates***	***t-stats***	***Estimates***	***t-stats***
DISTKM	-0.0064	-4.327	-0.2161	-8.629	-0.0602	-7.06
DINDUSTR	-0.3572	-2.398	0.3424	2.707	-0.2095	-2.372
DRECREAT	0.06	3.439	_	_	_	_
DOFFICE	_	_	0.1249	7.629	0.4253	4.824
DAGRICUL	_	_	_	_	0.5686	5.126
DLANDMIX	0.3623	4.37	0.2218	2.15	_	_
DSCHOOL	0.1168	2.167	0.2825	3.832	_	_
DCIVICCE	0.4525	15.562	_	_	0.4666	5.319
DINCOME	0.2031	2.605	0.287	2.836		
DGOVMNTB	_	_	_	_	0.3659	3.698

In the segment specific model estimation, we employed several destination characteristics. A cursory examination of the results clearly highlights how the variables (and parameter sign/magnitude) influencing the destination choice models across the various segments are quite different. The result provides strong support to our study hypothesis for the presence of population heterogeneity.

In all models, travel distance has a negative coefficient. While a direct comparison of the travel distance across segments needs to be judiciously conducted, a preliminary examination highlights intuitive trends. A low magnitude for the impact of destination is observed for weekend recreational destinations, indicating the higher spatial flexibility over weekends for such trips. A high negative magnitude is observed for the weekday shopping segment highlighting inherent preference for shorter distance trips on weekdays.

In segment 1 destination tract recreational area, land use mix, number of schools, number of civic centers and per-capita income are found to have significant positive impact on the destination alternative. On the other hand, the increased presence of industrial area is likely to reduce the preference for the destination.

In segment 2 industrial area, office area, land use mix, number of schools and income of the destination census are found to have significant positive impact on the individual choice of destination. The results are intuitive considering segment 2 is predominantly weekday shopping destinations. The positive impact of number of schools and office areas variables can be related to the fact that people on weekdays do not leave home only for shopping, rather they prefer shopping on their way to office or in some cases near schools.

For segment 3 we find the variables for office area, agricultural area, number of civic centers and government buildings in the destination census are found to have significant positive impacts (Table 6.5).

It must be noted that our panel structure was unbalanced, meaning that the number of repeated observations for individuals (trips made by individuals) varies across the dataset (from 1 trip to 1823 trips with the mean of 98.6 and median of 31 trips). Please note that while we correct for the panel effect in the standard error estimation we did not explicitly consider unobserved heterogeneity due to the repetitions. However, the PLSMNL performed better comparing to simple trip-specific MNL. We developed four different MNL models: one model for all trips and three models by activity purpose for recreational, shopping and other trips. The log-likelihood values for these models were found to be -48688.757, -20595.791, -2078.21 and -20969.19 respectively. The overall log-likelihood for all observations for trip purpose models was -43,643.19 (-20595.791, -2078.21 and -20969.19). The log-likelihood for the PLSMNL model was

-34,752.8 significantly lower than the overall MNL model or the trip purpose based model suite. Therefore, it is clear that the PLSMNL model provides superior fit.

6.5 Summary

In this chapter we have demonstrated a way of creating joint database by combining social media data with traditional census tract based socio-economic, landuse and infrastructural data for using in the context of transportation demand studies. We propose a panel based endogenous segmentation MNL model to analyze the destination choice preferences of the residents. With three segments we have found out the segment specific MNL models to explain the characteristics of the residents under each segment. Proposed PLSMNL outperformed the trip specific MNL models both quantitatively in terms of goodness of fit and qualitatively by providing better interpretation of the results.

CHAPTER SEVEN: CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

In this study, we presented methods to extract and analyze large-scale data for tourists' and residents' travel related information from Twitter. Filtering steps are followed to remove social bots from the dataset and prepare a reliable sample for analysis. From the filtered Twitter data, we identified tourists and residents using a simple heuristic classification approach. The proposed algorithm outperformed some of the widely used supervised classification methods. When compared to some of the state of art ensemble classification techniques, AdaBoost classifier performed better than the proposed heuristic. All the features used to train these advanced classification techniques are drawn from geographical coordinates (from geo-tagged posts) without making a time intensive content-based analysis.

To find spatial patterns of destination choices made by tourists and residents, we applied three common clustering techniques, i.e. K-Menas, Mean-Shift and DBSCN. From the tourists clustering results, locations are found to be clustered around some of the famous tourist spots, reserved forests and wetlands, airports and beaches in Florida. The number of unique users and total number of coordinates within each cluster indicate tourist attractions in Florida. On the other hand, resident locations are found to be clustered around the residential apartment complexes with some schools, shopping centers and small golf courses within the 3 kilometer radius. Based on some of the widely used validation measures, the performances of the clustering methods are measured. From these indices, K-Means clustering method performed best among the three clustering methods.

To predict the next destination types of the tourists, we have applied CRF model with the

extracted temporal and spatial features of the geo-tagged tweets. The model had a good performance with overall accuracy, precision, recall and f1-score of 65%, 64.7%, 64.9% and 64.8% respectively.

To understand the destination choice behavior of the residents, we proposed a PLSMNL model. The model had best fit with three segments and outperformed the trip specific MNL models. The qualitative assessments of the models indicated that the proposed PLSMNL successfully represented different types of trips (shopping, recreational and others) into different segments. The data integration part of these models will be of great interest for future works using social media data for transportation modeling.

Our analysis of tourists' and residents' destination patterns has significant implications. *First*, it shows how to collect and prepare reliable data on tourist travel behavior from social media. Extraction and analysis of most recent data are required for the planning of large states, especially tourism dependent states such as Florida. Where traditional surveys are highly expensive and difficult to conduct; social media can become a useful cost-effective source providing the most recent travel data of growing region. *Second*, our analysis shows how to infer different patterns from tourist destination choices. Combining the spatial clusters in temporal windows, it is possible to find out traffic impacts of tourists in a study area. *Third*, with extensive data merging techniques, this study presents a framework to understand individual level travel behavior for tourists and residents. Thus, this study showed a promising direction towards using social media data for understanding tourist travel behavior and the methods and findings from this study will be significantly useful in future studies on tourist travel behavior.

7.2 Limitations of the Study

There are some limitations of this study. We have thoroughly noted down the limitations as it will help the researchers to advance this research forward and to make more useful contributions transportation planning studies.

The data set used in this study was collected by setting a boundary for the central Florida region only. A more detailed data set with more tourist and resident users can be extracted by setting the boundary for the whole Florida state.

In the process of BOT filtering, it is possible that we might have excluded many individual user profiles along with the social BOTs as the filtering process is not 100% accurate. For instance, there are some individual users with BOT score greater than 0.4, and also there are some actual BOTs with BOT score less than or equal to 0.4. A more detailed procedure of BOT filtration can be adopted to overcome this limitation.

Self-declared location information, used as a ground truth, can also become erroneous in some cases. Although very few in number, we observed that some users actually reside in different places instead of the locations posted in their profiles. A better filtering procedure will help to introduce more users (residents and tourists) and find more diverse destination patterns.

In resident destination choice model, we considered all the trips as home-based trips, i.e. all the trip's origin is home. This can be avoided if we could have extracted the travel start time from the tweet posted time. But the problem is that people do not tweet at the exact time when they leave one place or reach to a place. For instance, during leaving for office one will not tweet just before starting his trip, and also one may not tweet instantly after reaching office.

7.3 Recommendations and Future Research

The main data source of the study is Twitter and as explained in (Zheng et al., 2015) social data is always evolving with time. Therefore, systematic data fusion approaches are needed to connecting social media data with different geographical and infrastructural database to add more information to models.

Using spatio-temporal clustering, it is possible to find out traffic impacts of tourists in a study area. But in that case researchers must be careful while using the tweet posted time for the clusters as individuals often do not post tweets at the exact starting or ending time of their activity.

For residents' destination choice, there are possibilities to enhance the models by incorporating more types of trips such as school trips, office/work trips etc. In this case, collecting a significant sample size may prove to be difficult as individuals are less likely to post geo-tagged tweets from these locations.

Lastly, to keep the analyses simple we did not use any tedious text mining process. As text is an important part of Twitter data, future studies can include features extracted from tweet texts and include them in the destination prediction models.

APPENDIX: RESIDENT DESTINATION CHOICE MODEL

TABLE A1: MNL for all Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.0296063	0.000178	-166.27
DAGRICUL	0.0428767	0.015523	2.76
DRESIDEN	0.0948184	0.016194	5.86
DOFFICE	0.0669291	0.007265	9.21
DLANDMIX	0.2210507	0.00904	24.45
DGOVMNTB	0.1654784	0.010166	16.28
DHOSPITA	0.0878285	0.005896	14.9
DSCHOOL	0.1403953	0.006081	23.09
DCIVICCE	0.2403403	0.008739	27.5
DINCOME	0.208048	0.012313	16.9
Male_DAGRICUL	0.0631316	0.017523	3.6
Male_DRESIDEN	-0.060991	0.019041	-3.2
Male_DOFFICE	0.0272397	0.009159	2.97
Male_DGOVMNTB	-0.0836544	0.012501	-6.69
Male_DCIVICCE	-0.0573614	0.01061	-5.41
Male_ DINCOME	-0.0755817	0.016081	-4.7

TABLE A2: MNL for Recreational Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.02354	0.000217	-108.34
DLANDMIX	0.319271	0.019527	16.35
DGOVMNTB	0.136443	0.00835	16.34
DCIVICCE	0.328753	0.007436	44.21
DRECREATION	0.070236	0.012205	5.75
DINCOME	0.326134	0.010863	30.02
Male_DOFFICE	0.14297	0.007433	19.23
Male_DCIVICCE	-0.0636	0.011016	-5.77
Male_DLANDMIX	-0.06161	0.024924	-2.47

TABLE A3: MNL for Shopping Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.25664	0.006075	-42.25
DINSTITUTIONAL	1.103844	0.453156	2.44
DRESIDENTIAL	0.200517	0.065837	3.05
DOFFICE	0.242138	0.032857	7.37
DINDUSTRIAL	0.163188	0.08158	2
DGOVMNTB	0.422072	0.06524	6.47
DSCHOOL	0.321163	0.02738	11.73
DINCOME	0.195943	0.036871	5.31
Male_DOFFICE	-0.27463	0.04256	-6.45
Male_DLANDMIX	0.279822	0.043544	6.43
Male_DCIVICCE	0.156366	0.023694	6.6
Male_DGOVMNTB	-0.62596	0.078539	-7.97

TABLE A4: MNL for Other Trips.

Parameters	Estimates	Standard Error	t-stat
DISTKM	-0.02933	0.000274	-106.97
DAGRICULTURAL	0.102649	0.018747	5.48
DRESIDENTIAL	0.068642	0.015537	4.42
DOFFICE	0.029873	0.012748	2.34
DRECREATIONAL	0.085015	0.013012	6.53
DLANDMIX	0.260602	0.013769	18.93
DGOVMNTB	0.266412	0.014414	18.48
DSCHOOL	0.255883	0.015177	16.86
DCIVICCE	0.133883	0.007995	16.75
DINCOME	0.162022	0.021221	7.63
Male_DAGRICULTURE	-0.03851	0.013864	-2.78
Male_DOFFICE	-0.10365	0.017929	-5.78
Male_DGOVMNTB	-0.77994	0.067714	-11.52
Male_DRECREATION	-0.20834	0.018956	-10.99
Male_DSCHOOL	0.07935	0.020457	3.88
Male_DINCOME	-0.11151	0.02659	-4.19

TABLE A5: Segment Shares for PLSMNL.

Variables	Mean of Independent Variables			Variable Mena in Overall Sample
	Segment 1	Segment 2	Segment 3	
AGE15	0.00590	0.00258	0.00712	0.00444
AGE1625	0.09010	0.05563	0.09502	0.07291
AGE2640	0.57565	0.62839	0.55796	0.59929
AGE4155	0.27527	0.24973	0.24511	0.25371
AGE56	0.05233	0.06343	0.09395	0.06915
FEMALE	0.42563	0.26684	0.50669	0.36171
PSHOP	0.00303	0.32223	0.00326	0.17415
PREC	0.63685	0.36170	0.55391	0.46774
POTHER	0.36011	0.31608	0.44283	0.35812
HAGRICULTURAL	0.09390	-0.01161	0.00780	0.01487
HINDUSTRIAL	0.80618	0.19245	0.19874	0.31865
HINSTITUTIONAL	-0.02250	-0.02226	-0.02332	-0.02258
HRECREATION	0.01137	-0.02817	-0.02666	-0.01975
HRESIDENTIAL	0.45286	0.17878	0.11802	0.21854
HOFFICE	11.65662	1.63968	2.19330	3.81717
HBUA	0.19221	0.00053	0.01760	0.04389
HLANDMIX	0.84659	0.33081	0.45246	0.46725
HGOVMNTBUILDING	0.58614	0.51493	0.57211	0.54432
HHOSPITAL	0.04214	0.12884	0.13467	0.11277
HSCHOOL	0.91779	1.00072	0.70210	0.90590
HCIVICCENTER	6.85813	1.78206	1.87492	2.83649
HINCOME	0.01268	0.03979	0.12637	0.05690
DAGRICULTURAL	0.10177	-0.04344	0.00111	-0.00233
DINDUSTRIAL	0.56183	0.20329	0.27183	0.29396
DINSTITUTIONAL	-0.01528	-0.01992	-0.01848	-0.01860
DRECREATION	0.01803	-0.00512	0.01657	0.00524
DRESIDENTIAL	0.42787	0.16326	0.20760	0.22854
DOFFICE	8.77714	3.11301	4.09335	4.51855
DBUA	0.17819	-0.00979	0.04542	0.04278

Variables	Mean of Independent Variables			Variable Mena in Overall Sample
	Segment 1	Segment 2	Segment 3	
DLANDMIX	0.64311	0.45145	0.48407	0.49887
DGOVMNTBUILDING	0.48698	0.32611	0.45881	0.39341
DHOSPITAL	0.07109	0.18085	0.17306	0.15654
DSCHOOL	0.78526	0.75808	0.72269	0.75435
DCIVICCENTER	5.26990	2.23893	2.74115	2.98521
DINCOME	0.04375	-0.02437	0.03959	0.00616
WEEKEND	0.49882	0.25468	0.34367	0.32747
DISTKM	45.83628	24.39721	35.74828	31.71260

REFERENCES

Abbasi, A., Rashidi, T.H., Maghrebi, M., Waller, S.T., 2015. Utilising Location Based Social Media in Travel Survey Methods. Proc. 8th ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN'15 1–9. doi:10.1145/2830657.2830660

Alvarez-Lozano, J., García-Macías, J.A., Chávez, E., 2013. Learning and user adaptation in location forecasting. Proc. 2013 ACM Conf. Pervasive ubiquitous Comput. Adjun. Publ. - UbiComp '13 Adjun. 461–470. doi:10.1145/2494091.2495978

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., Thom, D., 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. Comput. Sci. Eng. 15, 72–82. doi:http://dx.doi.org/10.1109/MCSE.2013.70

Beyer, M.A., Laney, D., 2012. The importance of "big data": a definition. Gartner: Stamford, CT, USA.

Bhat, C.R., 1997. An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel. Transp. Sci. 31, 34–48. doi:10.1287/trsc.31.1.34

Bholowalia, P., Kumar, A., 2014. EBK-Means : A Clustering Technique based on Elbow Method and K-Means in WSN. Int. J. Comput. Appl. 105, 17–24. doi:10.5120/18405-9674

Bolivar, C.T., 2014. City Usage Analysis using Social Media. Delft University of Technology.

Botometer [WWW Document], 2014. URL https://botometer.iuni.iu.edu/#!/ (accessed 12.26.17).

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. doi:10.1007/BF00058655

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. Methods 3, 1–27.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., Soltani, K., 2014. A Scalable

Framework for Spatiotemporal Analysis of Location-based Social Media Data. doi:10.1016/j.compenvurbsys.2015.01.002

Chan, J.C.W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens. Environ. 112, 2999–3011. doi:10.1016/j.rse.2008.02.011

Chang, H.W., Lee, D., Eltaher, M., Lee, J., 2012. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). pp. 111–118. doi:10.1109/ASONAM.2012.29

Chen, Y., Mahmassani, H.S., Frei, A., 2017. Incorporating social media in travel and activity choice models: conceptual framework and exploratory analysis. Int. J. Urban Sci. 0, 1–21. doi:10.1080/12265934.2017.1331749

Cheng, Y., 1995. Mean Shift, Mode Seeking, and Clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17, 790–799. doi:10.1109/34.400568

Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z., 2011. Exploring Millions of Footprints in Location Sharing Services. Icwsm 2010, 81–88. doi:papers3://publication/uuid/0C46BD5D-4908-4A8A-BD06-5BCB2F1DE282

Chi, E.H., 2008. The social web: Research and opportunities. Computer (Long. Beach. Calif). 41, 88–91. doi:10.1109/MC.2008.401

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: User Movement in Location-Based Social Networks. Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. 1082–1090. doi:10.1145/2020408.2020579

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619. doi:10.1109/34.1000236

Cramer, H., Mattias, R., Lars Erik, H., 2011. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare, in: 13th International Conference on Human Computer Interaction with Mobile Devices and Service. ACM, pp. 57–66. doi:10.1145/2037373.2037384

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press. doi:doi:10.1017/CBO9780511801389

Davies, D.L., Bouldin, D.W., 1979. A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1, 224–227. doi:10.1109/TPAMI.1979.4766909

Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F., 2016. BotOrNot: A System to Evaluate Social Bots 4–5. doi:10.1145/2872518.2889302

Dietterich, T.G., 2010. Ensemble methods in machine learning, in: Multiple Classifier Systems. pp. 1–15.

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern. 4, 95–104. doi:10.1080/01969727408546059

Edwards, D., Griffin, T., Hayllar, B., 2008. Urban tourism precincts: An overview of key themes and issues., in: Hayllar, B., Griffin, T., Edwards, D. (Eds.), City Spaces, Tourist Places: Urban Tourism Precincts. Amsterdam: Butterworth-Heinemann., pp. 95–105.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: In KDD. pp. 226–231.

Ester, M., Kriegel, H.-P., Xu, X., 1995. A Database Interface for Clustering in Large Spatial

Databases. Kdd-95.

Eye, O., Beach, C., Kingdom, M., Park, L.E., 2017. Central Florida Visitor Study. Florida, USA.

Florida Geographic File Database [WWW Document], 2008. URL ftp://ftp1.fgdl.org/pub/state/ (accessed 12.15.17).

Florida Transportation Trends and Condition 2012, 2012. doi:10.1016/S0962-8924(12)00187-0

Flyvbjerg, B., Holm, M.S., Buhl, S.L., 2005. How (In) accurate Are Demand Forecasts in Public Works Project? The Case of Transportation. J. Am. Plan. Assoc. 71, 131–146. doi:10.1080/01944360508976688

Freund, Y., Schapire, R.E., 1996. Experiments with a New Boosting Algorithm. Proc. Int. Conf. Mach. Learn. 148–156. doi:10.1.1.133.1040

Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez, E., 2012. Characterizing urban landscapes using geolocated tweets. ASE/IEEE Int. Conf. Soc. Comput. Soc. 2012 239–248. doi:10.1109/SocialCom-PASSAT.2012.19

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014. The potential of social media in delivering transport policy goals. Transp. Policy 32, 115–123. doi:10.1016/j.tranpol.2014.01.007

Geohash 1.0 [WWW Document], 2015. URL https://www.elastic.co/guide/en/elasticsearch/guide/current/geohashes.html (accessed 2.20.18).

Girardin, F., Dal Fiore, F., Blat, J., Ratti, C., 2007. Understanding of tourist dynamics from explicitly disclosed location information. 4th Int. Symp. LBS Telecartography 58.

Gladstone, D.L.. ., Fainstein, S.S.. ., 2001. Tourism in US Global Cities: A comparison of New York and Los Angeles. J. Urban Aff. 23, 23–40. doi:10.1111/0735-2166.00073

Gursoy, D., Jurowski, C., Uysal, M., 2002. Resident attitudes - A structural modelling approach. Ann. Tour. Res. 29, 79–105.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. 44. doi:10.1016/j.trc.2014.04.003

Hasan, S., Ukkusuri, S. V., 2017. Reconstructing Activity Location Sequences From Incomplete Check-In Data: A Semi-Markov Continuous-Time Bayesian Network Model. IEEE Trans. Intell. Transp. Syst. 1–12. doi:10.1109/TITS.2017.2700481

Hasan, S., Ukkusuri, S. V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. PLoS One 10, 1–19. doi:10.1371/journal.pone.0124819

Hasan, S., Ukkusuri, S. V., Zhan, X., 2016. Understanding Social Influence in Activity Location Choice and Lifestyle Patterns Using Geolocation Data from Social Media. Front. ICT 3, 1–9. doi:10.3389/fict.2016.00010

Hendrik, H., Perdana, D.H.F., 2014. Trip Guidance: A Linked Data Based Mobile Tourists Guide. Adv. Sci. Lett. 20, 75–79. doi:https://doi.org/10.1166/asl.2014.5285

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24, 881–892. doi:10.1109/TPAMI.2002.1017616

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53, 59–68. doi:10.1016/j.bushor.2009.09.003

Kennedy, L., Naaman, M., 2008. Generating Diverse and Representative Image Search Results for Landmarks.

Kim, J.H., 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out

and bootstrap. Comput. Stat. Data Anal. 53, 3735–3745. doi:10.1016/j.csda.2009.04.009

Koppelman, F.S., Bhat, C., 2006. A Self Instructing Course in Mode Choice Modeling : Multinomial and Nested Logit Models by with technical support from Table of Contents. Elements 28, 501–12. doi:10.1002/stem.294

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., Shoor, I., 2017. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. Transp. Res. Part C Emerg. Technol. 77, 275–291. doi:10.1016/j.trc.2017.02.003

Lafferty, J., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01 Proc. Eighteenth Int. Conf. Mach. Learn. 8, 282–289. doi:10.1038/nprot.2006.61

Lee, J.H., Davis, A.W., Yoon, S.Y., Goulias, K.G., 2016. Activity Space Estimation with Longitudinal Observations of Social Media Data. Transportation (Amst). 43, 955–977.

Lee, J.H., Mcbride, E., Mcbride, E., Goulias, K.G., 2017. Exploring Social Media Data for Travel Demand Analysis : A comparison of Twitter , household travel survey and synthetic population data in California. 95th Annu. Meet. Transp. Res. Board 500.

Lian, D., Xie, X., 2011. Collaborative activity recognition via check-in history. Proc. 3rd ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN '11 1. doi:10.1145/2063212.2063230

Liao, L., Fox, D., Kautz, H., 2007a. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. Int. J. Rob. Res. 26, 119–134. doi:10.1177/0278364907073775

Liao, L., Fox, D., Kautz, H., 2006. Location-based Activity Recognition. Adv. Neural Inf. Process. Syst. 7870794. doi:10.1007/978-3-540-74565-5_6

Liao, L., Patterson, D.J., Fox, D., Kautz, H., 2007b. Learning and inferring transportation routines. Artif. Intell. 171, 311–331. doi:10.1016/j.artint.2007.01.006

Lin, L., Ni, M., He, Q., Gao, J., Sadek, A.W., 2015. Modeling the Impacts of Inclement Weather on Freeway Traffic Speed. Transp. Res. Rec. J. Transp. Res. Board 2482, 82–89. doi:10.3141/2482-11

Liu, Y., Sui, Z., Kang, C., Gao, Y., 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS One 9. doi:10.1371/journal.pone.0086026

Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, S.T., 2015. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC 2015–Octob, 208–213. doi:10.1109/ITSC.2015.43

Majid, A., Chen, L., Chen, G., Mirza, H.T., Hussain, I., Woodward, J., 2013. A context-aware personalized travel recommendation system based on geotagged social media data mining. Int. J. Geogr. Inf. Sci. 27, 662–684. doi:10.1080/13658816.2012.696649

Manca, M., Boratto, L., Morell Roman, V., Martori i Gallissà, O., Kaltenbrunner, A., 2017. Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. Online Soc. Networks Media 1, 56–69. doi:10.1016/j.osnem.2017.04.002

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval, Computational Linguistics. doi:10.1162/coli.2009.35.2.307

Mathew, W., Raposo, R., Martins, B., 2012. Predicting future locations with hidden Markov models. Proc. 2012 ACM Conf. Ubiquitous Comput. - UbiComp ’12 911. doi:10.1145/2370216.2370421

McLachlan, G., Kim-Anh, D., Ambroise, C., 2005. Analyzing microarray gene expression data.

John Wiley & Sons, Hoboken, New Jersey.

McNeill, G., Bright, J., Hale, S.A., 2016. Estimating Local Commuting Patterns From Geolocated Twitter Data 1–17.

Molloy, J., Moeckel, R., 2017. Improving Destination Choice Modeling Using Location-Based Big Data. ISPRS Int. J. Geo-Information 6, 291. doi:10.3390/ijgi6090291

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. doi:10.1007/978-3-319-05579-4_10

North Florida Travel Survey [WWW Document], 2017. URL https://www.northfloridatravelsurvey.com/northfloridahtsweb/pages/privacy?locale=en-US

Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Struct. Equ. Model. 14, 535–569. doi:10.1080/10705510701575396

Perrin, A., 2005. Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites--a Nearly Tenfold Jump in the Past Decade. Pew Res. Cent.

Rashidi, T.H.; Abbasi, A.; Maghrebi, M.; Hasan, S.; Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. 75, 197–211.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. 75, 197–211. doi:10.1016/j.trc.2016.12.008

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: Promises, progress and prospects. Int. J. Urban Sci. 18, 31–60. doi:10.1080/12265934.2013.835118

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-Validation, in: Encyclopedia of Database System. Springer, p. (pp. 532-538.

Rokach, L., 2010. Pattern classification using ensemble methods. Ser. Mach. Percept. Artif. Intell. 75.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Sadri, A.M., Hasan, S., Ukkusuri, S. V., 2017. Joint Inference of User Community and Interest Patterns in Social Interaction Networks.

Safavian, S.R., Landgrebe, D., 1990. A Survey of Decsion Tree Clasifier Methodology. IEEE Trans. Syst. Man. Cybern. 21, 660–674.

Simma, A., Schlich, R., Axhausen, K.W., 2002. Destination choice modelling of leisure trips: The case of Switzerland. Monit. Manag. Visit. Flows Recreat. Prot. Areas 150–158.

Sobhani, A., Eluru, N., Faghih-Imani, A., 2013. A latent segmentation based multiple discrete continuous extreme value model. Transp. Res. Part B.

Sutton, C., McCallum, A., 2011. An Introduction to Conditional Random Fields. Mach. Learn. 4, 267–373. doi:10.1561/2200000013

The Google Places API Web Service [WWW Document], 2017. URL https://developers.google.com/places/web-service/intro (accessed 12.25.17).

Theobald, W.F. (Ed.), 2005. Global Tourism, Third. ed. Elsevier Butterworth-Heinemann, Maryland Heights, MO.

Twitter by the Numbers: Stats, Demographics & Fun Facts [WWW Document], 2017. . Omnicore. URL https://www.omnicoreagency.com/twitter-statistics/.

Twitter Developer Documentation: REST API [WWW Document], 2006. URL

https://dev.twitter.com/rest/public (accessed 12.30.17).

Twitter Developer Documentation: Streaming API [WWW Document], 2006. URL https://dev.twitter.com/streaming/overview/request-parameters#filter_level (accessed 12.30.17).

United States Census Bureau, n.d. TIGER/Line® Shapefiles and TIGER/Line® Files [WWW Document]. URL https://www.census.gov/geo/maps-data/data/tiger-line.html (accessed 3.15.18).

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Woolley, S.C., 2016. Automating power: Social bot interference in global politics. First Monday 21. doi:http://dx.doi.org/10.5210/fm.v21i4.6161

Wu, L., Zhi, Y., Sui, Z., Liu, Y., 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. PLoS One 9. doi:10.1371/journal.pone.0097010

Yin, Z., Cao, L., Han, J., Luo, J., Huang, T., 2011. Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media. Siam Icdm 980–991. doi:10.1137/1.9781611972818.84

Zhang, Y., Mohammadian, A. (Kouros), 2008. Bayesian Updating of Transferred Household Travel Data. Transp. Res. Rec. J. Transp. Res. Board 2049, 111–118. doi:10.3141/2049-13

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., Yang, L., 2015. Big Data for Social Transportation. IEEE Trans. Intell. Transp. Syst. 17, 620–630. doi:10.1109/TITS.2015.2480157

Zheng, Y.-T., Zha, Z.-J., Chua, T.-S., 2012. Mining Travel Patterns from Geotagged Photos. ACM Trans. Intell. Syst. Technol. 3, 1–18. doi:10.1145/2168752.2168770