IDENTIFYING FUNCTIONAL URBAN REGIONS FROM BLUETOOTH DATA:

A CASE STUDY OF AUSTIN, TEXAS

by

Jacob H. Combs, Bachelor of Science, Anthropology

A directed research project submitted to the Graduate Council of

Texas State University in partial fulfillment

of the requirements for the degree of

M.A.Geo

with a Major in Geographic Information Science

May, 2018

Committee Members:

Dr. Yihong Yuan, Chair

Dr. Nathan Currit

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

This research would not have been possible without the help and guidance of my committee members, Dr. Yihong Yuan and Dr. Nate Currit. I would also like to thank my wife for her infinite patience and buying me time to spend 10 hours at a time at the desk all day and night.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

ABSTRACT

Worldwide, the rate of urbanization has increased over the last several decades and the need to adequately identify how urban areas are used increases importance with every year. This study applies dynamic time warping and hierarchical clustering methods to a Bluetooth data set to identify functional urban regions in Austin, Texas. Examining the distribution of the functional urban regions and their spatial configuration allows inferences to be made in relation to the way that people use the urban area of Austin.

Identifying Functional Urban Regions from Bluetooth Data: A Case Study of Austin

1. Introduction

From road closures to new construction and rapid population growth, the urban landscape is in near constant flux. The United Nations reports in 2014 an urbanization rate of 54% globally with North America and Europe much higher at 82% and 73% respectively and yet there are few techniques that are suitable to understand the functional configuration and capture the dynamic nature of urban land use which are necessary information to build the cities of tomorrow (Sagal, Loidl, and Beinat 2012). This research posits a new methodology that can extract and visualize the rapidly growing and changing urban environment by incorporating dynamic time warping and cluster analysis to identify and define functional urban regions (FUR).

Though similar research has been conducted using big data to identify FUR's (Yuan and Raubal 2012; Gao et al. 2017), this research seeks to use Bluetooth data at surface street intersections to identify clustered patterns instead of measuring speeds or road use patterns. Previous research using Bluetooth has focused on finding patterns in road network usage or average highway speeds at various hours of the day (Bachmann et al. 2013:34). Previous research to define FUR's has been conducted using social media data (Liu et al. 2015; Zhi et al. 2016; Gao et al. 2017) or calls made through cell towers (Ahas et al. 2015; Yuan and Raubal 2012). This research seeks to bring together these two bodies of research, identifying FUR's and Bluetooth data analysis, to develop a dynamic methodology to define FUR's in Austin, Texas. We aim to identify clusters of similar time series in Austin, Texas. The clusters will then be used to identify into various functional regions such as: residential, work, and recreational (including but not limited to – bars, restaurants, movie theatres, etc.). The results will be based on the output of a clustering algorithm that takes the output from the distance time warping algorithm to build clusters of similar time series across the city. Our main goal is to prove the efficacy of this specific workflow and methodology to cluster city regions based on Bluetooth data as opposed to traditional remote sensing techniques. Classifying urban areas based on remotely sensed images is difficult due to the distribution of heterogeneous urban land-use types and the similarity of the

spectral response from different urban land-use types (Herold et al. 2002). While there are difficulties with image-based classification of urban areas, this research is more specifically concerned with developing a dynamic, data-based approach. This approach is more sensitive to the daily patterns of human interactions and movements on the individual level. Furthermore, this research program seeks to build upon existing social sensing paradigm by using Bluetooth traffic counts as the primary data source. We make use of hourly Bluetooth data collected in 2016 by Bluetooth sensors installed at major intersections across Austin. Using the principles of social sensing, and clustering and time series analysis, we intend to identify and describe functional urban regions across the city of Austin. The benefits of a study of this nature lie in the dynamic way that the data can reflect a road closure, or the opening of a new bar and restaurant section of the city. The final result of the data analysis will then be cross referenced using traditional remote sensing accuracy assessment techniques to determine the overall effectiveness of the data analysis results.

2. Background

Big data is normally characterized by the four V's – velocity, variety, volume and veracity (Goodchild 2013). The term velocity is part of the definition of big data because of the tremendous amount of data that is constantly being created every day from many different sources from phone records to sales at a corporation like Target or even healthcare information. The proliferation of devices with internet connection and global positioning systems (GPS) sensors is yet another source that unlike the previously mentioned examples are a direct connection to individuals and their behavior in both virtual (online) and real (GPS) worlds (Gandomi and Haider 2014). Variety refers to the myriad sources of data and data types. Sources vary widely in their type and are not confined to technology such as warehouses full of reports that contain huge amounts of data on location and contents of archaeological sites (many states in the U.S. have digitized these records, see Texas Historical Commission Atlas Map). Another source of huge amounts of data are media such as the millions of photos that sites like Flickr and Facebook process every hour or the even larger catalogue of videos hosted on sites like YouTube. Volume refers to the sheer amount of data that is generated by these myriad sources. Finally, veracity refers to the reliability of the accuracy of the data which rely heavily on haw the data is collected, structured and disseminated.

The four V's are what allow researchers from politics to marketing to machine learning to leverage the volume and near real-time flow of data into well-informed decisions about the way that people speak, choose, and move through the world. Specifically, in the field of geography, many big data sources are drawn from social media (Liu et al. 2015; Zhi et al. 2016) where geolocated posts and/or photos are used to examine the movement behavior of people and with them, their ideas and interests. The rise of large data sets that contain geographic information such as twitter, foursquare, and other social networking sites led researchers to examine how those big data sets could be leveraged to gain understanding of how individuals move through the world. In a social sensing framework, the individual is the analog of the spectral sensor (Liu et al. 2015) on remote sensing platforms. This research uses social sensing techniques as a basis for our methodology. Specifically, the use of data that is derived from actual human interaction with the physical world and use of geo-located technology and leveraging the geo-located data to make inferences about the nature of the world and the way that humans interface with it. The built-up urban environment is the manifestation of humans engineering the word around them to move the human-nature interface to from a wild space to something slightly more controlled.

2. Literature Review

2.1. The Nature of Bluetooth Data

The use of data collected through Bluetooth specifically for geographical research is relatively new and therefore there is a paucity of scholarship on the subject. Scholarship in the domain of civil engineering has yielded several articles detailing the deployment of Bluetooth systems in several major US cities and Toronto since 2008 (Bachmann et al. 2013:34). Bachmann et al. describe Bluetooth systems to capture probe vehicles to model traffic flow and calculate velocity (2013:35). Probe vehicles were not used in the Austin Bluetooth (ATXBT) data collection. Nevertheless, the process and geometry associated with Bluetooth traffic monitoring serves to examine how the ATXBT system functions and guide the rest of the current research program.

One prominent difference between the ATXBT data and the data and processes outlined in the literature is the geometry of the roads and BT sensors (Bachmann et al. 2013). In several articles the sensors were placed next to a road at a fixed distance to account for the specific

geometry of the BT sensor. The ATXBT sensors were placed in traffic control boxes which are highly variable in their orientation with respect to the roadway. Another difference is the placement of sensors next to highways which indeed experience variability in traffic flowrate, but much less so than the stop and go traffic on city surface streets as is the case with the traffic signal loaded ATXBT sensors. Bachmann et al concluded in their study that there is a statistically significant affect between different traffic states from free-flowing to congested where congested roadways made it difficult for accurate sensing (2013:46). Since the sensors geometry and adjacent road type are different in the ATXBT system stop and go traffic might not be subject to the same effect.

Another consideration is Bluetooth penetration into the population. Recent studies have put the percentage of devices held by the total population at around 60% of phones in the U.S. and up to 80% in Canada (Friesen & McLeod 2015). Since the use of Bluetooth systems currently can only represent a percentage of the total number of cars there will always be significant unknown quantities of cars that are on the roads but are not counted. Furthermore, although there are predictions for Bluetooth penetration in the phone market, a Bluetooth system relies on a device's Bluetooth mode to be turned on but also not be paired with any other device (Friesen & McLeod 2015). Many people that have both a Bluetooth enabled phone and car pair them for music and phone calls and thus are not being counted by the system. This certainly has a significant effect on the percentage of total vehicles captured by the system. The total number of unique device addresses in the ATXBT dataset is 11,139,076 which leads to the conclusion that the under sampling of the population of Austin in this data set is not an issue. Oversampling does not seem to be an issue either since Austin is such a diverse city that attracts myriad groups of people from all over the country. The largest contributors to the device address total could be attendees of the Austin City Limits music festival and the South by Southwest festival.

2.2 Social Sensing and its Applications in the Big Data Era

Social sensing combines the techniques used in data science with techniques used in remote sensing to identify and define functional urban regions (FUR) in cities (Yuan and Raubal 2012; Ahas et al. 2015; Zhang et al. 2017, Gao et al. 2017). There are many articles that apply the concept of social sensing to identify space-time patterns (Yuan and Raubal 2012; Liu et al. 2015; Gao et al. 2017). The most important aspect of these studies is the time component. By using algorithms

to measure the similarity of time trajectories researchers can show how people move through urban landscapes and thus can define FURs based on the temporal patterns and spatial clustering (Yuan and Raubal 2012; Ahas et al. 2015; Zhang et al. 2017, Gao et al. 2017). This type of analysis provides beneficial information to city planners and business owners that is dynamic in nature and demonstrates the persistence of human mobility patterns over time. Typically, in social sensing points of interest (POI) are used as the basic geographic unit to which geolocated social media posts and check ins define. The spatial patterning between POI's is then used to define the functional urban regions such as shopping areas, business districts, education districts and tourist attractions (Zhi et al. 2016; Gao et al. 2017).

This research contrasts much of social sensing research since it uses the absolute location of Bluetooth sensors at intersections instead of mobile phones which have an average of ±5 meters accuracy or use cellular towers to triangulate the position of phones. However, the process and underpinning theory of this research is very similar to that of other social sensing research projects and reports.

2.3 Dynamic Time Warping - Its Theory and Uses

Time is the most important part of this research because variations in temporal signature help to decipher discrete FUR's. Time forms the basis for how most people schedule their day and thus time dictates when and where people begin and end their movements. For this research, the time component is comprised of time-stamped Bluetooth pings at surface street intersections across Austin. This research will use the distance time warping (DTW) algorithm to compare the time series between each intersection in the data. The DTW algorithm was chosen because it allows for comparison between two different times and can also compensate for time lags and distorted time series or time series of different lengths. Studies have used the DTW algorithm to analyze time series patterns (Yuan and Raubal 2012; Ying et al. 2016, Chen et al. 2017) concluding that it is a robust technique to deploy to analyze temporal patterns.

For example – if the time curves at 8:00 AM for two intersections do not match perfectly the algorithm can compare 8:00 AM from one intersection to 7:00 AM or 9:00 AM at another intersection so that the overall pattern of the series can be assessed. The algorithm works by

creating a matrix and then finding the shortest possible distance across the matrix providing both a vector of the path the algorithm took across the matrix and a number that gives the distance across the matrix (Ying et al. 2016). See Figure 1 below for a graphical representation. This number is used to determine the similarity or the dissimilarity between two time series. The closer to zero the more related and the further from zero the less related. The threshold at which two intersections will be deemed related or unrelated remains to be determined. That decision will be based on a cursory running of the algorithm comparing at least 20 intersections.

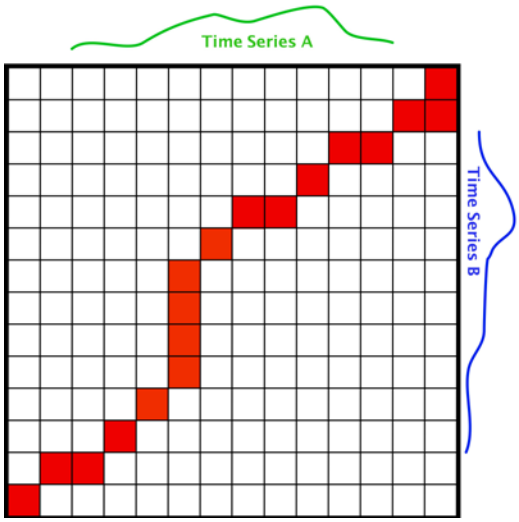

Figure 1. Example of DTW algorithm

The temporal flexibility of DTW makes it the ideal choice for this research since the overall patterns between time series are of most importance, rather than the strict hour by hour relationship between two given time series. This leads to a malleable threshold for clustering and therefore heightens the probability of larger and more robust clusters and FURs. Another reason DTW is a fitting choice for analyzing the intersection time series is its tolerance for outliers (Yuan and Raubal 2012). There is a large amount of variability in the total counts between different intersections in the ATXBT data set. The IH 35 service road and Riverside intersection is an example of an outlier on the upper end of the data as it has a consistent three-figure count for each hour of the day and the total number of unique record id counts is almost twice that of any other intersection. The DTW algorithm allows for the difference in magnitude to be secondary to the overall shape of a time curve instead of the actual counts.

2.4 Big (Geo) Data Quality Considerations

Devillers and Jeansoulin outline several types of error the can be present in any data including measurement, assignment, class and spatial generalization, data entry errors, temporal, and processing (2006:49). These errors can arise out of necessity (generalization) or user error (measurement, assignment and data entry). Another error quality is how vague a definition of an object is and how that can affect the quality of a geographic object. Vagueness errors are more difficult to control for in data since they are not empirical errors and cannot be examined using probability theory or other mathematical solutions (Devillers & Jeansoulin, 2006). Another quality issue in geographic data is precision (Devillers & Jeansoulin, 2006). This can refer to spatial precision (±10 meters for a DEM), temporal or numeric precision as in the difference in programming or database design between float and double precision.

The assurance of high-quality data is a necessary condition for any research program on or data-based decision process. In the past, data quality from sources like the United States Census Bureau were made available "scrubbed" and ready for scientific use (Goodchild 2013:281). To the contrary, Big Data cannot be produced or scrubbed in any way like the Census data is since two of the key components of Big Data are velocity and volume; too much too fast or said differently, too much data to reliably and quickly scrub and shape into useable form. Therein lies the catch-22 of Big Data. Its volumes of potentially insight-filled data points make it attractive to myriad business, intellectuals and governments, but effective process for handling the quality of such large datasets have not yet been fully developed (Cai & Zhu 2015:3).

The ATXBT data set used in this research is subject to quality considerations as well. This includes the potential for multiple counts of one vehicle as in the case of a Bluetooth enabled car and Bluetooth enabled hand-held device registering in the system as separate pings or pedestrians being captured by the system. There are also quality considerations about the percentage of Austin drivers that are captured by the system as not every car has Bluetooth and not every individual has

a Bluetooth enabled device. Furthermore, a ping can only occur and be registered in the system if the device or vehicle has the Bluetooth sensor turned on and discoverable, that is, not paired with another Bluetooth device. Though these are serious quality considerations, the methodology of this research is more interested in the time series of the number of pings and although some counts might not be accurate or capture too many or too few drives, it is not anticipated that that will have a negative effect on the results. This is because this research is specifically designed to test a specific methodology and not as a basis from which to advise city planning or engineering decisions. This methodology could certainly benefit those in civil engineering and city planning but the data set itself would need to be more fully examined to address the concerns stated above for the results to be used in that capacity.

3. Methodology

3.1 Data Pre-processing and Cleaning

The data used in this project was collected by the City of Austin, Texas over the course of 2016. The data collected totals just over 81.6 million records with each record consisting of two time stamps – field read and host read, a unique id alpha numeric code, an intersection (i.e. ih_35_riverside), and a unique MAC address tied to a Bluetooth. The MAC addresses are generated and are not tied to any identifiable information but they can be used to track a single device through the sensor locations through time. This research will use a dynamic time warping algorithm to compare the temporal sequences between two or more Bluetooth sensors located across Austin, Texas. To achieve this, we will return the cumulative counts of Bluetooth pings for each hour (8:00 AM – 8:59 AM etc.) of a 24-hour period and then create a matrix with counts at a sensor comprising the rows and time comprising the columns.

From the nine months contained in the original data set, one week (July 3 – 9, 2016) was selected as a sample for this analysis. This week was selected because there were no major events in the city (normal traffic patterns could be assumed) and this date range captured the highest number of intersections than any other week (125). Once the sample week was selected, Server Query Language (SQL) was used to calculate the total number of pings that occurred at each intersection by hour of the day – for example, the total number of pings during the 08:00 hour at

5th and Trinity was 588. The counts were then normalized by dividing each hour total count by the largest total count for that intersection. The choice as made to normalize the data because it negates the magnitude of the counts so that the temporal curve can be accurately analyzed.

The 125 sensor locations will serve as the geographical unit to which each time series will be assigned. This will allow for analysis of the patterns of similarity and dissimilarity across space-time. Further reading will help to decide how to properly divide the areas around each sensor into polygons. This might prove difficult as often in urban areas residential areas and commercial areas are right next to each other. We used Voronoi polygons derived from the sensor locations to spatially visualize the FURs. This polygon layer was clipped using the Austin city limits shapefile.

3.2 Temporal Signature Similarity and Clustering Calculations

After the matrices were calculated, the results were analyzed to find meaningful patterns. The analysis aimed to identify clustering of similar temporal patterns which, in turn, were used to define functional regions such as: work, personal spaces (i.e. bars, restaurants, music venues etc.), and residential areas. The analysis also examined outlier patterns that display counter-intuitive temporal patterns or do not fit well with the majority of other sensor temporal patterning.

The results of the DTW algorithm were then used as the input to the hierarchical cluster algorithm, part of the SciPy python package to find and define clusters of similar time series in the data set. The method used to build the clusters was Ward's method, which calculates clusters based on the minimum variance in each cluster. This process yielded 8 primary clusters. The intersections belonging to each cluster were then averaged to create an average temporal curve that was then run through the DTW and clustering algorithms to see if any of the 8 primary clusters could be grouped together. This process further narrowed the cluster count from 8 to 4 as seen in Figure 2 below. Cluster 1 is an outlier and consists of only one intersection and cluster 4 is another outlier that consists of only two intersections. Final cluster 2 is formed by primary clusters 2, 3, and 4 and cluster 3 is formed by clusters 6, 7, and 8. The dendrogram was cut based on an examination of the final cluster temporal signatures as seen in Figure 4. By cutting the dendrogram in this location allowed for less ambiguity in the two clusters that exhibit a typical diurnal pattern while allowing for the outliers to stand alone for analysis and not skew the other two clusters.

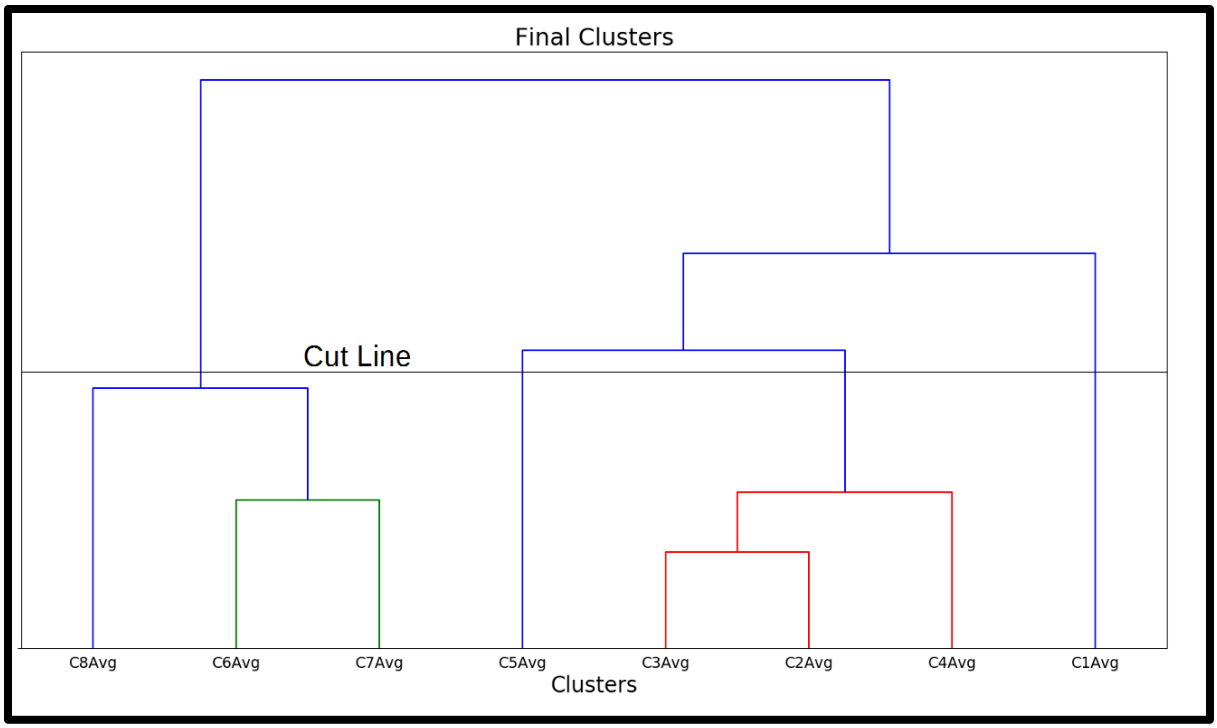

Figure 2. Hierarchical Clusters of 8 Original Clusters

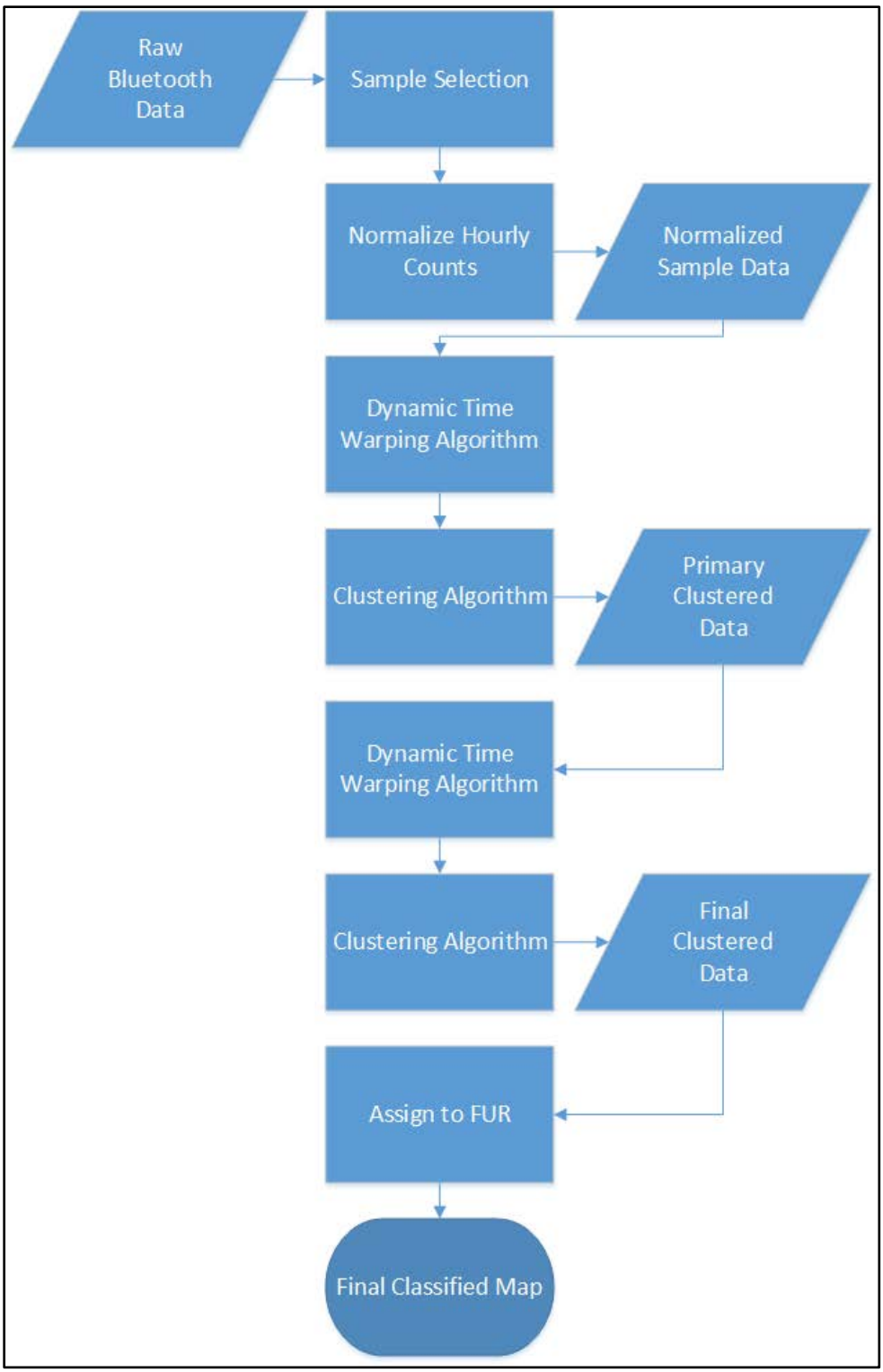

Figure 3. Methodology Flow Chart

3.3 Results

The results of the analysis show four clusters; two clusters (final clusters 2 and 3) that demonstrate a general curve that follows the diurnal cycle where activity rapidly increases beginning around 06:00, with a morning peak occurring near 08:00. The other two clusters are outliers that do not fit within an expected temporal pattern. The curve holds relatively constant through the day with a small peak at mid-day around lunch, and then a final and highest daily peak between 16:00 and 18:00. This overall temporal curve is not surprising for a large urban area like Austin. However, there are slight variations in curve shape that represent different FUR's across the city. Figure 4 below shows the average curve for the final clusters. The results have forced a reconsideration of the hypothesis that there will be three identifiable FUR's. The home and work FUR's have been collapsed into one because a significant difference between the two was not supported by the data for this research. Furthermore, two interesting outlier patterns were observed and are discussed below.

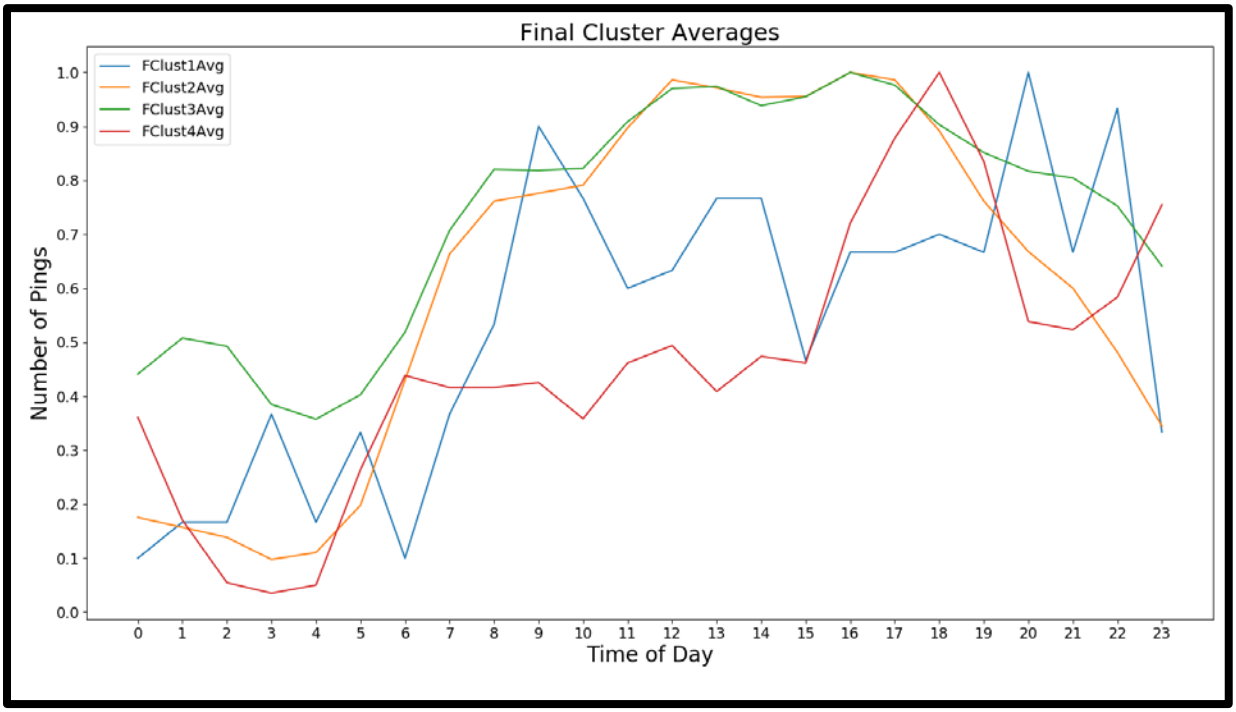

Figure 4. The temporal signatures of the four final clusters

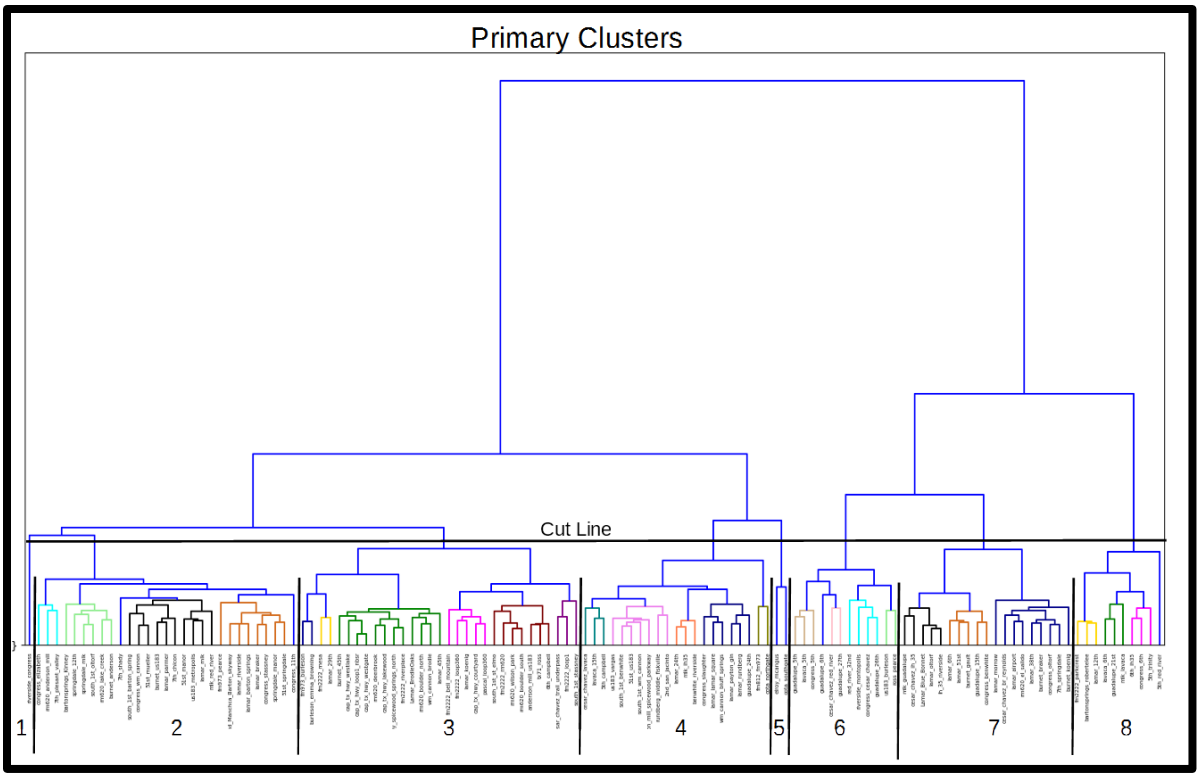

Figure 5. Primary Cluster Assignments

The primary clusters were assigned using the equation $x = \sqrt{\frac{n}{2}}$ where n = the number of intersections (Jung et al. 2003, Fouedjio 2016) in the sample dataset (125). In this case x = 7.9. The dendrogram was cut where 8 clusters could be achieved as seen in in Figure 5. The average temporal signature for each cluster were compared using the same DTW and clustering algorithms, which resulted in four final clusters (Figure 2). The temporal curve of the final four clusters were analyzed to discern which curve could represent each FUR – home, work, and recreation/other. The final results of the research show four final clusters. Two of the clusters represent two distinct patterns that suggest a day use /night use dichotomy. The two other clusters are outliers that exhibit unusual patterns that do not fit with expected results. The FUR's were determined by examining the patterning of each cluster time curve before 08:00 and after 15:00. The average curve for all clusters is similar between these two times, which represent the beginning and end of the typical workday.

Final cluster 2 (orange line in Figure 6 below) was assigned to the home/work FUR for several reasons. First, the low frequency of pings in the early morning and late evening, the steep

curve between 05:00 and 08:00 and after 18:00, and the peaks at 12:00 and 17:00. The steep curve between 06:00 and 08:00 is indicative of many people leaving home to go to work in the morning and the opposite is true for the precipitous drop in pings after 18:00. The peak at 17:00 indicates people heading home after work.

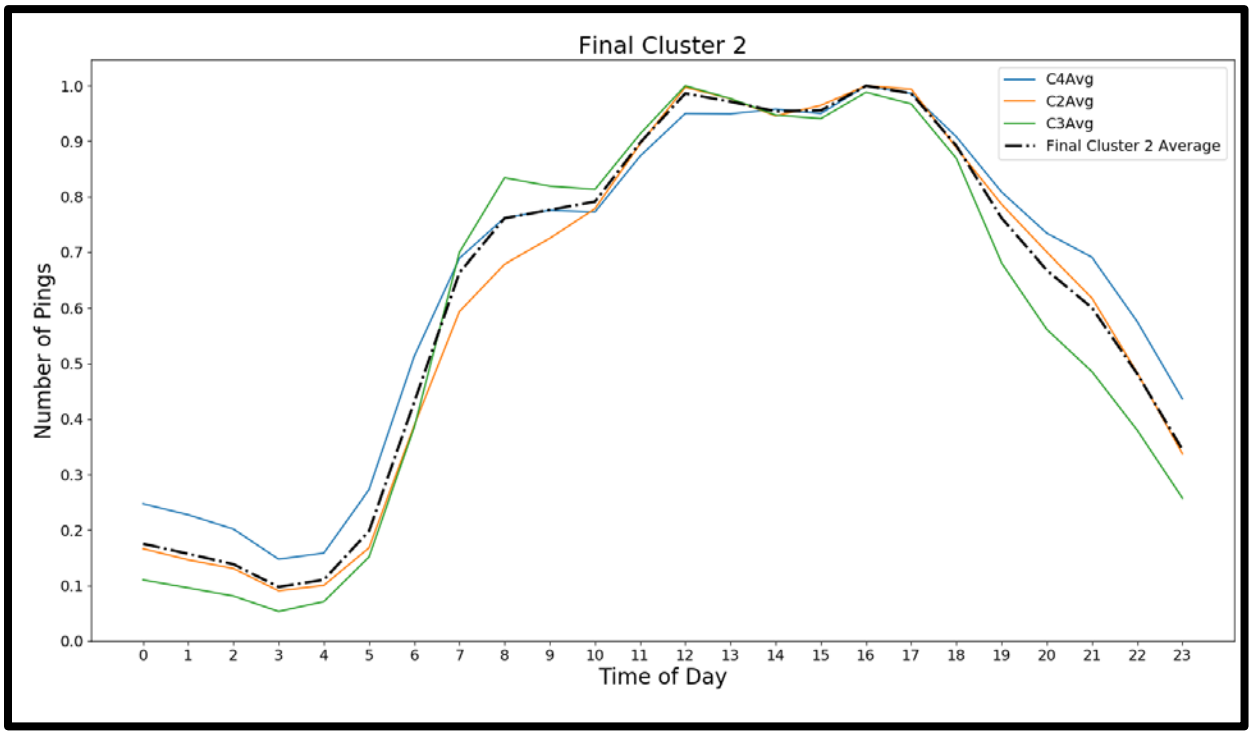

Figure 6. Final Cluster 2 and the average for clusters 2, 3, and 4.

Final cluster 3 (Figure 8) was demonstrated a pattern that allowed it to be identified as distinct from final cluster 2. It was assigned to the recreation/other FUR. There is a distinct peak between 01:00 and 02:00, which suggests activity around the time when bars close in Austin or concerts conclude and patrons head home. The final cluster 3 temporal signature differs from the other cluster in the lack of a precipitous drop in ping frequency after the peak at 17:00. This pattern reflects that the intersections in final cluster three represent places where traffic is heavier in the traditional after work hours where people would go to dinner, movies, concerts, bars, etc.

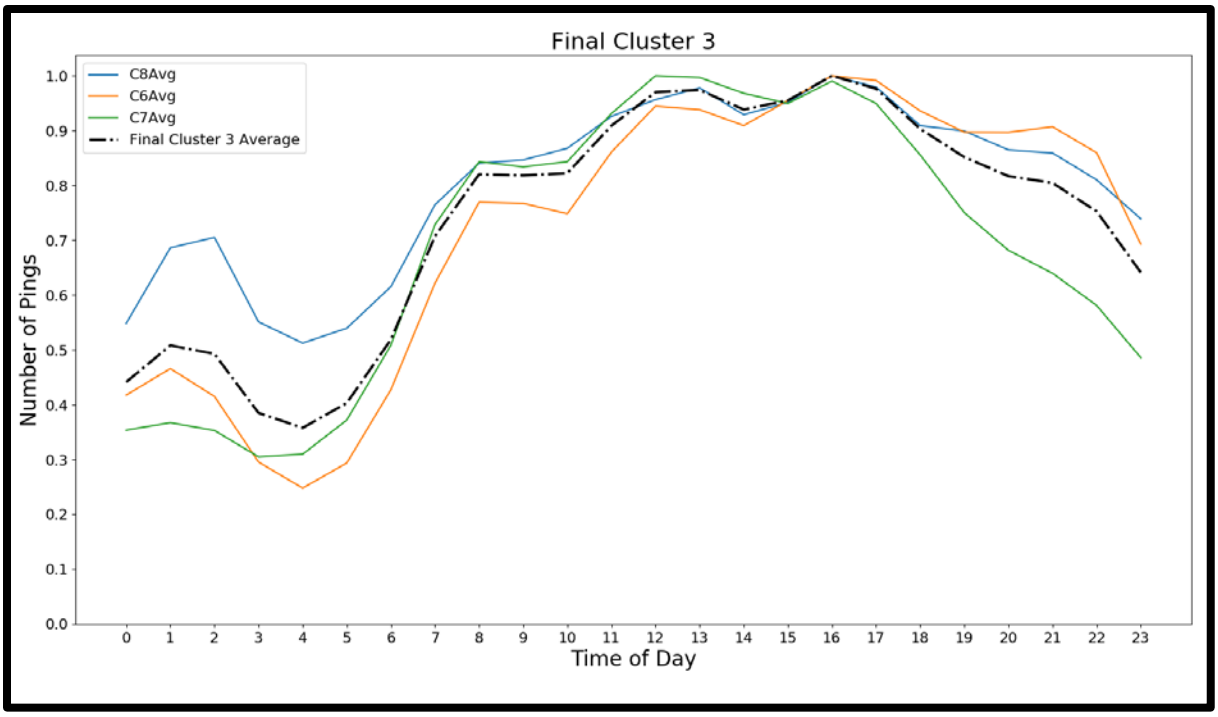

Figure 7. Final Cluster 3 and the average curve for primary clusters 6, 7, and 8.

Though all three clusters exhibit similar overall curves, specific intersections in the data set are outliers based on their shape. These intersections are Riverside and Congress, Circuit of the Americas Southgate, and Elroy and McAngus. These intersections form final cluster 4 and Riverside and Congress is the sole member of final cluster 1. Its pattern is so distinct from all other intersections that it was never put into a cluster with any other intersection. Likewise, the intersections of final cluster 4 persisted as a distinct cluster together through to clustering and DTW iterations. These intersections are all members of primary cluster 7 (below), which is part of final cluster 2 – the work FUR. In Figure 4, these intersections clearly show how dissimilar they are from the curve of all the other clusters. Most notable is how they do not follow the distinct diurnal pattern exhibited by final clusters 2 and 3. The final cluster 4 intersections are both located in far east Austin, where traffic patterns are more variable than the rest of the city. Riverside and Congress is an interesting and unexpected result. It is located near the center of Austin and it is difficult to hypothesize why such a result occurred. Given the location of this intersection in Austin, this intersection could represent a mixed-use area that does not exhibit patterns that can be readily identified by the broadly defined temporal signatures in this study. Further research could examine this type of outlier but cross referencing the area with a land use map or census data to

get a clearer picture of how this area is used. More research should be conducted to examine how the final cluster 4 outlier intersections might change in cluster membership and average temporal curve over time since they are on the periphery of the Austin city limits.

The final map shows the location of each of the two FUR and two outlier patterns in Austin. The polygons representing each region were built using Delaunay triangulation. This process was used because it is the best way to represent area coverage based on points but it does introduce so uncertainty since home areas, work areas and recreation areas can and often are closely spaced meaning that the polygons in the final map might not accurately represent the on-the-ground use spaces. Expanding the research to include all intersections in the data and using different techniques to spatially visualize the data might produce more accurate results. However, the main purpose of this research was to ascertain the efficacy of this methodology instead of striving for the most accurate representation of FUR. If a follow up study is possible, making a more accurate map of FUR will be a higher priority.

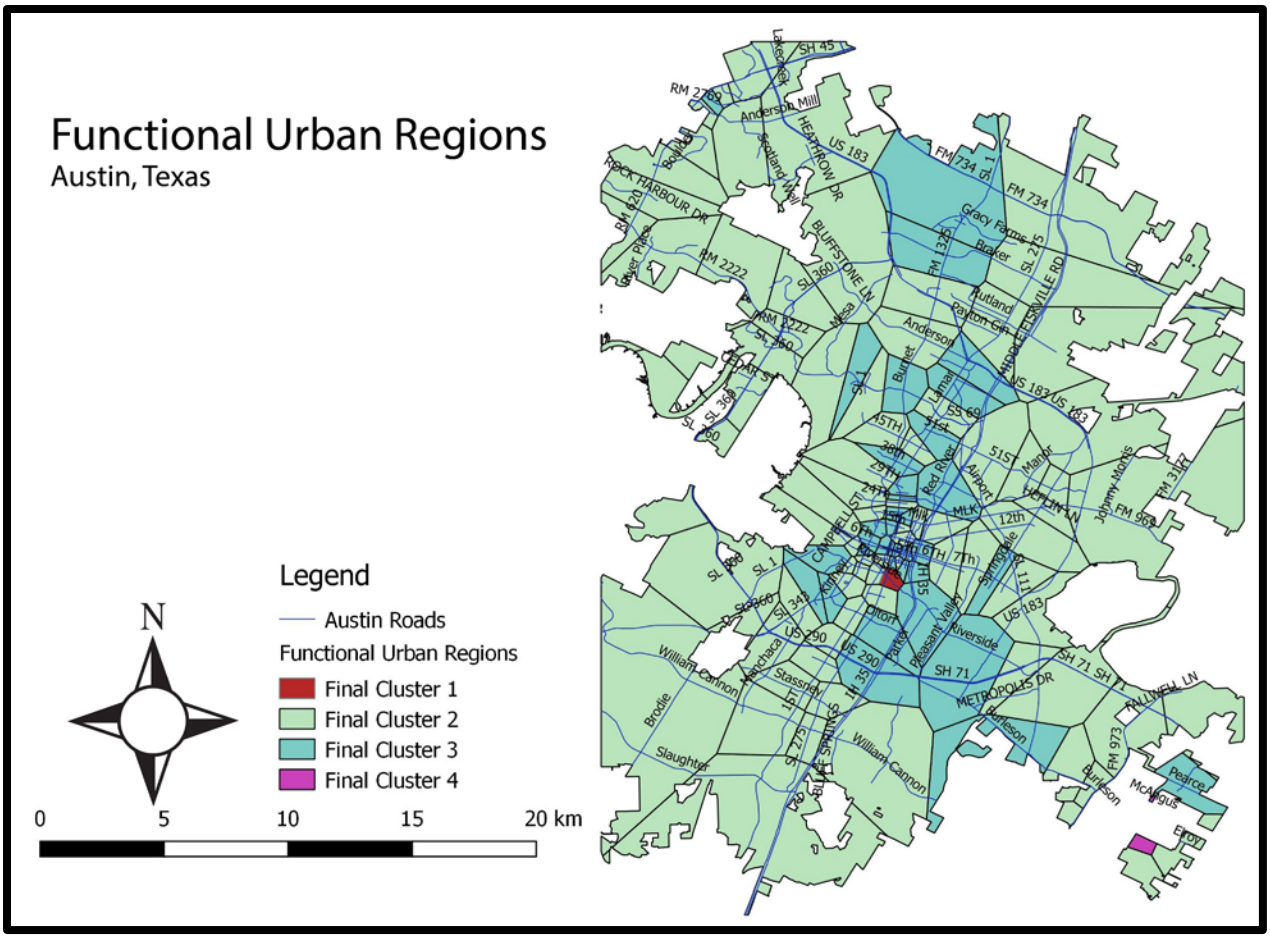

Figure 8. Final Classified map of FURs in Austin.

4. Conclusions

This research used Bluetooth data acquired from the City of Austin to identify FUR across the city. This dynamic method can demonstrate different patterns and show changes in the location of FURs over time with higher accuracy and at a shorter turn around than traditional remote sensing techniques. The results of this research should be of interest to city planners, marketing strategists and entrepreneurs to guide where city infrastructure, advertisement and new business should be located respectively. This research should also interest other researchers in myriad fields who are interested in spatio-temporal modelling and big data analysis. The use of Bluetooth data shows the efficacy of another big data source that can be leveraged to achieve similar results to location-based social media sources.

Using Bluetooth data instead of location-based social media makes use of the exact location of every sensor instead of triangulated cell tower locations or GPS error in mobile phones. The exact locations allow for more precise measurement at exact locations. However, there exists the potential for errors in the Bluetooth data as well, including multiple counts for what should be a single ping (a Bluetooth enabled phone in a Bluetooth enabled car counting as two instead of one) or odd patterns such as one device only going back and forth between two proximate intersections multiple times a day. The latter error type does exist in the dataset but represents less than 0.00003% of the total number of records and therefore does not pose a significant obstacle to the accuracy of the results. Further research into the sources for error in Bluetooth data collection should be addressed in another study to refine a technique that has enormous potential for providing insights into the mobility patterns of individuals and the way that individuals and the population use the road network.

Only one week was used as a sample from this large data set and further research should be conducted to examine if the patterns present in the sample are indicative of the entire data. Austin hosts several large festivals throughout the year and it would be interesting to examine how these festivals affect the traffic patterns of the city. Studying the temporal curve for intersections in this way is sensitive to changes in traffic patterns so it would make sense that some intersections would change from one cluster to another as traffic patterns change over time. Further research using this same data set would most likely produce interesting results about how the temporal signatures change over time. This could allow for a comparison between low traffic and high traffic events (e.g. a normal traffic week and a week where the pattern is augmented by an event like South by Southwest). Further research could also examine potential errors in the data and attempt to calculate the percent of the Austin population captured by the Bluetooth system. This would allow researchers to explore how representative this data is of Austin traffic trends.

References

Ahas, R., A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook. 2015. *Every Day Space-time Geographies: Using Mobile Phone-Based Sensor data to Monitor Urban Activity in Harbin, Paris, and Tallinn.* International Journal for Geographic Information Science. 29 (11): 2017-2039.

Bachmann, C., B. Abdulhai, M. J. Roorda, and B Moshiri. 2012. *A Comparative Assessment of Multi-Senor Data Fusion Techniques for Freeway Traffic Speed Estimation using Microsimulation Modeling.* Transportation Research Part C. 26: 33-48.

Cai, L., and Y. Zhu. 2015. *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.* Data Science Journal, 14 (2): 1-10.

Chen, Y., X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei. 2017. *Delineating Urban Functional Areas with Building-level Social Media Data: A Dynamic Time Warping (DTW) Distance Based k-medoids Method.* Landscape and Urban Planning. 160: 48-60.

Devillers, R., and R. Jeansoulin. 2006. In *Fundamentals of Spatial Data Quality.* 43-59

Fouedjio, F. 2016. *A Hierarchical Clustering Method for Multivariate Geostatistical Data*. Spatial Statistics. *18*, 333-351.

Friesen, M. R., and R. D. Mcleod. 2015. *Bluetooth in Intelligent Transportation Systems: A Survey.* International Journal of Intelligent Transportation Systems Research. 13:143-153.

Gandomi, A., and M. Haider. 2014. *Beyond the Hype: Big Data Concepts, Methods, and Analytics.* International Journal of Information Management. 35:137-144.

Gao, S., K. Janowicz, H. Couclelis. 2017. *Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-Based Social Networks.* Transactions in GIS 21(3):446-467.

Goodchild, M. F. 2013. *The Quality of Big (geo)Data.* Dialogues in Human Geography 3(3):280-284.

Herold, M., J. Scepan, K. C. Clarke. 2002. *The Use of Remote Sensing and Landscape Metrics to Describe Structures and d Changes in Urban Areas.* Environment and Planning A 34(8):1443-1458.

Hu, T. J. Yang, X. Li, and P. Gong. 2016. *Mapping Urban Land Use by Using Landsat Images and Open Social Data.* Remote Sensing 8 (2)151:170.

Jung, Y., Park, H., Du, D. Z., & Drake, B. L. 2003. *A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering*. Journal of Global Optimization *25*(1): 91-111.

Liu, H. and Q. Zhou. 2003. *Accuracy analysis of remote sensing change detection by rule-based rationality evaluation with post-classification comparison.* International Journal of Remote Sensing. 25(5):1037-1050.

Liu, Y, X. Liu, S. Gao, L. Gong, C. Kang, and Y. Zhi. 2015. *Social Sensing: A New Approach to Understanding our Socioeconomic Environments.* Annals of the Association of American Geographers. 105 (3):512-530

Sagal, G., M., Loidl, and E., Beinat. 2012. *A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic.* ISPRS International Journal of Geo-Information. 1(3):256-271.

United Nations, Department of Economic and Social Affairs, Population Division. 2014. *World Urbanization Prospects: The 2014 Revision, Highlights* (ST/ESA/SER.A/352).

Ying, R., J. Pan, K. Fox, and P. Agarwal. 2016. *A Simple Efficient Approximation Algorithm for Dynamic Time Warping*. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 21. ACM, 2016.

Yuan, Y. and M. Raubal. 2012. *Extracting Dynamic Urban Mobility Patterns from Mobile Phone Data.* GIScience: International Conference on Geographic Information Science 7th International Conference. 354-367.

Yuan Y. 2017. *Exploring the Spatial Decay Effect in Mass Media and Location-Based Social Media: A Case Study of China*. Advances in Geocomputation. Advances in Geographic Information Science. 133-142.

Zhang, Y., Q. Li, H. Huang, W. Wu, X. Dui, and H. Wang. 2017. *The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China.* Remote Sensing 9(9):865-888.

Zhi, Y., H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan and Y. Liu. 2016. *Latent Spatio-Temporal Activity Structures: A New Approach to Inferring Intra-Urban Functional Regions Via Social Media Check-in Data.* Geo-spatial Information Science 19(2): 94-105

Zimmerman, P.L., I.W. Housman, C.H. Perry, R.A. Chastain, J.B. Webb, and M.V. Finco. 2013. *An accuracy assessment of forest disturbance mapping in the western Great Lakes.* Remote Sensing of the Environment. 128:176-185.