

EXPLORATION OF MUNICIPAL MOBILITY USING SMARTPHONE GPS DATA

Xiaoyi Peng¹, Ce Wang¹, Ke Liu¹, Agustin Quinones¹, Shuo Li², Jie Shan¹ *

¹Lyles School of Civil Engineering, Purdue University, USA – {peng183, wang4101, quinona, jshan}@purdue.edu;
keliu0530@gmail.com

²Research Division, Indiana Department of Transportation, USA - sli@indot.in.gov

Commission IV WG IV/10

KEY WORDS: Smartphone GPS data, Municipal mobility, Flow map, Kernel density function, Traffic volume prediction

ABSTRACT:

Due to the advances in location-acquisition techniques, smartphone GPS location data has emerged with opportunity for research, development, innovation, and business. A variety of research has been developed to study human behaviour through exploring patterns from these data. In this paper, we use smartphone GPS location data to investigate municipal mobility. Kernel density estimation and emerging hot spot analysis are used to the GPS dataset to demonstrate GPS user (point) distribution across space and time. Flow maps are capable of tracking clustering behaviours and direction maps drawn upon the orientation of vectors can precisely identify location of the events. Case study with Indianapolis metropolitan area traffic rush hour verifies the effectiveness of these methods. Furthermore, we identify smartphone GPS data of vehicles and develop a concise and effective method for traffic volume estimation for county highway network. It is shown that the developed smartphone GPS data analytics is powerful for predicting reliable annual average daily traffic estimation. The study showcases the capability of GPS location data in identifying municipal mobility patterns for both citizens and vehicles.

1. INTRODUCTION

Understanding mobility of human beings at the municipality scale is a key factor for understanding the mechanism of urban expansion and smart cities. It is quickly becoming a key task in many GIScience applications (Dodge et al., 2012). Applications such as urban planning, relieving traffic congestion, and effective location recommendation systems are important objectives worldwide and have received increasing attention these past years (Wu et al., 2018). In this paper, smartphone GPS data is used to investigate patterns of municipal mobility.

One way to investigate distribution of human activities across space and time is by using Kernel Density Estimation (KDE). To further express human movement, flow maps have been considered as an effective means for spatial interaction visualization and such data can form a location-to-location network (Kim et al., 2017). In a flow map, the connection between two locations is usually represented by a straight or curved flow line and magnitude of flow could be visualized by varying colours or width of the flow line (Guo et al., 2009).

Another aspect of investigating municipal mobility is to evaluate the traffic volume around the city. Municipal traffic volume is a primary factor to reflect spatial pattern of human mobility. Meanwhile, it is also a key element of transportation system design. Federal Highway Administration (FHWA). State Department of Transportation (DOTs) invest a vast amount of expense on traffic monitor program to collect traffic count data by traditional means. The most wide-applied method is generating annual average daily traffic (AADT) from permanent and short-term traffic count station installed at representative roadway sites (Chowdhury et al., 2019). Intuitively, collecting continuous traffic volume for every road segment is an ideal way to get perfect traffic count data. But it is not economically efficient for most road segments, especially for rural roads with low traffic volume. It is highly demanded to estimate reliable AADT with least data collection effort. With the popularization of smartphones over the past two decades, millions of trajectories

with location and timestamp information recorded when users travel along roads can be applied to traffic pattern research. Hence, mining the big smartphone GPS data is becoming an emerging approach for human movement as well as vehicle movement pattern understanding.

This paper is organized as follows: Section 2 shows related work with GPS location data. Section 3 describes the smartphone GPS location data and the study area used in this paper, while Section 4 presents the methods for a density difference-based flow extraction model and for traffic volume estimation. Section 5 shows the results of municipal mobility pattern with experimental work. Section 6 summarizes the findings of this study.

2. RELATED WORK

For the past two decades, GPS data extracted from GPS-enabled devices such as smartphones, floating cars, taxis, GPS logs, etc., has increased the opportunity for researchers to collect large amount of spatio-temporal data for human mobility study and diverse applications. For example, Zheng et al. (2009) and Li et al. (2008) mine interesting locations, user similarity and classical travel sequences to help users understand surrounding locations and enable travel recommendations. They use a tree-based hierarchical graph (TBHG) to model individuals' location histories, a hypertext induced topic search (HITS) inference model to infer users' travel experiences and interests, and a hierarchical graph-based similarity measurement (HGSM) to model each individuals' location history and to efficiently measure the similarity among users to retrieve information with high relevance. Ahas et al. (2007) use mobile phone data to analyse the seasonality of foreign tourists' space consumption in Estonia, where mobile phone data has the potential to record invisible flows of tourists that can be used for better marketing and services. Movement and mobility are important aspects of human behaviour. Siła-Nowicka et al. (2016) propose an approach that combines volunteered GPS trajectories and contextual spatial information to analyse human mobility. Each

* Corresponding author

individual GPS trajectory is segmented by a spatio-temporal kernel window (STKW) statistics. The segments are classified by a feed-forward neural network. Then locations of significant places are identified by stop segments and contextual spatial information. Kim et al. (2017) propose a gravity-based flow extraction method, which extracts and visualizes population movement patterns of non-directional GPS data. This model generates data flows from heat maps of GPS points density at different time steps. As such, it does not require any trajectory information. Xia et al. (2018) quantitatively represent human movement pattern on weekends and weekdays in Shanghai through analysing subway smart card transactions and taxi GPS trajectories. Bayesian weights are calculated by logarithmic binning and data fitness to find the best fitting distributions for spatiotemporal data. Shen et al. (2018) propose a deep neural network to predict spatial-temporal mobility event. This model simultaneously considers all correlated spatial-temporal mobility patterns and switch mobility events over time in an area into an event video.

As a one of the most important variables to quantify traffic volume, AADT has been computed using probe data for non-surveyed road sections by converting the probe volume of target road section to AADT using the nonlinear relationship between geographical neighbourhoods composed of observed AADT volumes and annual average daily probe volumes (Chang et al., 2019). Castro-Neto et al. (2009) propose support vector machine for regression with data-dependent parameters (SVR-DP) to forecast AADT. They compare its performance with those of Holt exponential smoothing (Holt-ES) and of ordinary least-squares linear regression (OLS-regression). Caceres et al. (2012) use cellular phone data to develop a set of models to infer the number of vehicles moving from one cell to another to estimate the traffic flow. Wang et al. (2013) represent a method to estimate AADTs for local roads by parcel-level trip generation model that estimates the trips generated by each parcel. Zhan et al. (2017) integrate machine learning techniques and well-established traffic flow theory to estimate citywide traffic volume, as an alternative to traffic volume estimation with road sensors. Sun et al. (2019) use GPS trajectory of floating cars to predict traffic congestion using the Hidden Markov Model to match GPS trajectory data to the road network, estimate the average speed of a road section for map matching and predict the traffic congestion using deep learning techniques. Zhang et al. (2013) and Fan et al. (2019) have come up with a method of estimating VMT and AADT of local roads and state-wide roads from vehicle GPS data with trajectory information.

3. DATA AND STUDY AREA

Indianapolis, the most populous city in Indiana State and the seat of Marion county, is taken as an example of municipality for human mobility exploration. The GPS data used in this research is provided by SafeGraph, a company that collects location information from numerous smartphone apps of anonymous users that have given permission to have their location data shared. The dataset consists of GPS records of more than 57.92 million GPS points from 914,809 users for the study area (bounding box in coordinates: $\min(x) = 86.40^\circ$ W, $\min(y) = 39.63^\circ$ N, $\max(x) = 85.88^\circ$ W, $\max(y) = 39.98^\circ$ N). This dataset includes the following 7 attributes: "safegraph id", an anonymous ID for each user, geographical coordinates latitude and longitude in the WGS 84 reference system, the horizontal accuracy in meters, the id type, the timestamp in UTC and the geohash. The extracted time period used here to analyse traffic pattern and human movement pattern is "2017-08-25 00:00:00" to "2017-08-31 23:59:59" in Local Time (EST).

The horizontal accuracy index of the data is a radius about a GPS point, implying that the true location is somewhere within the circle formed by the given point as a centre and the horizontal accuracy index as a radius. Points with a larger horizontal accuracy index are less likely to be accurate, which would affect the results of spatial analysis. Therefore, points with horizontal accuracy index beyond a threshold should be eliminated. On the other hand, if the threshold is set to small, the number of remaining points would be not enough to do a valid spatial-temporal analysis. Thus, it is critical to select an appropriate threshold for horizontal accuracy. Fig. 1 shows the relationship between different values of threshold and corresponding number of remaining points after filtering. As Fig.1 shown, when threshold decreases from 50 m, the number of remaining points decreases slightly. However, once threshold is less than 20m, the data size would decrease significantly. Therefore, 20 m are selected as the threshold for the horizontal accuracy index. Points whose horizontal accuracy index is larger than 20 m would be filtered out.

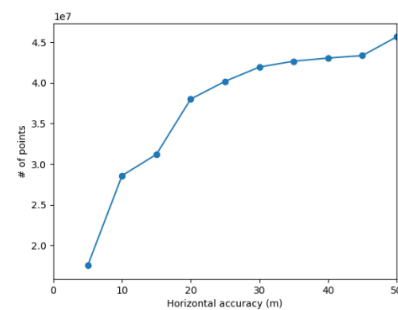


Figure 1. Threshold selection for the GPS horizontal accuracy

Human movement pattern analysis will only use GPS data on weekdays since there are only two weekends in this dataset, which may not be sufficient to draw any solid conclusion.

For traffic pattern analysis, two additional datasets are used: Indiana highway network and 2017 traffic volume shapefile. The highway network data is from INDOT for Marion county (Indianapolis). Indiana DOT collects annual average daily traffic (AADT) data by two traffic-monitoring systems: Statewide Traffic Monitoring System and Statewide Coverage Count Program, which respectively collect traffic counts data continuously and in short term. AADT shapefile is generated from these permanent and portable traffic sites by INDOT internal GIS inventory system. In this study, 103 road segments with a total of 135 miles long in Marion county state highway, interstate highway and US highway are extracted from 2017 AADT shapefile.

4. METHODOLOGY

4.1 GPS User Distribution

This is inferred through the smartphone GPS data, including both pedestrians and aboard vehicles. To obtain continuous distribution from discrete smartphone GPS data, kernel density estimation (KDE) is applied (Simonoff, 2012). Generally, a multivariate KDE can be expressed

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_H(x - X_i) \quad (1)$$

$$K_H(x) = |H|^{-1/2} K(H^{-1/2}x) \quad (2)$$

where $x = (x_1, x_2, \dots, x_d)^T$, $i = 1, 2, \dots, n$

$X_i = (X_{i1}, X_{i2}, \dots, X_{id})^T$, $i = 1, 2, \dots, n$

$H =$ symmetric and positive definite $d \times d$ bandwidth matrix

$K =$ kernel function

KDE is capable of smoothing out each data point into a smooth bump. The shape of the smooth bump is determined by the kernel function $K(x)$, whose main role is to confer differentiability and smoothness properties on the resulting estimate. The kernel function selected in this work is Gaussian kernel, which is the most widely used one:

$$K_H(x) = (2\pi)^{-d/2} |H|^{-1/2} e^{-\frac{1}{2}x^T H^{-1}x} \quad (3)$$

It is important to find an optimal data-driven bandwidth when KDE is applied. The measurement of the performance of \hat{f} is the mean integrated squared error (MISE) criterion (Wand et al., 1994):

$$MISE(H) = E \int_{\mathbb{R}^d} (\hat{f}(x; H) - f(x))^2 dx \quad (4)$$

The selection of H is usually determined by the minimization of MISE over the space of all symmetric positive definite $d \times d$ matrices:

$$H_{MISE} = \underset{H}{\operatorname{argmin}} MISE(H) \quad (5)$$

4.2 Population Flow

The flow extraction model is built upon the difference of density distribution between two-time stamps. This model assumes that when the number of users increase or decrease at a location, they either come from or go to surrounding cells. Spatial distribution in this model is expressed by a probability density matrix, which is estimated by KDE and visualized as a heat map. To illustrate this model, heat maps of population at two-time stamps are shown in Fig. 2a and 2b, respectively. The KDE difference between two heat maps is shown in Fig. 2c. Location i is picked as a sample location and locations j and k are neighbouring cells of location. From t_1 to t_2 , the population at location i decreases so it imposes an incoming flow to location i . Direction of this flow is from j to i and magnitude is the KDE difference at location j . Similarly, the flow from i to k is generated since population in location k increases from t_1 to t_2 . The final flow at location i could be formed by adding all of flows formed by all neighbouring cells. The formula for calculating flow vector (u, v) at any location i is given by

$$\begin{bmatrix} u \\ v \end{bmatrix} = \sum_{p=-W}^W \sum_{q=-W}^W \frac{KDE(x_i, y_j)|_{t_2} - KDE(x_i, y_j)|_{t_1}}{d_{ij}} \begin{bmatrix} p \\ q \end{bmatrix} \quad (6)$$

where W is size of window that defines the farthest surrounding cells for a given location
 $[p, q]^T$ is the vector from (x_i, y_i) to (x_j, y_j) .

This equation could be divided into two parts, which calculate flow along the x -axis and y -axis separately:

$$U = (KDE|_{t_2} - KDE|_{t_1}) \times kernel_p / kernel_d \quad (7)$$

$$V = (KDE|_{t_2} - KDE|_{t_1}) \times kernel_q / kernel_d \quad (8)$$

where U is the matrix of all x coordinates in a vector field
 V is the matrix of all y coordinates in a vector field

The calculations are implemented with kernels of $(2W+1) \times (2W+1)$ matrix.

W is a critical parameter in this model and could affect the flow extracted from heat maps. If W is not large enough, it only extracts local flow patterns rather than the global flows. However, if a very large W is picked, distant cells which could not affect the KDE at centre point would be regarded as surrounding cells. Therefore, it is an important task to select an

appropriate W . In this work, in order to recognize population movement pattern, W should depend on time span between two heat maps and speed of population movement.

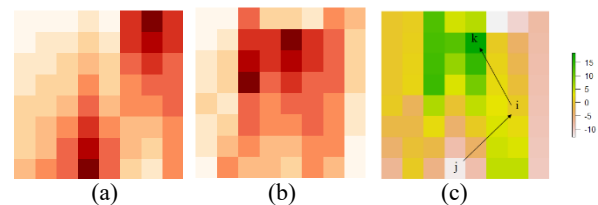


Figure 2. Flow extraction by density difference-based method. (a) Heat map at t_1 . (b) Heat map at t_2 . (c) KDE difference between t_1 and t_2 .

4.3 Traffic Volume

The steps of our traffic volume estimation method are filtering, map matching, aggregation and regression. More detailed description on each step is shown as below.

Filtering is to calculate the speed of temporal consecutive recording of the same user, then exclude those that are not driving. Smartphone GPS data can be recorded from any smartphone user allowing location sharing at any time. But it does record more frequently when users are moving. The recording frequency varies from 1s to 400s depending on users and time period. To only involve users on a vehicle into traffic volume estimation, only points with absolute speed higher than 2m/s from last temporal consecutive point are extracted as pedestrians can barely walk/run faster than this speed. Meanwhile, a user must be recorded more than 5 times in a day.

The aim of **map matching** is to locate GPS recordings as spatial points, and assign points within highway network buffer to the nearest AADT road segments. Highway network is clipped by Marion county from Indiana highway double line network shapefile. Raw GPS recordings are converted to spatial points and projected to NAD 1983 UTM zone 16. Then, a 12m buffer on both sides of the highway network was created to intersect with traffic points of users aboard vehicles on these roads. The single line road network with traffic volume information shares the same shape with highway network shapefile but have its own traffic sections as segments of roadway that are considered to have relatively uniform travel characteristics. The approach of matching selected traffic points to traffic volume shapefile road segment is simply assigning points to the nearest traffic sections. At the end, the total number of traffic points in our area of interest is 5.23 million.

In the **aggregation** step, we combine the number of unique users to each road segment. The number of traffic points assigned to road segments everyday was counted as Daily Traffic (DT). Ideally, the daily traffic of all these 190 road segments should be compared with actual AADT in next step. However, there are some road segments that have short length that tend to be skipped in an individual trip resulting in statistical uncertainty. Sampling threshold is selected by making a trade-off between the coverage of road network and regression assessment.

Finally, the **regression** step finds linear correlation between traffic count from the GPS data and actual AADT. From the data used, 7 days, including 5 weekdays and 2 weekend days can best represent the yearly average traffic. Their average daily traffic referred to as GPS_ADT is calculated. As GPS data record both dynamic and static devices, we can estimate sampling ratio from regression between GPS_ADT and real AADT.

5. RESULTS AND EVALUATION

5.1 Patterns of Human Mobility

In order to explore human movement patterns during traffic rush hour, the hour between 7:30 am and 8:30 am is selected to study morning rush hour. Similarly, the evening rush hour is selected from 5 pm (17:00) to 6 pm (18:00). GPS data in the range of ten minutes before and after each timestamp are also included since data size at these exact moments is not large enough. In order to show the movement pattern in the morning rush hour of the weekday, a heat map based on the Gaussian kernel is derived from discrete GPS data as shown in Fig. 3.

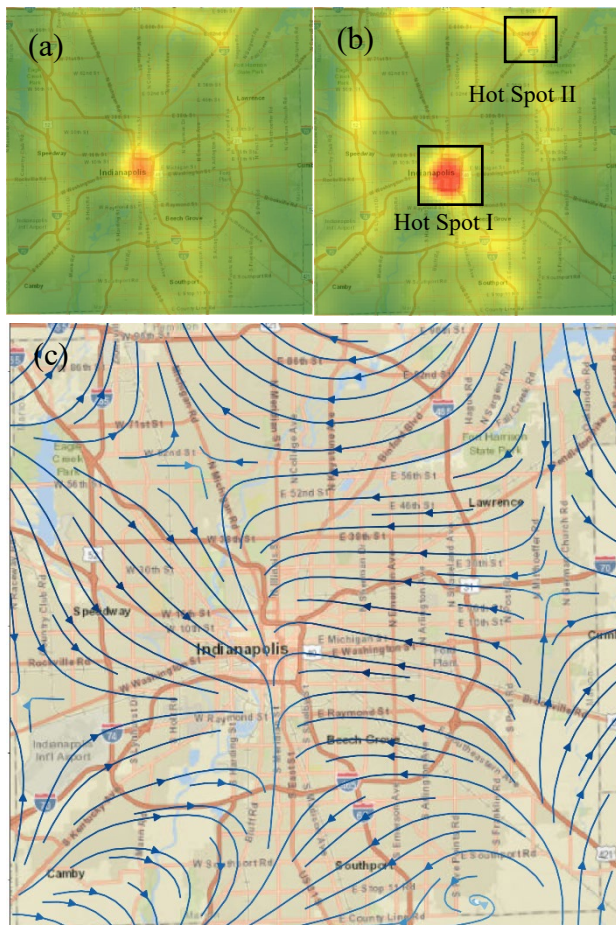


Figure 3. Heat map before and after morning rush hour. (a) Heat map at 7:30 am. (b) Heat map at 8:30 am. (c) Flow map during morning rush hour.

Each pixel in the heat map represents a 200 m by 200 m area and the corresponding pixel value represents the population (GPS user) density evaluated by KDE at the centre of each cell. As shown in Fig. 3a and 3b, there are two regions that change significantly from 7:30 am to 8:30 am and both are hotspots in the heat maps. As shown in Fig. 3a, hotspot I is around the centre of Marion County where a university and hospital campus resides. Both GPS points density and size of this hotspot increase, which indicates that the population aggregates in this area during 7:30 am to 8:30 am. This is because a large number of people work or study there but live elsewhere. They are expected to travel to this area during the morning rush hour. Hotspot II is located at north-eastern Marion County and exhibits the same pattern. There is an interstate highway intersection where many vehicles would pass, resulting in a significant increase of GPS points density during morning rush hour. However, at both of

7:30 am and 8:30 am, most GPS points get clustered in the downtown area of Indianapolis, forming an irregular distribution. The rest of the points scatter around the map and reveal some small hot spots near highway intersections. However, it is still hard to recognize where and how people move.

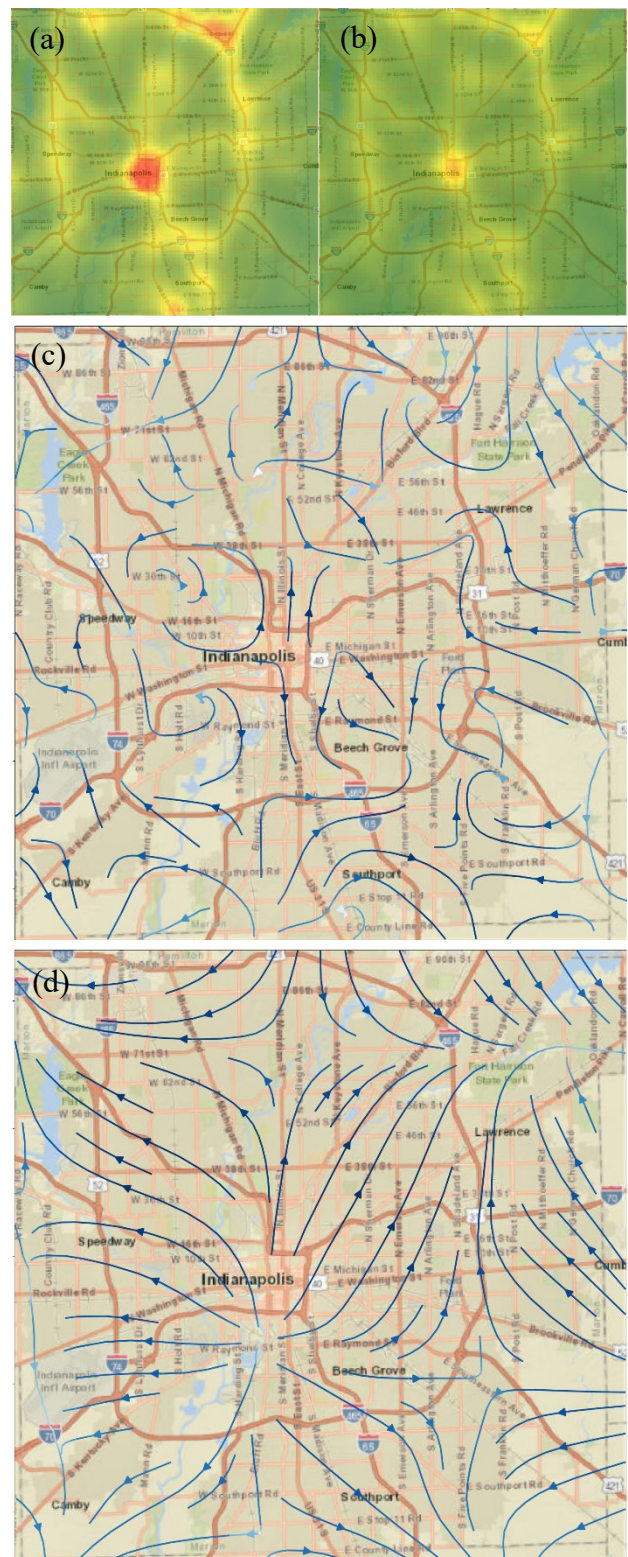


Figure 4. Heat map before and after evening rush hour. (a) Heat map at 17:00. (b) Heat map at 18:00. (c) Flow map during evening rush hour with $W = 15$. (d). Flow map during evening rush hour with $W = 40$.

Density difference-based flow extraction method is applied to form flow maps from the heat maps at 7:30 am and 8:30 am. As Fig. 3c shows, the most noticeable movement pattern is that population converge in the downtown of Indianapolis during morning rush hour, resulting in an increase of GPS points density in this area. However, no convergence is observed near the interstate highway intersection. Thus, density of GPS points increases there only because there are more vehicles go through the highway during rush hours. Therefore, flow maps can illustrate different patterns of population movement while heat maps could not.

Heat maps at 17:00 and 18:00 are shown in Fig. 4a and Fig. 4b. Compared to Fig. 3, during this period, the density of GPS points in the downtown area decreases significantly, indicating that people are leaving from the downtown area. The flow maps during evening rush hour are extracted by the same method with $W = 15$ and $W = 40$. As shown in Fig. 4c and Fig. 4d, a small W results in shorter flow lines and less obvious population movement pattern. As mentioned in Section 4, the value of W is key to flow map and should be taken carefully in accordance with the data. When W is too small, flow maps extracted by density difference-based method can better reflect the local movement pattern while neglects the global trend. On the other hand, for a given cell, contribution from cells too far away should be excluded from the sum. Otherwise, the cell will be calculated wrongly and the flow will be over-smoothed. As Fig. 4d shown, compared to flow map in the morning rush hour, the population movement shows an opposite pattern, that is, population diverges from the central area of Indianapolis. The second noticeable location in Fig. 4c is the highway at the northeast side of Indianapolis, the population converges in this area during the evening rush hour.

5.2 Traffic Volume Estimation

To optimize the correlation between GPS_ADT (average daily traffic) in 7 days and the INDOT 2017AADT, a threshold of road segment length should be selected to keep as many sample roads as possible and get good regression result.

Fig. 5 shows the number of samples and R squares with respect to the road segment length threshold. Compromising the coverage of road network and regression assessment, 0.5 mile is the best sampling threshold. 103 out of 190 road segments occupying 86.11% length of road network is remained. The positive Pearson correlation coefficient between GPS_ADT and 2017 AADT are found as 0.9355, $p < 0.000$. The linear regression between these two values are found to be significant with an R^2 of 0.875. The coefficient of GPS_ADT calculated from linear regression is 53.20, $p < 0.000$.

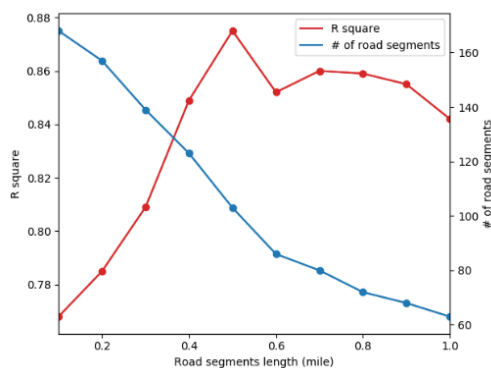


Figure 5. Threshold selection for road segment length

Fig. 6 shows the scatterplot and linear regression line between GPS_ADT and 2017 AADT. Sampling ratio of GPS recordings in actual traffic flow on highway in Marion county is around 1.8%. To evaluate the accuracy of the prediction, absolute percentage error (APE) of each road segment was calculated. The spatial distribution of Marion county highway traffic volume estimation errors in terms of APE are shown in Fig. 7. The road segments with high APE are mostly located in suburban areas where could exist a spatial bias of collecting data. This regression result reveals that one-week average daily traffic estimated from smartphone GPS data can be an effective explanatory variable for yearly traffic volume estimation and verifies that smartphone data is valuable for depicting municipal traffic mobility.

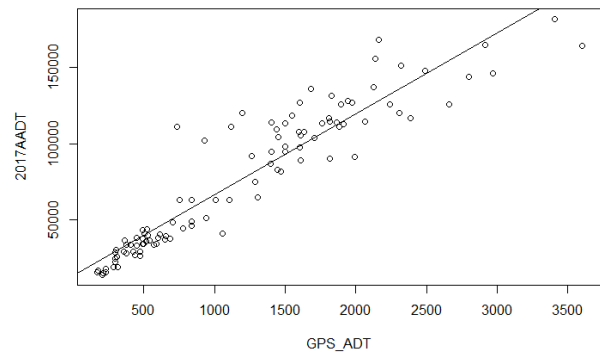


Figure 6. Scatterplot and linear regression line between GPS_ADT and AADT

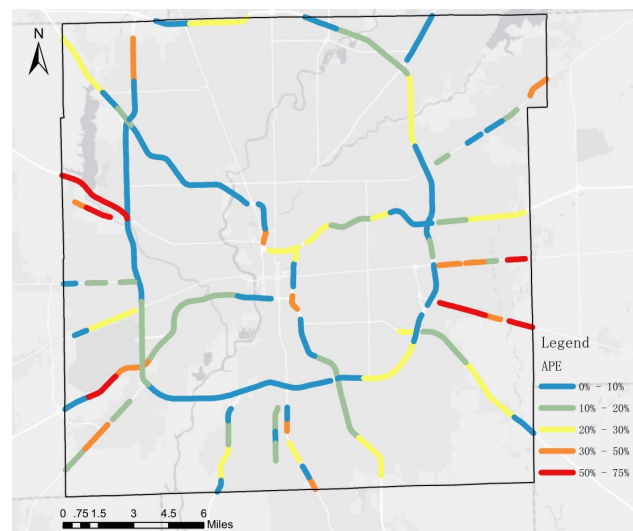


Figure 7. Spatial distribution of AADT prediction errors

But there are still some errors that can be addressed in the following three respects: 1) AADT value of traffic sections that are considered to have relatively uniform travel characteristics are generated from permanent and short-term traffic count sites, which have different counting accuracy. 2) Due to the limitation of data collection time period, traffic volume during the week of data collection could be slightly different from yearly traffic volume, which can cause sampling bias. 3) There could be deficiency on map matching method, like users recorded at road crossing area may be counted twice for both road segments.

We also evaluate our work with other independent studies. For example, Fan et al. (2019) have also proposed workflow where vehicle GPS data was used to estimate traffic volume. They used Apache Spark-based geo-computing framework to estimate Maryland AADT accurately and stably. Table 1 shows a

comparison of Fan et al. (2019) and this study. The GPS data used by Fan et al. (2019) was acquired only from vehicles, while our dataset contains unclassified smartphone users. The objective of study for Fan et al. (2019) were traffic count sites, but we used highway traffic sections for AADT estimation study. Comparing to Fan et al. (2019), our traffic volume estimation method does not need to reconstruct trips from GPS data. It gets R^2 value of linear regression between ADT and AADT as high as Fan et al. (2019), reflecting that both approaches can predict reliable traffic volume.

	Fan et al. 2019	This study
Regression objects	Maryland continuous traffic count sites	Marion county highway traffic sections
Number of cases	84	103
R Squared	0.878	0.875
Sampling ratio	1.5% - 2%	~1.8%
Number of GPS Timespan	19.8 million June, July, Feb. and Oct., 2015	5.23 million August 25 – 31, 2017
Vehicle defined Method	defined trip based	undefined point based

Table 1 Comparison of data and methods between Fan et al. (2019) and this study.

6. CONCLUSION

Human mobility study has already showed that people moved to and leave from downtown during morning rush hour and evening rush hour respectively. Heat maps are able to replicate such pattern with the change of GPS points density but lack the information of the direction of population movement. However, our flow maps generated by density difference-based method depict that people converge in the downtown area from all directions during the morning rush hour and left from the downtown area in the evening rush hour. Also, flow maps could help interpret the pattern of population movement. For example, it shows the capability of determining an increase of GPS points (users) density is caused by convergence or an increased number of people passing through. From the visualization perspective, the selection of window size W is very critical for the density difference-based method. When the value of W is too small, it will lose its effectiveness of presenting the movement pattern for the whole area.

For municipal traffic mobility, we provided a concise and effective method of estimating highway AADT using smartphone GPS data, where no trip information is generated or referred. This method has advantage over trip reconstruction from GPS data with trajectory information in terms of efficiency and accuracy. Comparing to other recent reported studies, our work showed that this traffic volume estimation method without trip information performed as well as other complex, cluster-based solutions. By using 103 road segments with actual AADT and compared with 7 days estimated average daily traffic, we found that there was a strong linear correlation between the GPS data estimated traffic and the INDOT 2017 AADT. The sampling ratio of GPS recordings in actual traffic flow on highway in Indianapolis is around 1.8%. The strong correlation between GPS data based traffic and AADT indicates that the estimating traffic volume using smartphone GPS data is promising, particularly with the growing population of smartphone users and the development of location sharing apps.

ACKNOWLEDGEMENTS

We thank SafeGraph, Inc. for providing the smartphone GPS data. The work was in part funded by Indiana Department of Transportation.

REFERENCES

- Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal Tourism Spaces in Estonia: Case Study with Mobile Positioning Data. *Science Direct*, 898-910.
- Caceres, N., Romero, L. M., Benitez, F. G., & del Castillo, J. M. (2012, September). Traffic Flow Estimation Models Using Cellular Phone Data. *IEEE Transactions On Intelligent Transportation Systems*, 13, 1430-1441.
- Castro-Neto, M., Jeong, Y., Jeong, M. K., & Han, L. D. (2009). AADT prediction using support vector regression with data-dependent parameters. *Expert Systems with Applications*, 36(2), 2979-2986.
- Chang, H. H., & Cheon, S. H. (2019). The potential use of big vehicle GPS data for estimations of annual average daily traffic for unmeasured road segments. *Transportation*, 46(3), 1011-1032.
- Chowdhury, M., Huynh, N., Khan, S. M., Khan, M. Z., Brunk, K., Torkjazi, M., ... & Shiri, S. (2019). Cost Effective Strategies for Estimating Statewide AADT (No. FHWA-SC-18-10).
- Dodge, S., Laube, P., & Weibel, R. (2012). Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science*, 1563-1588.
- Fan, J., Fu, C., Stewart, K., & Zhang, L. (2019). Using big GPS trajectory data analytics for vehicle miles traveled estimation. *Transportation research part C: emerging technologies*, 103, 298-307.
- Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 1041-1048.
- Kim, S., Jeong, S., Woo, I., Jang, Y., Maciejewski, R., & Ebert, D. S. (2017). Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *IEEE transactions on visualization and computer graphics*, 24(3), 1287-1300.
- Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W.-Y. (2008). Mining User Similarity Based on Location History. In *Proc. of GIS '08* (pp. 298-307). Santa Ana, CA: ACM Press.
- Shen, B., Liang, X., Ouyang, Y., Liu, M., Zheng, W., & Carley, K. M. (2018, July). Stepdeep: a novel spatial-temporal mobility event prediction framework based on deep neural network. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 724-733).
- Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30(5), 881-906.
- Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Sun, S., Chen, J., & Sun, J. (2019, April 4). Traffic congestion prediction based on GPS trajectory data. *International Journal of Distributed Sensor Networks*, 15(5).

Wand, M. P., & Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2), 97-116.

Wang, T., Gan, A., & Alluri, P. (2013). Estimating annual average daily traffic for local roads for highway safety analysis. *Transportation research record*, 2398(1), 60-66.

Wu, R., Luo, G., Shao, J., Tian, L., & Peng, C. (2018). Location prediction on trajectory data: A review. *Big data mining and analytics*, 1(2), 108-127.

Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., & Liu, C. (2018). Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine*, 56(3), 142-149.

Zhan, X., Zheng, Y., Yi, X., & Ukkusuri, S. V. (2017, February). Citywide Traffic Volume Estimation Using Trajectory Data. *IEEE Transactions On Knowledge and Data Engineering*, 29, 271-285.

Zhang, L., & He, X. (2013). Feasibility and advantages of estimating local road vehicle miles traveled on basis of global positioning system travel data. *Transportation research record*, 2399(1), 94-102.

Zheng, Y., Zhang, L., Xie, X., & Ma, W.-Y. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, (pp. 791-800). Madrid, Spain.