Final Report

How Mobility and Accessibility Affect Crime Rates: Insights from Mobile Device Location Data

Lei Zhang, Ph.D.
Maryland Transportation Institute
Department of Civil and Environmental Engineering
University of Maryland
Phone: 301-405-2281; Email: lei@umd.edu

Mofeng Yang
Maryland Transportation Institute
Department of Civil and Environmental Engineering
University of Maryland
Email: mofeng@umd.edu

Guangchen Zhao
Maryland Transportation Institute
Department of Civil and Environmental Engineering
University of Maryland
Email: gczhao@umd.edu

Aref Darzi
Maryland Transportation Institute
Department of Civil and Environmental Engineering
University of Maryland
Email: adarzi@umd.edu

Sepehr Ghader, Ph.D.
Maryland Transportation Institute
Department of Civil and Environmental Engineering
University of Maryland
Email: sghader@umd.edu

Date
2021

Prepared for the Urban Mobility & Equity Center, Morgan State University, CBEIS 327, 1700 E. Cold Spring Lane, Baltimore, MD 21251

ACKNOWLEDGMENT

This research was supported by the Urban Mobility & Equity Center at Morgan State University and the University Transportation Center(s) Program of the U.S. Department of Transportation.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle How Mobility and Accessibility Affect Crime Rates: Insights from Mobile Device Location Data		**5. Report Date** 11/10/2020
		6. Performing Organization Code
7. Author(s) *Include ORCID #* Lei Zhang # 0000-0002-3372-6321 Mofeng Yang # 0000-0002-0525-7978 Guangchen Zhao # 0000-0002-0803-207X Aref Darzi # 0000-0003-2558-5570 Sepehr Ghader # 0000-0003-1938-7914		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Maryland Transportation Institute, Department of Civil and Environmental Engineering, University of Maryland College Park College Park, MD 20742		**10. Work Unit No.**
		11. Contract or Grant No. 69A43551747123
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary-Research UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Oct 2019 – Jan 2021
		14. Sponsoring Agency Code
15. Supplementary Notes		
16. Abstract This research study investigates the possible correlations between mobility, accessibility, and crime rate. A rich mobile device location dataset including detailed anonymized location traces of the mobile devices observed in the City of Baltimore was combined with the police arrest records to study how mobility and accessibility affect neighborhood safety. The research team first processed and analyzed the mobile device location dataset to obtain measures of mobility and accessibility. These measures differed from the traditional measures in that they were obtained based on the empirically observed location data. The research team then built statistical and machine learning tools to model crime rates at the census-tract levels, using the calculated mobility and accessibility measures, land-use variables, and socioeconomic-related variables as the covariates. Subsequently, the team focused on the correlation of the crime rates with the mobility and accessibility variables. Results indicated that the mobility and accessibility measures can help improve the performance of crime rate prediction. Also, non-motorized travel might be positively related to burglary. The study seeks to inform decision-makers about the transportation-related issues contributing to the lack of safety and offer transportation solutions to crime-related problems, especially in the neighborhoods suffering from high crime rates.		
17. Key Words : Mobile Device Location Data; Mobility; Accessibility; High-Crime Neighborhoods		**18. Distribution Statement**

19. Security Classif. (of this report) :	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	23	

How Mobility and Accessibility Affect Crime Rates: Insights from Mobile Device Location Data

Project PI: Dr. Lei Zhang
Herbert Rabin Distinguished Professor of Civil Engineering
Director, Maryland Transportation Institute, University of Maryland (UMD)
Phone: (301)405-2881; Email: lei@umd.edu

Project Team: Mofeng Yang, Guangchen Zhao, Aref Darzi, Sepehr Ghader.

ABSTRACT

This research study investigates the possible correlations between mobility, accessibility, and crime rate. A rich mobile device location dataset including detailed anonymized location traces of the mobile devices observed in the City of Baltimore was combined with the police arrest records to study how mobility and accessibility affect neighborhood safety. The research team first processed and analyzed the mobile device location dataset to obtain measures of mobility and accessibility. These measures differed from the traditional measures in that they were obtained based on the empirically observed location data. The research team then built statistical and machine learning tools to model crime rates at the census-tract levels, using the calculated mobility and accessibility measures, land-use variables, and socioeconomic-related variables as the covariates. Subsequently, the team focused on the correlation of the crime rates with the mobility and accessibility variables. Results indicated that the mobility and accessibility measures can help improve the performance of crime rate prediction. Also, non-motorized travel might be positively related to burglary. The study seeks to inform decision-makers about the transportation-related issues contributing to the lack of safety and offer transportation solutions to crime-related problems, especially in the neighborhoods suffering from high crime rates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Crime has been one of the biggest social problems in the United States, with a sharp increase after the 1900s and decreasing since the early 1990s (Truman and Planty, 2012). In 2019, the Federal Bureau of Investigation (FBI) reported a total of 2,109.9 property crimes per 100,000 people and 379.4 violent crimes per 100,000 people in the U.S. (Gramlich, 2020). To reduce or even potentially prevent crimes from happening, previous studies have analyzed the occurrence of crime with the support of routine activities theories, which include three essential elements: a motivated offender, an attractive target, and the absence of capable guardianship (Cohen and Felson, 1979). Though proven by empirical evidence, these approaches cannot differentiate between how local characteristics – such as land use and socio-demographics – and, more importantly, the population-level human activities influence crime patterns (Caminha, 2017). With the technological advancement in mobile sensors and mobile networks, mobile device location data (MDLD) has been growing drastically in terms of data coverage and data size, which makes it possible for researchers to look at crime from another perspective.

This report summarizes the study of analyzing how mobility and accessibility affect crime rates in Baltimore City. The mobility and accessibility are estimated from a large-scale MDLD dataset including detailed anonymized location traces of the mobile devices observed in Baltimore, including the multimodal mobility patterns and accessibility measures. These measures differ from the traditional measures in that they were obtained based on the empirically observed location data that is able to depict individual-level mobility. To estimate the crime rate, the open-source crime records data provided by the Baltimore Policy Department (BPD) were collected. The research team has built a data-driven modeling framework to model crime rates at the census-tract level, using the calculated mobility and accessibility measures, land-use variables, and economy-related variables as the covariates. The relationship between these covariates and the crime rate is discussed. Results indicated that the mobility and accessibility measures can help improve the performance of crime rate prediction. Also, non-motorized travel might be positively related to burglary. The final result can inform decision-makers about the transportation-related issues contributing to the lack of safety and offer transportation solutions to crime-related problems, especially in the neighborhoods suffering from high crime rates. The rest of the report is organized as follows. In section two, we review the literature on the application of mobile device location data in the realm of transportation and the relationship between crime rate and transportation-related variables. In section three, we introduce the methodology of deriving mobility and accessibility measures from the mobile device location data, and the data-driven modeling framework of high-crime neighborhoods with Geographically Weighted Regression (GWR). In section four, we describe the data used in this study, including the MDLD and the mobility and accessibility measures derived from it, the Baltimore crime record, and the American Community Survey (ACS). In section five, we use the data-driven approach results to discuss how these variables affect the Burglary crime rate across different regions in the City of Baltimore. Section six offers concluding remarks and provides policy indications.

2. LITERATURE REVIEW

Previous research reveals that transportation-related variables have contributed to various types of crimes, such as rail stations (Cahill and Mulligan, 2007) and interstate highways (Marton, 1995).

By considering the location of these transportation components, the correlation of certain types of crimes (e.g., Burglary) and the transportation components can be quantified. In recent years, along with the technological advancements in the internet, mobile sensors, and mobile networks, the movement data has also been brought into the criminology domain to support crime rate analysis. For instance, Wang et al. 2016 leveraged the taxi flow and Point-Of-Interest (POI) data to infer the crime rate in the Chicago, Illinois, USA. The taxi trip records were aggregated at the community areas to capture the socio interactions among various community areas. The results suggested that by considering the POI and taxi flows, the crime rate inference error can be reduced by 17.6% (Wang et al., 2016). Zhao and Tang built a framework that captures temporal-spatial correlations for crime prediction (Zhao and Tang, 2017). They collected the check-ins from the POI dataset and pick-up and drop-off points from the taxi trajectories dataset to represent human mobility. Hanaoka (2018) examined the relationships between the occurrences of snatch-and-run offenses and hourly population estimated from approximately 500,000-700,000 mobile phone users. The time of day relationship between the snatch-and-run offenses and the population density was discussed. However, either the check-ins data or the taxi flow data are not able to capture the entire mobility patterns without an appropriate computational algorithm to expand these samples to population-level statistics. Caminha et al. (2017) estimated the floating population to represent human mobility for each census tract (number of people who pass through a census tract) and used the City Clustering Algorithm (CCA) to measure the relationship with crime rates. The result suggested that a disproportional number of property crimes occurs in regions where an increased flow of people occurs in the city of Fortaleza, Brazil (Caminha et al., 2017). Even though the floating population can represent the population-level human movement, it does not separate the movement by travel modes, such as vehicle, transit, and non-motorized travel, which each have distinct characteristics.

The literature review suggests that limited studies have been able to estimate the population-level human mobility in order to support crime rate analysis. This report is among the first to leverage the rich MDLD dataset and Location-based Service (LBS) data, including detailed anonymized location traces of the mobile devices observed in Baltimore, and combines that with the police arrest records to study how mobility and accessibility affect neighborhood safety. The Location-based Service (LBS) data is generated when a mobile application updates the device's location with its most accurate sources, based on the currently available location-providing technologies such as Wi-Fi, Bluetooth, cellular tower, and GPS (Gonzalez et al., 2008; Wang et al., 2019). The LBS data can reflect the exact location of the device, providing invaluable location information describing population-level mobility patterns. Lots of applications have been developed using the LBS data. For instance, the Maryland Transportation Institute (MTI) at the University of Maryland (UMD) developed the COVID-19 Impact Analysis Platform (https://data.covid.umd.edu/) to provide insight into COVID-19's impact on mobility, health, economy, and society across the U.S. (Zhang et al., 2020; Zhang et al., 2020; Xiong et al., 2020; Xiong et al., 2020; Hu et al., 2021).

The research team first processes and analyzes the mobile device location dataset to obtain measures of mobility and accessibility. These differ from the traditional measures in that they will be obtained based on the empirically observed location data. The research team then builds statistical and machine learning tools to model crime rates at the census-tract levels, using the calculated mobility and accessibility measures, land-use variables, and economy-related variables as the covariates. Subsequently, the team focuses on the correlation of the crime rates with the mobility and accessibility variables. The study seeks to inform the decision-makers about the transportation-related issues contributing to the lack of safety and offer transportation solutions to crime-related problems, especially in the neighborhoods suffering from high crime rates.

3. METHODOLOGY

3.1. Deriving Multimodal Mobility Pattern and Accessibility Measures from Mobile Device Location Data

In this section, we briefly discuss the methodologies to derive multimodal mobility patterns and accessibility measures with MDLD. A suite of computation algorithms that includes home and work location imputation, weighting, trip identification, and travel mode identification is introduced (Zhang et al., 2020).

3.1.1. Home and Work Location Identification and Weighting

Identiying the home and work locations is crucial to understand the travel pattern of each sample in the MDLD and a necessary step to expand the sample to population-level statistics. In this project, we use the Density-based spatial clustering of applications with noise (DBSCAN) to identify the location of residence and destinations for employment (if exists) for all individuals included in the MDLD (Ester et al., 1996), as shown in **Figure 1**. For home identification, we use every day of the study period, from 7 PM to 7 AM the next day. For work location identification, we use every working day in the study period, from 11 AM to 5 PM. For home location, among all the clusters obtained from the DBSCAN algorithm for each device ID, the center of the cluster with the most observation points is identified as its home location. If a device does not have any cluster, the device is entirely removed from the sample. For the work location, the cluster with the most observation points during the assumed work hours may still be the home location, as some people might work with a different schedule or just work from home. Therefore, in order to identify a fixed location, we first identify if the device has one or more clusters during the daytime which are at least 500 meters away from its identified home location. Among them, the center of the cluster with the most observation points is identified as the device's work location (Alexander et al., 2015).

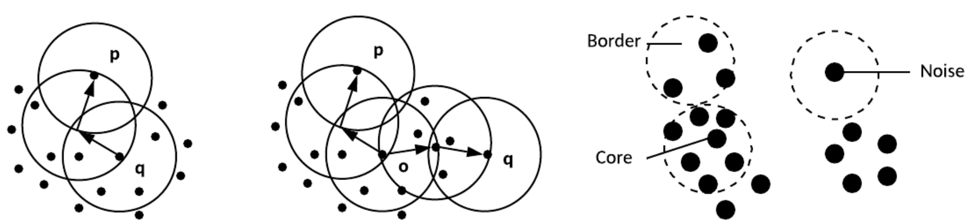


Figure 1. Illustration of DBSCAN algorithm.

After identifying the home locations from the MDLD, the spatial join is performed to match each device's home location to a census tract. Then, the weight of each mobile device can be calculated with the census tract population, as shown below:

$$w_i = \frac{pop_i}{n_i}$$

where w_i represents the weight for devices with home location at census tract i, pop_i represents the population at census tract i, and the n_i represents the number of mobile devices with home location at census tract i. With the weight of each mobile device generated, the sample can be expanded to the population by multiplying the measures with the weight.

3.1.2. Trip Identification

Trips are the unit of analysis in our study. Mobile device location data does not include trip information. Location observations are continuously being generated while the device moves, stops, stays static, or starts a new trip. As a result, we developed a trip identification algorithm that can detect which location observations form a trip together. We take the following steps to identify trips. The algorithm runs on the observations of each device separately.

1. Pre-Processing: We first sort device observations by time. The algorithm assigns a random ID to each trip it identifies. Many location points in the dataset may not belong to trips. The algorithm assigns "0" to the trip ID of these locations to tag them as static points. For every location point, we calculate distance, time, and speed between the point and its immediate previous and next points, if they exist. Three hyperparameters need to be set for the algorithm: distance threshold (300 meters), time threshold (5 minutes), and speed threshold (1.4 meters per second). The speed threshold is used to identify if a location point is recorded on the move. The distance and time thresholds are used to identify stay locations and trip ends. At this step, the algorithm identifies the device's first observation with $speed\ from \geq speed\ threshold$. This identified location point is recorded on the move, so a hashed trip ID is generated and assigned to this point. All points recorded before this point, if they exist, are set to have "0" as their trip ID. Next, a recursive algorithm identifies if the next points are on the same trip and should have the same trip ID.

2. Iterative Algorithm. This algorithm checks every point to identify whether they belong to the same trip as their previous point (**Figure 2**). If they do, they are assigned the same trip ID. If they do not, they are either assigned a new hashed trip id (when their $speed\ from \geq speed\ threshold$) or their trip ID is set to "0" (when their $speed\ from < speed\ threshold$). Identifying whether a point belongs to the same trip as its previous point is based on the point's "speed to," "distance to" and "time to" attributes. If a device is seen in a point with $distance\ to \geq distance\ threshold$ but is not observed to move there ($speed\ to < speed\ threshold$), the point does not belong to the same trip as its previous point. When the device is on the move at a point ($speed\ to \geq speed\ threshold$), the point belongs to the same trip as its previous point, but when the device stops, the algorithm checks the radius and dwell time to identify if the previous trip has ended. If the device stays at the stop (points should be closer than the distance threshold) for a period of time shorter than the time threshold, the points still belong to the previous trip. When the dwell time reaches above the time threshold, the trip ends, and the next points no longer belong to the same trip. The algorithm does this by updating "time from" to be measured

from the first observation in the stop, not the point's previous point. The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips that are shorter than 300 meters are removed.

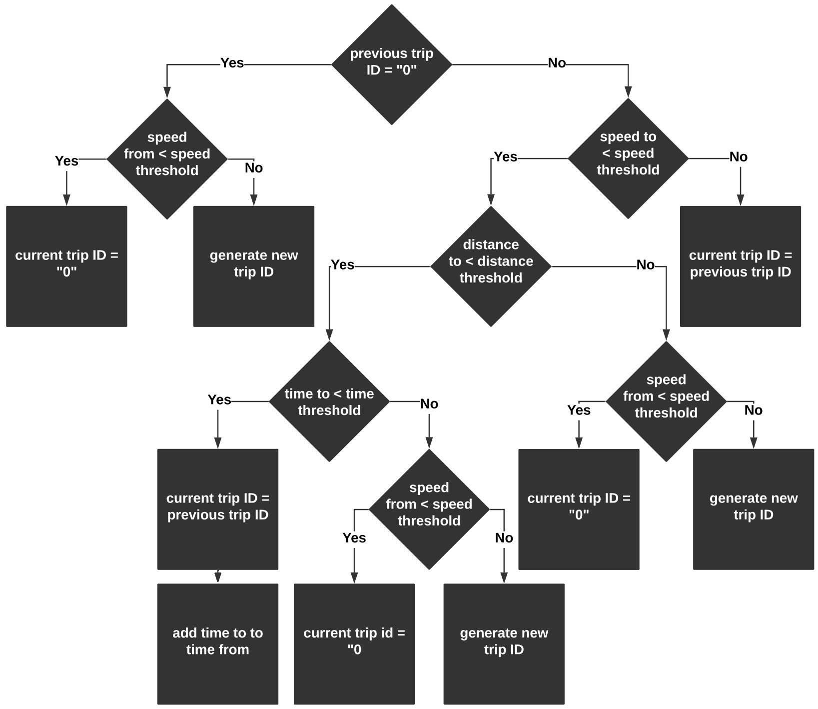


Figure 2. Recursive algorithm for trip identification.

3.1.3. Travel Mode Imputation

Machine learning models are used to impute five travel modes (drive, bus, rail, bike and walk) from trips identified from the previous step. The five modes are further combined into three modes: vehicle (drive), transit (bus and rail), and non-motorized (walk and bike). Five machine learning models are examined in terms of prediction accuracy, including K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), eXtreme Gradient Boosting (XGB), Random Forest (RF), and Deep Neural Network (DNN). Feature set construction directly affects the model performances. Three types of features – sample rate, trip, and multimodal transportation network – are constructed from the MDLD, as shown in **Table 1**. The sample rate feature, represented by the average number of records per minute, indicates the location service usage during a trip. The trip features can show the characteristics of each trip, including trip distance, origin-destination distance, trip time, average speed, minimum speed, maximum speed, median speed, and 5th, 25th, 75th, 95th percentile speeds. The multimodal transportation network features are important to distinguish between different travel modes (Bohte and Maat, 2009; Gong et al., 2018). The multimodal transportation network data is collected including drive, bus, rail networks, and bus stop locations to construct network-related features. The drive network is collected from the Highway Performance Monitoring System (HPMS) that includes national freeway and arterial roads in the U.S. The national bus and rail network and the bus stops data are collected from the United States Department of Transportation (U.S. DOT) Bureau of Transportation Statistics (BTS)

National Transit Map (NTM). **Figure 3** illustrates the multimodal transportation networks used in this study. Here, the min-max, median, and 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile distances to the rail and bus network are calculated. A 50-meter buffer is generated for all bus stops to obtain the percentage of location points for each trip that falls within the buffer respectively.

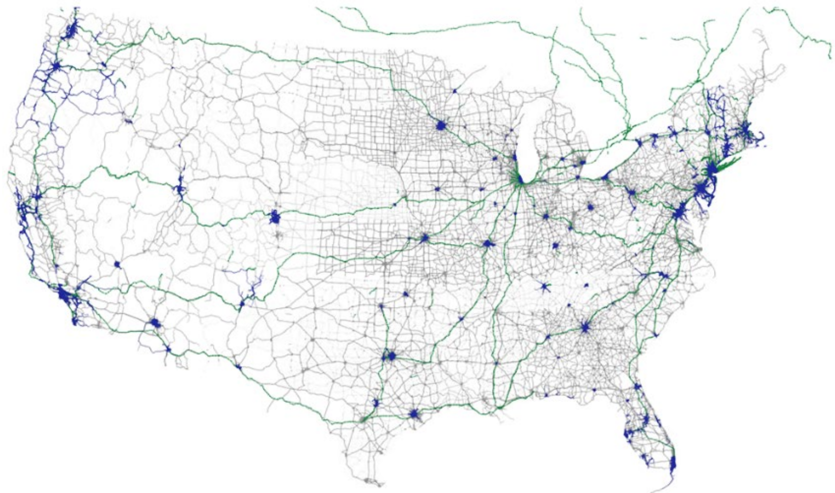


Figure 3. Multimodal transportation networks: drive (grey), rail (green), bus (blue).

Table 1. Features constructed from mobile device location data.

Features	Unit
Sample Rate Feature	
Average # of records per minutes	number / minute
Trip Features	
Trip distance	meters
Origin-destination distance	meters
Trip time	minutes
Min, max, median, 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile speeds	meters / second
Multimodal Transportation Network Features	
Min, max, median, 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile distances to the rail network	meters
Min, max, median, 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile distances to the bus network	meters
Percentage of records within 50-meter of bus stops	meters

3.1.4. Accessibility Measures

For each census tract, we calculate the average observed travel time for all "home-based work" trips of the census tract residents as our data-driven measure of accessibility to jobs. Similarly, we calculate the average observed travel time for all "home-based food" and "home-based healthcare" trips of the census tract residents as the data-driven measures of accessibility to food and healthcare. Therefore, we take the following steps:

1. For each device, find the average trip duration of home-based trips for job/food/healthcare, as its representative travel time to job/food/healthcare.
2. For each census tract, find the average travel time to job/food/healthcare of all the devices that are imputed to reside there as the census tract's accessibility to job/food/healthcare.

3.2. Data-Driven Modeling Framework of High-Crime Neighborhoods

In this section, we introduce the data-driving modeling framework of high-crime neighborhoods. This framework will apply the Geographically Weighted Regression (GWR) to estimate the burglary crime rates with the calculated mobility and accessibility measures, land-use variables, and economy-related variables as the covariates.

3.2.1. Multicollinearity

Before we estimate the model, the multicollinearity of all the explanatory variables needs to be examined. Multicollinearity refers to the situation in which the explanatory variables are highly correlated with each other, which might cause biased estimates of the regression model (Farrar and Glauber, 1967). To solve the multicollinearity issue, we calculated the Variable Inflation Factor (VIF) for each explanatory variable (Neter and Kutner, 1989). VIF can effectively quantify the multicollinearity by estimating the linear relationship between one explanatory variable and the others. For each explanatory variable, the VIF can be calculated with the following equation:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where VIF_i represent the VIF value for explanatory variable i, R_i^2 represents the coefficient of determination for explanatory variable *i*. If the VIF value is greater than 10, the variable is assumed to be a multicollinear variable and should be removed from the regression model.

3.2.2. Geographically Weighted Regression

The Geographically Weighted Regression (GWR) extends the generalized linear regression by incorporating the spatial heterogeneity when modeling the relationship between the response variable and the explanatory variables. In the generalized linear regression model, the estimated coefficients of variables are constant across the whole sample dataset, while the GWR model allows different relationships to exist at different locations over space (Brunsdon & Fortheringham, 1996; Brunsdon & Fortheringham, 1998). The GWR model can be generally represented by the following equation:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{n} \beta_{ik}(u_i, v_i) x_{ik} + \varepsilon_i$$

where the vector representing the weight for location *i*, $\hat{\beta}(i) = (\beta_{i0}, \beta_{i1}, \dots, \beta_{in})^T$, can be estimated with:

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i) X]^{-1} X^T W(u_i, v_i) Y$$

where X represents the $n \times n{+}1$ matrix of explanatory variables, Y represents the $n \times 1$ matrix of response variable, $W(u_i, v_i)$ is an $n \times n$ matrix whose diagonal elements denote the geographical weighting of data at location i. In GWR, $W(u_i, v_i)$ can be interpreted as a distance decay function, where $w_j(u_i, v_i)$ represents the weighted distance between location j and i. There are two commonly used kernel functions for the distance decay function, namely Gaussian and the bi-square kernel:

$$Gaussian: w_j(u_i, v_i) = exp\left[-\left(\frac{d_{ij}}{b}\right)^2\right]$$

$$Bi-Square: w_j(u_i, v_i) = \begin{cases} \left[1-\left(\frac{d_{ij}}{b}\right)^2\right], if\ d_{ij} \le b \\ 0, if\ d_{ij} > b \end{cases}, j = 1,2, \dots, n$$

where d_{ij} represents the distance between location i to location j, and b is the bandwidth. In this study, the adaptive bi-square is selected as the kernel function for the GWR model. The adaptive bi-square kernel function considers the bandwidth as the number of nearest-neighbors at which data is weighed to exactly zero and further observations do not influence each local regression (Oshan et al., 2020). The optimal bandwidth was selected by minimizing the corrected Akaike Information Criterion (AICc) in order to balance between model variance and bias (Sakamoto et al., 1986; Yu et al., 2020).

4. DATA

4.1. Response Variable: Burglary Crime Rate

The response variable used in this project is the Burglary crime rate of 2019 for each census tract in Baltimore City. The crime record data is obtained from the Open Baltimore website "BPD Part 1 Victim Based Crime Data" (https://data.baltimorecity.gov/datasets/part1-crime-data?geometry=-89.408%2C37.108%2C-68.479%2C40.112). The data is collected and updated by the Baltimore Policy Department (BPD) every Monday with a nine-day time lag, including the geocoded to the approximate latitude/longitude location of the incident and excluding those records for which an address could not be geocoded.

Figure 4 (a) summarizes the overall data quality and the trends of the data. This project will focus on Burglary crime. We further calculate the burglary crime rate:

$$CR_i = \frac{100{,}000 \cdot C_i}{pop_i}$$

where CR_i represents the burglary crime rate at census tract i, the C_i represents the number of burglaries at census tract i, and pop_i represents the population at census tract i.

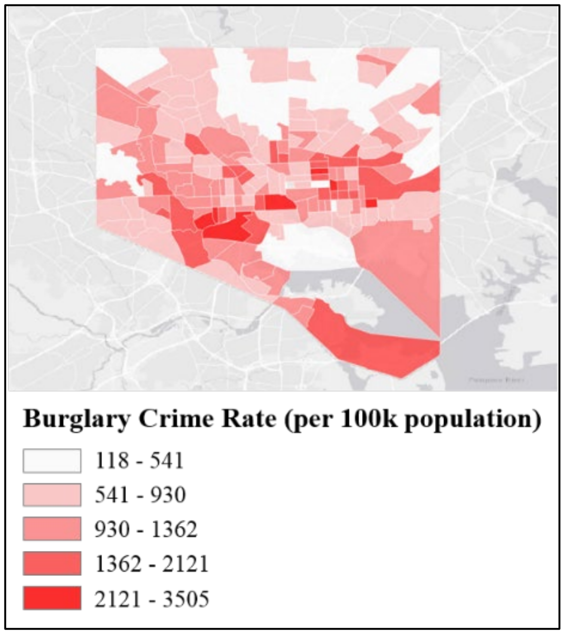


Figure 4. Baltimore police department victim-based crime data: burglary crime rate per 100k population.

4.2. Explanatory Variables

4.2.1. Multimodal Mobility Patterns

We used one month of mobile device location data for January 2019 covering the Baltimore Metropolitan Area. The data is obtained from different leading data vendors that collect anonymized mobile device location data with over 100 million monthly active samples. For this study, only the national samples with imputed home locations that have been observed in Baltimore City in January 2019 have been used, which includes a total of around 350,000 unique monthly samples. The raw mobile device location data includes an anonymized device ID, latitude, longitude, and timestamp for each location sighting at a high level of location accuracy (less than 30 feet). Sample rows of the data are shown in **Table 2**. We then applied the computation algorithms to identify home and work locations, generate weights, identify trips, and impute the travel mode in order to derive the multimodal travel patterns.

Table 2. Sample rows of the mobile device location data.

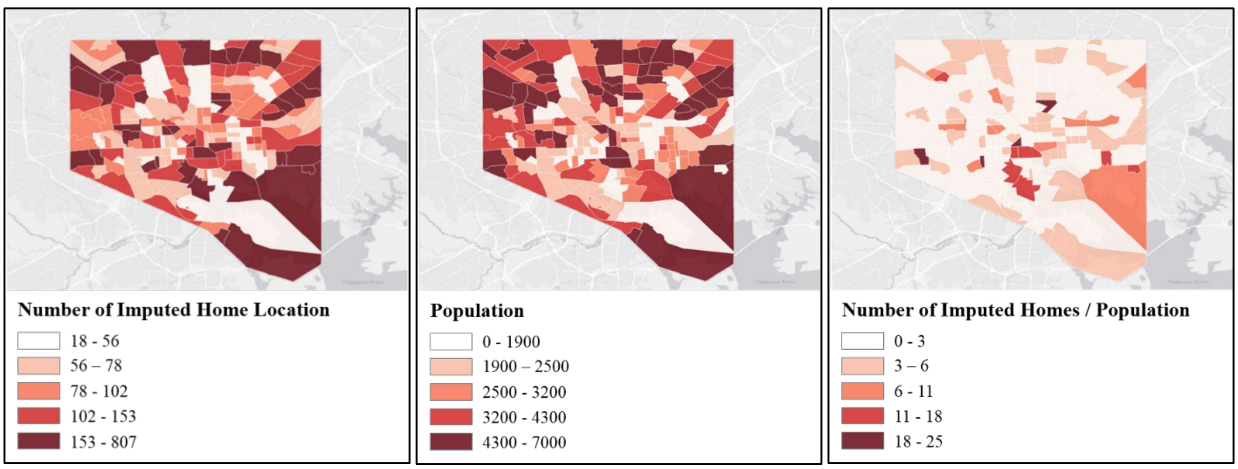
Device ID	Latitude	Longitude	Timestamp	Accuracy
12dsu85dbxxx	39.65441	-76.51234	1548438483	50
12dsu85dbxxx	39.65432	-76.51278	1548438682	69
2as6axf12123y	39.32136	-76.71222	1548425142	13
2as6axf12123y	39.32162	-76.70684	1548427307	5

Figure 5 shows the home location identification result. The overall pattern of the number of identified homes shows a distribution similar to the population distribution. A relatively higher penetration ratio (number of imputed homes/population) can be observed around the Inner Harbor area. **Figure 6** shows the multimodal mobility patterns by visualizing the location traces of different travel modes. The vehicle travel (in orange) covers the entire region of Baltimore City,

where all the major arterials and the Baltimore Beltway can be observed. The non-motorized travel (in green) is concentrated mainly in the Inner Harbor region and drops in the suburbs. The transit travel (in blue), including bus and rail, is highly aligned with the light rail and bus lines. To incorporate these multimodal mobility patterns into the census tract-level characteristics, we did the spatial join for the origins and destinations of the multimodal trip rosters and then aggregated them by the destination:

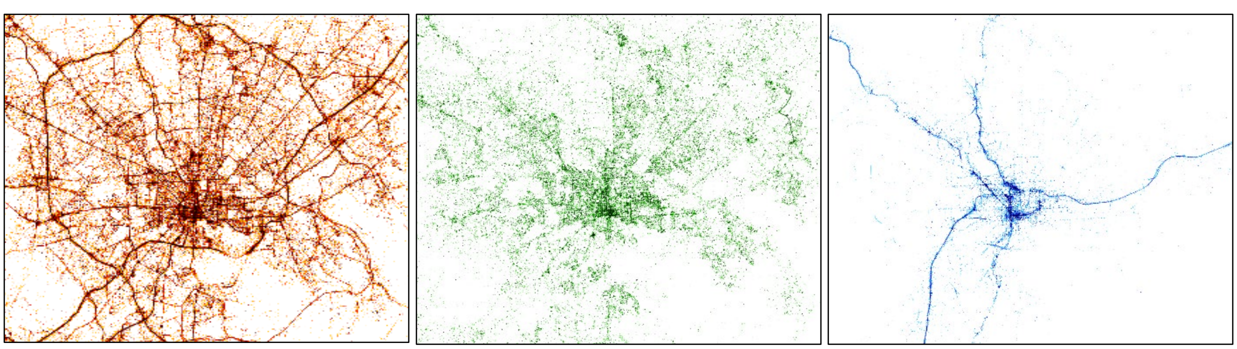
$$A_{i,m} = \frac{\sum_{j}^{N_i} w_j I_{j,m}}{pop_i}$$

where $A_{i,m}$ represents the total weighted trips attracted per person to census tract i for travel mode m, N represents the total number of trips observed ended in census tract i, w_j is the weight for each trip j (for trips conducted by the same person, they have the same weight), and $I_{j,m}$ is the indicator function to check whether the travel mode of this trip j is m. **Figure 7** shows the aggregation results.



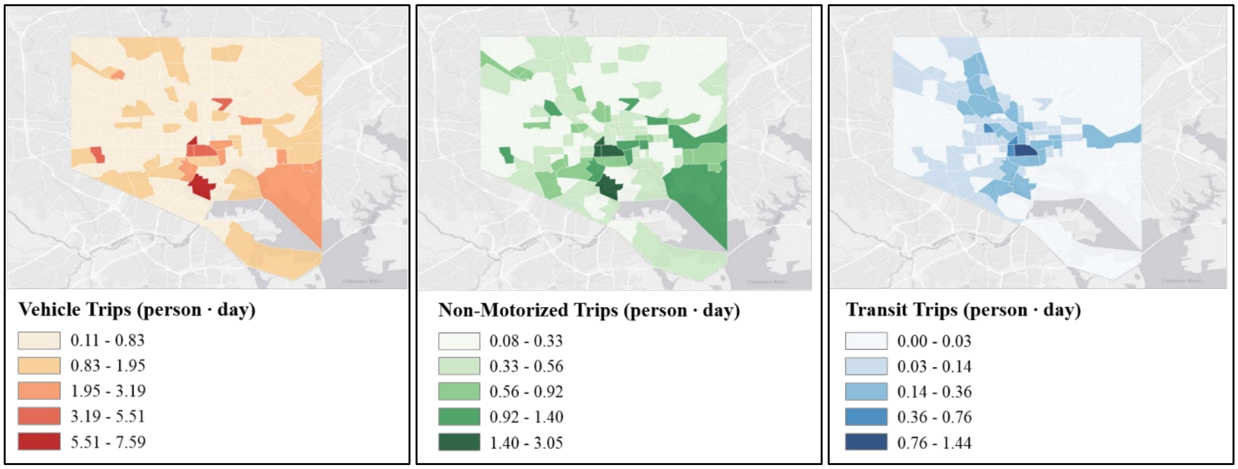
(a) (b) (c)

Figure 5. Home imputation results. (a) number of imputed home location; (b) population; (c) number of imputed homes / population (in percentage).



(a) (b) (c)

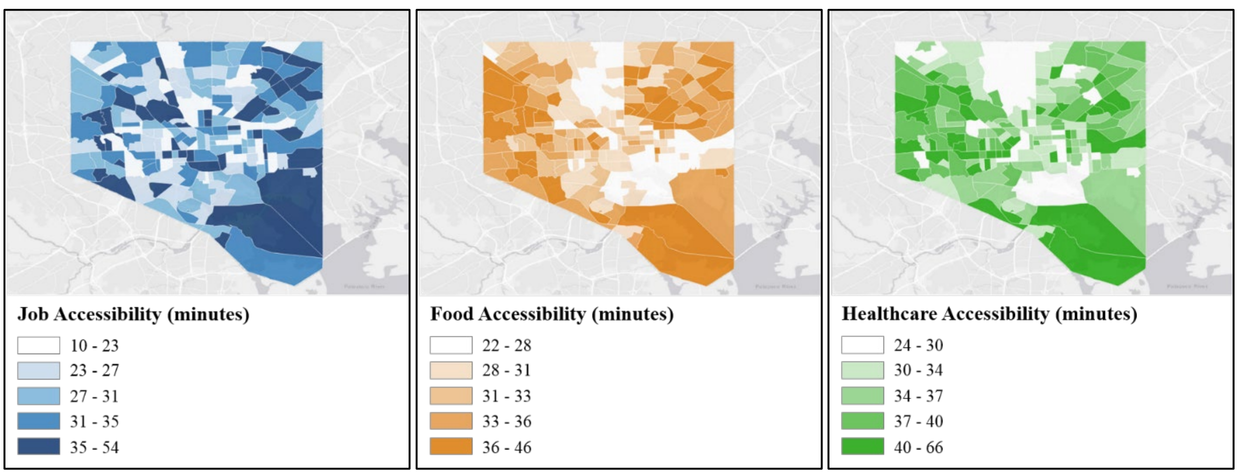
Figure 6. Multimodal mobility patterns in Baltimore City: (a) vehicle; (2) non-motorized (walk and bike); (3) transit (bus and rail).



(a) (b) (c)

Figure 7. Daily average number of person trips attracted to census tracts in Baltimore City: (a) vehicle; (2) non-motorized (walk and bike); (3) transit (bus and rail).

4.2.2. Accessibility Measures



(a) (b) (c)

Figure 8. Accessibility measures in Baltimore City. (a) job; (b) food; (c) healthcare.

Figure 8 shows the accessibility measures calculated from the MDLD. As shown in **Figure 8 (a)**, the job accessibility varies across the region and no significant clusters are observed. A low value of job accessibility does not mean the corresponding census tract is more affluent. In contrast, the high-income population might be able to access more jobs that are far from their home locations, while the low-income population has limited choices. **Figure 8 (b)** and **(c)** show the food and healthcare accessibility, respectively. It can be clearly observed that north Baltimore has a lower value of food and healthcare accessibility, as compared to west Baltimore. This corresponds to the fact that north Baltimore has a much larger high-income population than west Baltimore, and the high-income population has better access to food and job resources.

4.2.3. American Community Survey

The land-use variables and economy-related variables are collected from the 2014-2018 5-year American Community Survey (ACS) published by the U.S. Census Bureau. The description of all variables used in this project is summarized in **Table 3**. The descriptive statistics of all variables are summarized in **Table 4**.

Table 3. Description of all variables used in this study.

Variables	Description
Response Variables	
Burglary	Number of Burglary per 100k population of the given census tract in 2019.
Multimodal mobility patterns	
Vehicle	Daily number of vehicle trips attracted divided by census tract population.
NonMotor	Daily number of non-motorized trips attracted divided by census tract population
Transit	Daily number of transit trips attracted divided by census tract population
Accessibility Measures	
avgt2work	Average travel time to work for the given census tract.
avgt2food	Average travel time to food for the given census tract.
avgt2healthcare	Average travel time to healthcare service for the given census tract.
American Community Survey	
med_age	Median age of the given census tract.
med_inc	Median household annual income (in dollar) of the given census tract.
1524_perc	Percentage of population with age between 15 and 24 of the given census tract.
unemployed_perc	Percentage of unemployed population of the given census tract.
mvd5yr_perc	Percentage of the population moved within 5 years of the given census tract.
gpq_perc	Percentage of group quarter population of the given census tract.
black_perc	Percentage of African American population of the given census tract.
hisp_perc	Percentage of Hispanic population of the given census tract.
college_perc	Percentage of population with bachelor or higher degree of the given census tract.
hschool_perc	Percentage of population with high school degree of the given census tract.
huvac_perc	Percentage of vacant housing units of the given census tract.
avg_veh	Average number of vehicles per household of the given census tract.
sphh_perc	Percentage of single-parent households of the given census tract.
pov_perc	Percentage of population under 100% of the poverty level of the given census tract.
snap_perc	Percentage of households receiving food stamps of the given census tract.

Table 4. Descriptive statistics of all variables used in this study.

Variables	Obs.	Mean	Std.Dev.	Min.	Q1	Med.	Q3	Max.
Response Variables								
Burglary	198	962.650	553.490	117.647	608.076	883.227	1211.964	3504.673
Multimodal mobility patterns								
Vehicle	198	0.819	1.037	0.112	0.335	0.458	0.794	7.592
NonMotor	198	0.429	0.392	0.084	0.226	0.299	0.461	3.048
Transit	198	0.059	0.140	0.001	0.004	0.014	0.048	1.439
Accessibility Measures								
avgt2work	198	29.557	7.717	10.383	24.953	28.835	34.017	54.376
avgt2food	198	32.008	4.615	21.520	28.781	32.219	35.072	45.920
avgt2healthcare	198	35.707	6.493	23.537	31.142	35.626	39.652	66.471
American Community Survey								
med_age	198	36.321	6.563	17.800	31.900	35.500	40.275	64.000
med_inc	198	51646	29545	13074	33763	42251	60437	213200
1524_perc	198	0.126	0.085	0.013	0.089	0.115	0.141	0.741
unemployed_perc	198	0.550	0.132	0.145	0.489	0.571	0.643	0.796
mvd5yr_perc	198	0.184	0.090	0.012	0.118	0.173	0.236	0.494

gpq_perc	198	0.029	0.090	0.000	0.000	0.003	0.017	0.743
black_perc	198	0.627	0.337	0.004	0.330	0.764	0.930	0.994
hisp_perc	198	0.050	0.073	0.000	0.011	0.029	0.059	0.449
college_perc	198	0.217	0.189	0.010	0.070	0.147	0.298	0.726
hschool_perc	198	0.586	0.124	0.206	0.503	0.581	0.671	0.951
huvac_perc	198	0.201	0.120	0.031	0.110	0.166	0.274	0.608
avg_veh	198	1.061	0.360	0.229	0.783	1.044	1.346	2.034
sphh_perc	198	0.134	0.095	0.000	0.062	0.128	0.187	0.560
pov_perc	198	0.226	0.130	0.008	0.115	0.201	0.323	0.601
snap_perc	198	0.272	0.165	0.000	0.140	0.261	0.394	0.702

5. Results

5.1. Multicollinearity

The left column in **Table 5** shows the VIF values for all explanatory variables. It can be seen that some of the variables have high VIF values that should be considered as multicollinear variables. After carefully removing these variables, as shown in the right column in **Table 5**, the VIF values for all the variables are below 10, indicating the multicollinearity problem is resolved. These variables include NonMotor, Transit, avgt2work, mvd5yr_perc, gpq_perc, hisp_perc, college_perc, huvac_perc, sphh_perc and pop_dens.

Table 5. Variable inflation factor of explanatory variables.

Variables	Before Variable Inflation Factor	Variables	After Variable Inflation Factor
Vehicle	11.93072	NonMotor	5.62210
NonMotor	26.5437	Transit	2.65959
Transit	3.46095	avgt2work	8.25586
avgt2work	18.86203	mvd5yr_perc	8.66353
avgt2food	105.4974	gpq_perc	1.24776
avgt2healthcare	52.98927	hisp_perc	1.63469
med_age	130.3267	college_perc	4.79397
med_inc	25.20029	huvac_perc	4.87947
1524_perc	14.58747	sphh_perc	4.82582
unemployed_perc	114.9077	pop_dens	5.54659
mvd5yr_perc	10.48979		
gpq_perc	5.15734		
black_perc	40.4317		
hisp_perc	3.29156		
college_perc	40.39105		
hschool_perc	241.1167		
huvac_perc	7.43132		
avg_veh	61.463		
sphh_perc	11.92838		
pov_perc	23.0304		
snap_perc	35.16363		
pop_dens	7.59319		

5.2. Ordinary Least Squares Regression Results

After removing the multicollinear variables, we ran the OLS model to examine the global relationship between the explanatory variables and the response variable. **Table 6** shows the OLS model estimation result. Five out of 10 variables are considered to have a significant impact on the burglary crime rate, where NonMotor, hisp_perc, and huvac_perc show a positive relationship and gpq_perc and college_perc show a negative relationship. These relationships are relatively stationary, indicating that these variables are contributing to the burglary crime rate across all of Baltimore City.

Table 6. OLS result for burglary crime rate.

Variables	Coefficient	Std.Dev.	Z-value	P-value
Intercept	0.000	0.052	0.000	1.000
NonMotor	0.196	0.085	2.314	0.021 **
Transit	0.156	0.081	1.939	0.053
avgt2work	0.005	0.053	0.086	0.932
mvd5yr_perc	0.110	0.067	1.638	0.102
gpq_perc	-0.170	0.056	-3.043	0.002 ***
hisp_perc	0.251	0.055	4.592	0.000 ***
college_perc	-0.200	0.085	-2.339	0.019 **
huvac_perc	0.468	0.063	7.490	0.000 ***
sphh_perc	-0.100	0.073	-1.376	0.169
pop_dens	0.043	0.062	0.704	0.481

5.3. GWR Results

The variables used in the OLS are also used in the GWR model to analyze the spatial heterogeneity of these variables. In this study, we utilized the open-source Python Spatial Analysis Library (PySAL) (Ray and Anselin, 2010) and mgwr (Oshan et al., 2019) to estimate the GWR model. **Table 7** shows the model estimation results. **Figure 9** shows the R-Square, prediction results, and the residuals of the GWR model.

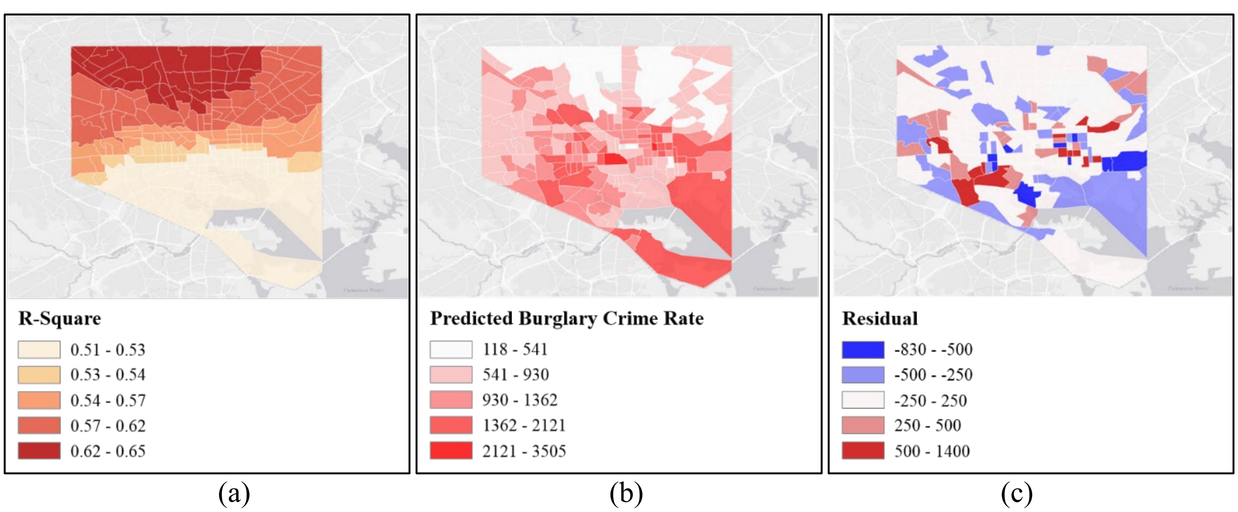


Figure 9. GWR model estimation result. (a) R-Square; (b) prediction; (c) residual.

Table 7. GWR result for burglary crime rate.

Variables	Mean	Std.Dev.	Min.	Median	Max.
Intercept	0.058	0.066	-0.109	0.049	0.182
NonMotor	0.260	0.061	0.130	0.245	0.408
Transit	0.101	0.056	-0.029	0.107	0.236
avgt2work	0.027	0.036	-0.062	0.039	0.081
mvd5yr_perc	0.092	0.019	0.066	0.087	0.151
gpq_perc	-0.189	0.044	-0.249	-0.203	-0.086
hisp_perc	0.422	0.194	0.184	0.351	0.802
college_perc	-0.191	0.078	-0.310	-0.208	-0.035
huvac_perc	0.473	0.069	0.366	0.471	0.563
sphh_perc	-0.097	0.057	-0.209	-0.094	0.022
pop_dens	0.010	0.089	-0.141	0.017	0.155

It can be seen that the R-Squared for each local model is larger than 0.51, indicating that these explanatory variables are able to explain more than 50 percent of the burglary crime rate. The model performs better in north and northwest Baltimore as compared to south Baltimore. In addition, as shown in **Figure 9 (b)** and **Figure 9 (c)**, the residuals of the model estimates present a random pattern. We further compare the GWR model with the OLS model. As shown in **Table 8**, It can be seen that both AICc and AIC are significantly smaller in the GWR model, indicating the GWR model can better explain the relationship between the variables and burglary crime rate.

Table 8. Comparisons.

Criterion	OLS	GWR
AICc	450.126	442.822
AIC	446.440	435.570

5.4. Discussion

The impact of explanatory variables on the burglary crime rate can be examined by the parameter surface which shows the coefficient value distribution across the region and the corresponding significance level. **Figure 10** shows the parameter surface for each explanatory variable for each census tract. The red color represents a significant positive relationship between the variable and the burglary crime rate; the blue color represents a significant negative relationship between the variable and the burglary crime rate; and the gray color represents the coefficient variable is not significantly different from zero. It can be observed that the estimated coefficient of each variable varies greatly across the census tracts in Baltimore City. The main reason is that each census tract has its local characteristics, and the determinants of the burglary crime rate change.

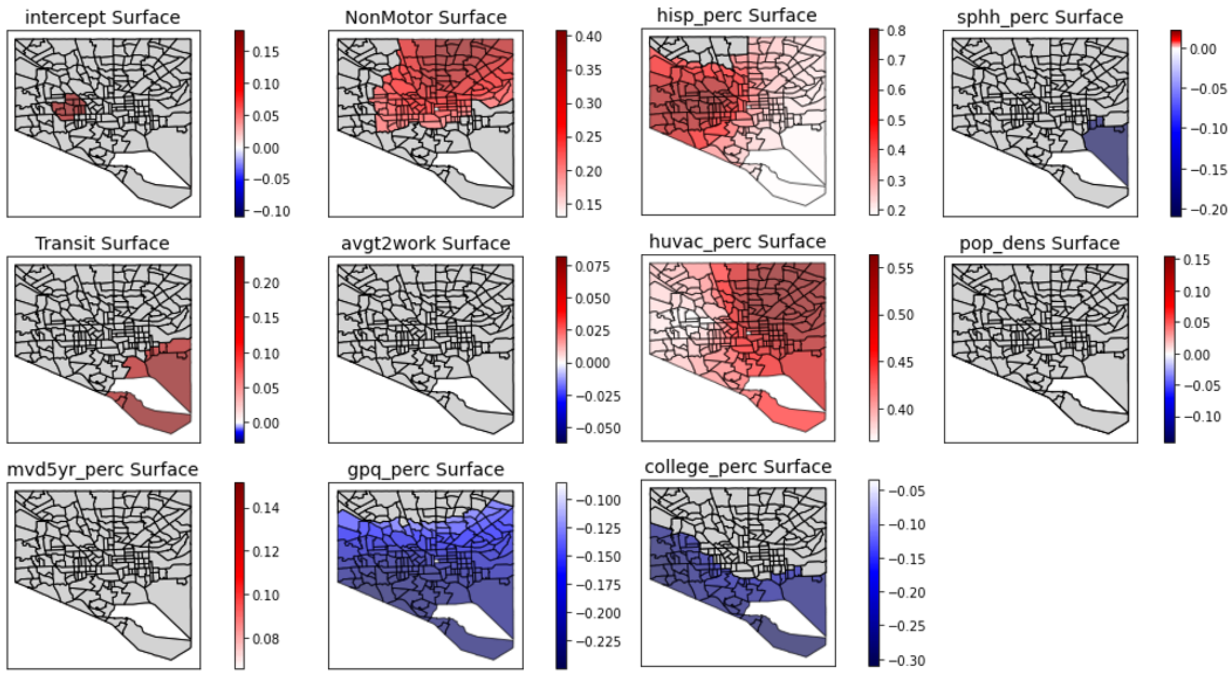
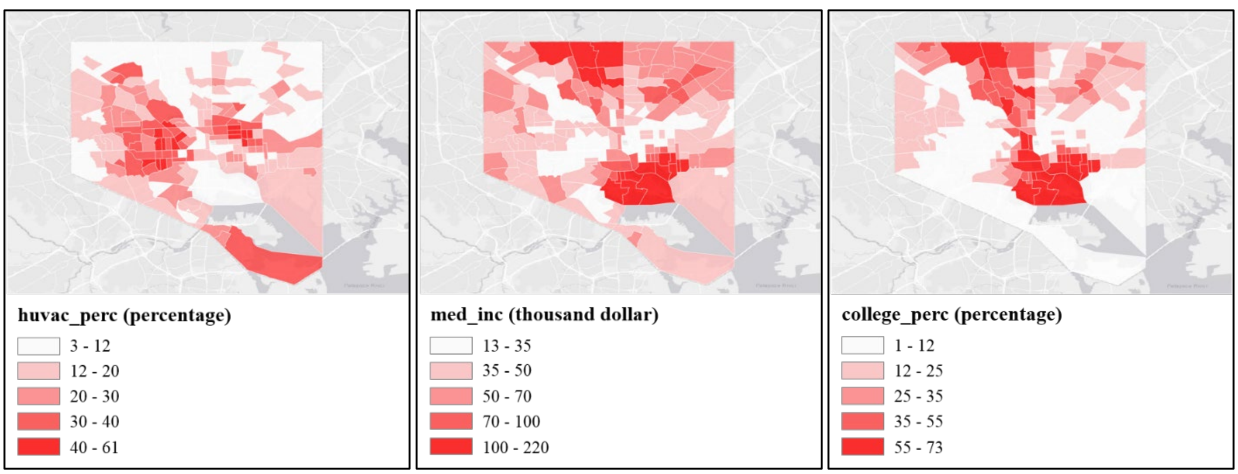


Figure 10. GWR model parameter surface.



(a) (b) (c)

Figure 11. Variables interface. (a) huvac_perc; (b) med_inc; (c) college_perc.

The intercept coefficient estimates are positively significant in the west Baltimore region, which means that this region has a significantly higher burglary crime rate compared to other parts of the city, which might be caused by other unobserved variables. This is consistent with common sense since west Baltimore is a high-crime neighborhood, and not just for burglary. It is interesting because that intercept coefficient is not statistically different from zero in the OLS regression estimates, which means that the GWR model captures the local characteristics of each census tract. Similarly, the hisp_perc coefficient estimates are positively significant in most parts of Baltimore City, where the correlation is stronger in the west Baltimore region. The huvac_perc coefficient estimates are positively significant for the entire Baltimore City. However, the correlation is less strong in the west Baltimore region. A main reason is that west Baltimore has a huge percentage of vacant housing units as compared to other regions (i.e., northeast Baltimore), thus playing a less important role in the west Baltimore burglaries (**Figure 11 (a)**). The impact of vacant housing units increases heading toward the more affluent northeast Baltimore region (**Figure 11 (b)**), where the percentage of vacant housing units presents as a strong indicator of burglary crime. These findings indicate that the government should enhance the living conditions of the minority community (i.e., Hispanic community) to keep the community from suffering frequent burglaries. In addition, the government should also deal with the vacant housing units, either by investments or providing incentives for future residents, to reduce the burglary crime rate not only in west Baltimore, but also in northeast Baltimore, even though the percentage of vacant housing units is low there.

Significant negative correlations can be found from the gpq_perc and the college perc. As for the gpq_perc, a higher percentage of the population lives in group quarters, indicating fewer housing units and households and more concentrated living conditions, where the chances for criminals to conduct burglary is low. For example, if 100% of the population lives in group quarters, it will be very unlikely to have burglaries since all people will live within the same building. Also, the negative parameter estimates for college_perc can be observed in both west and south Baltimore, where the percentage of the population with a college or higher degree is also lower than 25% on average (**Figure 11 (c)**). In light of this, to reduce the burglary crime rates in these regions, the government should provide more educational opportunities and improve the overall education level of the residents.

For the mobility and accessibility variables, we can observe a positive significant estimate for the NonMotor variable, and less or no significant estimates for Transit and avgt2work. The positive significant correlation between NonMotor and burglary crime rates can be observed in central, west, north, and northeast Baltimore. In central and west Baltimore, where the median household income is low **(Figure 11 (b)**), walking or biking tends to be the major travel mode for the residents. More non-motorized travel might include more criminals that search for targets on the streets. More importantly, we can observe that the non-motorized travel shows an even higher positive significant relationship with the burglary crime rate in the more affluent north and northeast Baltimore, where the major travel mode for the residents is driving. This finding indicates that too much non-motorized travel, which is usually conducted by non-residents, might not be a good indicator in the affluent neighborhood. The Transit coefficient estimates are not significant except for southwest Baltimore. Southwest Baltimore includes the big warehouses and port of Baltimore that have numerous freight movements, which makes them a potential target for burglaries. Though not significant, public transit might still be considered as an important factor in analyzing burglaries, and also other crimes. Public transit usually provides a cheap transportation service for people that reduces their travel cost and at the same time improves their accessibility to jobs, which will potentially lead to less poverty. Since poverty is usually positively related with crime, increasing public transit services might have a positive impact that leads to a better quality of life for people in the region. Therefore, decision-makers might consider decreasing the number of poor by providing them an improved public transit service, reducing their travel costs, improving their accessibility to opportunities, and eventually decreasing the crime rate.

6. CONCLUSION

In this project, we presented the methodology for deriving multimodal mobility patterns and accessibility measures from the mobile device location data (MDLD), and the data-driven modeling framework of high-crime neighborhoods with Geographically Weighted Regression (GWR). The crime of burglary in Baltimore City is studied using our proposed approach to show how the mobility and accessibility variables can give a different perception to understand the burglary crime rate. The results suggested that in the more affluent neighborhoods like north and northeast Baltimore, too much non-motorized travel might not be a good indicator and might increase the burglary crime rate. Though not significant, public transit might still be considered as an important factor in fighting againts burglaries, and also other crimes. Decision-makers might consider improving or expanding the public transit service in those regions suffering from poverty in order to provide the residents a cheaper transportation solution and at the same time reduce poverty and eventually decrease the crime rate. The job accessibility is not significant over the entire Baltimore region, which should be further analyzed and discussed in future research.

REFERENCES

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*. 68, 285-299, (2016).

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782, (2008).

Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*. 105, 183-202, (2019).

Wang, F., & Chen, C. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*. 87, 58-74, (2018).

Zhang, L., Sepehr, G., Michael, L. P., Chenfeng, X., Aref, D., Mofeng, Y., Qianqian, S., AliAkbar, K., and Songhua, H. An interactive COVID-19 mobility impact and social distancing analysis platform. *medRxiv* (2020).

Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A. and Zhao, G. Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends Based on Mobile Device Data. *Federal Highway Administration Exploratory Advanced Research Program* (2020).

Xiong, C., Hu, S., Yang, M., Luo, W., & Zhang, L. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences*, *117*(44), 27087-27089 (2020).

Xiong, C., Hu, S., Yang, M., Younes, H., Luo, W., Ghader, S., & Zhang, L. Mobile device location data reveal human mobility response to state-level stay-at-home orders during the COVID-19 pandemic in the USA. *Journal of the Royal Society Interface*, 17(173), 20200344 (2020).

Hu, S., Xiong, C., Yang, M., Younes, H., Luo, W., & Zhang, L. A Big-Data Driven Approach to Analyzing and Modeling Human Mobility Trend under Non-Pharmaceutical Interventions during COVID-19 Pandemic. *Transportation Research Part C: Emerging Technologies*, 102955 (2020).

Bohte, W., & Maat, K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*. 17(3), 285-297, (2009).

Gong, L., Yamamoto, T., & Morikawa, T. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. *Transportation Research Procedia*. 32, 146-154, (2018).

Ester, M., Kriegel, H.-P. , Sander, J., and Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, (Ed.)^(Eds.), *Kdd*, (1996).

Alexander, L., Jiang, S., Murga, M., and González, M. C. Origin–Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data, *Transportation Research Part C: Emerging Technologies* Vol. 58, pp. 240-250, (2015).

Farrar, D. E., & Glauber, R. R. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107, (1967).

Neter, J., Wasserman, W., & Kutner, M. H. Applied linear regression models, (1989).

Rey, S. J., & Anselin, L. (2010). PySAL: A Python library of spatial analytical methods. In *Handbook of applied spatial analysis* (pp. 175-193). Springer, Berlin, Heidelberg (2010).

Oshan, Taylor M., et al. "mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale." *ISPRS International Journal of Geo-Information* 8.6 (2019): 269.

Oshan, T., Smith, J., and Fotheringham, A. S. "Targeting the spatial context of obesity determinants via multiscale geographically weighted regression." *International Journal of Health Geographics* 19 (2020): 1-17.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel* 81 (1986).

Yu, Hanchen, et al. On the measurement of bias in geographically weighted regression models. *Spatial Statistics* (2020): 100453.

Cahill, M., and Mulligan, G. Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review* 25.2 (2007): 174-193.

Marton, D. Crime along rural interstate highways. *Free Inquiry in Creative Sociology* 23.2 (1995): 105-108.

Wang, Hongjian, et al. Crime rate inference with big data. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (2016).

Zhao, X. and Tang, J. Modeling temporal-spatial correlations for crime prediction. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. (2017).

Hanaoka, K.. "New insights on relationships between street crimes and ambient population: use of hourly population data estimated from mobile phone users' locations." *Environment and Planning B: Urban Analytics and City Science* 45.2 (2018): 295-311.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3 (1998): 431-443.

Brunsdon, C., Fotheringham, A. S. and Charlton, M. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis* 28.4 (1996): 281-298.

Gramlich, John. What the data says (and doesn't say) about crime in the United States. *Pew Research Center* (2020).

Truman, J., and Planty, M. Criminal Victimization, 2011. U.S. Department of Justice Office of Justice Programs Bureau of Justice Statistics. (2012).

Cohen, L. and Felson, M. Social change and crime rate trends: A routine activity approach. *American Sociological Review* (1979): 588-608.

Caminha, C., Furtado, V., Pequeno, T. H., Ponte, C., Melo, H. P., Oliveira, E. A., & Andrade Jr., J. S.. Human mobility in large cities as a proxy for crime. *PloS one*, *12*(2), e0171609 (2017).