ABSTRACT

Title of Thesis:	A NOVEL MEASUREMENT OF JOB ACCESSIBILITY BASED ON MOBILE DEVICE LOCATION DATA
	Guangchen Zhao, Master of Science, 2022
Thesis Directed By:	Professor, Lei Zhang, Department of Civil and Environmental Engineering

Mobile device location data (MDLD) can offer a new perspective on measuring accessibility. Compared with the traditional accessibility measures, MDLD is capable of capturing people's preferences with the observed locations. This study proposes a job accessibility measure based on the identified home and work locations from MDLD, evaluating the job accessibility by the proportion of workers identified working in zones within a certain time threshold. In the case study on the Baltimore region, the job accessibility from the MDLD-based measure is compared with the results from a widely-used traditional measure. Then, generalized additive models (GAM) are built to analyze the socio-demographic impact on job accessibility from a MDLD-based measure and a traditional measure, with a feature-to-feature comparison. Finally, the socio-demographic characteristics of regions where there are major disparities between the job accessibility from the traditional measure and the MDLD-based measure are also evaluated from the Student's t-test results.

A NOVEL MEASUREMENT OF JOB ACCESSIBILITY BASED ON MOBILE DEVICE LOCATION DATA

by

Guangchen Zhao

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2022

Advisory Committee:
Professor Lei Zhang, Chair
Professor Paul Schonfeld
Associate Research Professor Chenfeng Xiong

Acknowledgements

First, I would like to express my grateful appreciation to my advisor, Dr. Lei Zhang, who offered me the chance to study and work in this outstanding research group at the University of Maryland, and always being supportive in the past few years. His immense knowledge and enthusiasm for work have inspired me in my academic research and daily life.

I would also like to express my sincere gratitude to Dr. Paul Schonfeld and Dr. Chenfeng Xiong for serving on my master thesis committee and offering me their invaluable comments. Dr. Schonfeld's teaching style and professional for this topic made a strong impression on since my first semester here. I would like to express my special gratitude to Dr. Xiong for his insightful comments and suggestions at every stage of this research.

I also want to thank my friends and colleagues in our research group offering me kind help and support: Aliakbar Kabiri, Sepehr Ghader, Aref Darzi, Songhua Hu, Mofeng Yang, Weiyu Luo, and Yixuan Pan. Thank you all.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Improving accessibility is one of the ultimate goals of transportation planning [1], as the availability of meaningful, affordable, and accessible opportunities such as healthcare, employment, food, and education are critical to the well-being of any family and individual. Social inequalities and social exclusion to a large extent result from accessibility inequalities [2, 3]. In the field of urban planning, job accessibility is a critical subject to study among all those different types of accessibility, as a key index to understand the problems such as the spatial mismatch of jobs and housing [4, 5], and urban form [6], etc.

One of the earliest definitions of accessibility is "the potential of opportunities for interaction", by Hansen in 1959 [7]. Over the past few decades, many definitions of accessibility were proposed, mainly defining accessibility in terms of opportunities and travel impedance or resistance. Higher accessibility means more options to fulfill activity needs and lower generalized costs (travel time, monetary cost, etc.) to reach these options [8]. Traditional accessibility studies usually use land-use data in terms of population or employment, as a proxy for measuring the attractiveness of destinations [9], and rely on model estimates of travel time (or distance) as an indicator of impedance. Challenges in accessibility studies stem from the requirement of detailed knowledge of activity locations, transportation networks, and individual travel behaviors [9]. When it comes to job accessibility, cumulative opportunity measure, also

known as isochronic measure, contour measure, or proximity count, counting the number of jobs that can be reached within a given time threshold, is so far most widely used by land use and transportation practitioners [11]. The advantage of this measure is related to its high operationalization and interpretability for both researchers and policymakers, while the most significant disadvantage of this approach is the negligence of individual choices and preferences, considering all jobs are accessible as long as they can be reached [12, 13].

As a new data source, mobile device location data, MDLD in abbreviate, offers a novel paradigm for measuring and analyzing accessibility. MDLD typically records the unique device identifier, the geographic coordinates, and the timestamp for each of the location records, from various sources including Global Positioning Service (GPS) devices, cellular networks, Bluetooth, and Wi-Fi, etc. In accessibility studies, specifically, there are two inherent strengths make MDLD have great potential: first, observed location trajectories extracted from MDLD reveal where travelers choose to go and how they get there every single day. Therefore, land-use proxy, revealed-preference survey, or destination choice model are less relied on or not even needed. Second, the high spatial and temporal resolution, much larger sample size, and continuous observations day after day in MDLD to a large extent remove previous data constraints. With traditional accessibility measures such as the number of jobs reachable within a certain travel time budget (e.g., 20 minutes), a low-income neighborhood next to downtown could be considered a place with high accessibility to jobs, even if no one from the community really works downtown. Based on observed

destinations in MDLD, accessibility can be measured with greater accuracy, especially for underserved communities.

1.2 Research Objective

The objective of this study is to develop a zone-level job accessibility measure based on the MDLD data as the proportion of workers observed working in zones within a certain time threshold. The home and work locations of workers should be identified based on MDLD, revealing actual individual choices and preferences, while the proposed measure still stays easy to interpret. In addition, the results from the traditional and proposed measures should be compared, and the confounding socio-demographic effects on job accessibility should be analyzed.

1.3 Research Approach

In order to fulfill the research objective, the research approach of this study is identified as the following four steps: (1) evaluate the state-of-the-practice accessibility measures as well as the state-of-the-art methods based on MDLD data, and identify the key research gap from the literature review; (2) employ a suitable algorithm to identify home and work locations from the MDLD data, as the data input for the new measure; (3) propose an MDLD-based accessibility measure and both geographically and statistically evaluate the differences between job accessibility measured by the traditional and the proposed methods; and (4) analyze the impact of zone-level socio-demographics on job accessibility measured by the traditional and the proposed methods respectively with the Generalized Additive Model (GAM) and T-tests, then conduct a feature-by-feature comparison.

1.4 Outline

The rest of the thesis is organized as follows. In Chapter 2, the literature review on the state-of-the-practice accessibility measures, as well as the state-of-the-art methods of home and work location imputation and accessibility studies based on MDLD data is conducted. In Chapter 3, the MDLD used in this study, which is obtained from multiple leading providers, is described. In addition, the skim matrices produced by a statewide activity-based model, as well as other supplement datasets including LEHD Origin Destination Employment Statistics (LODES) data and American Travel Survey (ACS) data are also introduced. In Chapter 4, the methodology of the home and work location imputation, as well as the definition of the MDLD-based job accessibility measure are demonstrated. The traditional measure selected as a comparison is introduced. The feature engineering process, model selection, and application for the socio-demographic impact study are also included. In Chapter 5, the job accessibility results of the case study from the MDLD-based measure, as well as the results from the traditional measure are presented, and the difference is statistically compared. The generalized additive models are employed on those two results respectively to evaluate the socio-demographic impact, and a feature-by-feature comparison is conducted. Besides, the socio-demographic characteristics of regions whose job accessibility has major disparities between the traditional measure and the MDLD-based measure are also evaluated from the results of T-tests. Chapter 6 offers conclusions about this study, contributions this study make, and suggestions for future research.

Chapter 2: Literature Review

2.1 Accessibility

2.1.1 Importance of Accessibility

For all groups of individuals, the availability of opportunities such as employment, healthcare, food, and education, etc., which are essential for healthy living, should be studied and considered in the decision-making of public policies. Researchers have suggested that approaches such as benefit-cost analysis and multi-criteria analysis do not fully address equity concerns in the development and evaluation of public policies [17, 18]. Research also shows that the more vulnerable segments of the population suffer more from insufficient transportation services such as longer travel times, higher pollution exposure, and accident risks [19].

Accessibility, the ease with which any land-use activity can be reached from an origin given a specific transportation system [20], can reveal valuable information about the availability of social and economic opportunities for people in a particular neighborhood or population segment [12]. The contribution of accessibility inequalities in producing social inequalities and social exclusion has been highlighted in the literature [2, 3, 17]. Therefore, improving accessibility and spatial equity of opportunities are considered in evaluating land-use or transportation policies and prioritizing relevant projects, for instance, analyzing the inadequacies of transportation service and providing evidence on disparities among neighborhoods [21]. As a result, many planning agencies around the world have focused on, or at least taken the

relationship between accessibility and equity into consideration in their planning processes [21, 22].

2.1.2 Accessibility Measures

Graphical techniques are utilized in the early stage of accessibility studies [23, 24]. In this approach, the distribution of the proportion of opportunities accessible by time is represented by a cumulative distribution curve. These curves theoretically can be created for different study areas, modes, or socio-demographic groups, but the comparison is difficult in practice. As a results, single-value measures of accessibility are mostly applied.

Table 2-1. Summary of accessibility measures

Measure category	Infrastructure-based	Location-based	Person-based	Utility-based
Example	travel time; travel speed; level-of-service	number of opportunities within given time; gravity measures	space-time prism	logsum
Transportation component	Y	Y	Y	Y
Land-use component	N	Y	Y	Y
Time component	P	P	Y	P
Person component	N	N	Y	Y
Simplicity	Y	Y	N	P

a. “Y” = Yes, “N” = No, “P” = Possible.

Single-value measures of accessibility traditionally depend on four main components: (1) land use, representing the spatial distribution of opportunities; (2) transport,

representing the travel impedance to reach opportunities; (3) time, representing people's schedule as well as time constraints related to the availability of opportunities; and (4) person, representing people's needs and abilities to participate in activities [12, 25]. Focus on different components of accessibility has led to various indicators and methodologies for measuring accessibility, which can be divided into four categories [12], summarized in Table 2-1.

The first is the infrastructure-based measures, such as "average travel speed" or "level of congestion", which mainly measure mobility, describing the network performance or level of service [26, 27]; the second is the location-based measures, such as "number of opportunities accessible within 20 minutes", which quantify the accessibility of locations at macro scale, describing the accessibility to the spatially distributed opportunities [28-31]; the third is the person-based measures, such as "number of activities an individual is able to participate in a certain time", describing individual-level accessibility based on individual's spatial and temporal constraints [32-34]; and the fourth is the utility-based measures, such as logsum, measuring the economic benefit of accessing spatially distributed opportunities [9, 35, 36]. Infrastructure-based measures have the advantage of simplicity and interpretability, but such measures only cover the transport component of accessibility and are unsuitable for social analysis or economic assessment of land-use change. Location-based measures, such as "cumulative opportunity numbers" or gravity-based measures, cover both transport and land use components, but they are insensitive to the person component of accessibility [37, 38]. In addition, there are issues related to spatial aggregation, definition of attractiveness measure, and the construction of the friction coefficient [39]. Person-

based and utility-based measures theoretically better capture all accessibility components, but they become complex and require more individual-level observations.

Another dimension of the classification of accessibility is active and passive accessibility. Active accessibility refers to the easiness of the people in a certain zone to do their tasks, while passive accessibility refers to the easiness of the places to be reached by the users in a certain location [40]. This study is focused on active accessibility.

2.1.3 Job Accessibility

Job accessibility is a critical subject to study among all those different types of accessibility in the field of urban planning, and equity of job accessibility is always the research focus.

Levinson (1998) [41] established an ordinary least squares regression model to analyze the socio-demographic effect on the commuting duration, based on a regional travel survey. Although the data is limited at that time and the model is relatively simple, the findings are valid, that people who live in job-rich places and people who work in housing-rich areas have shorter commuting times. Hernandez (2017) [42] evaluates the accessibilities with the cumulative opportunity measure and analyzes the job accessibility of three income groups: low, mid, and high. The limitation of this work is that it only focuses on public transport accessibility, while the usage of public transport may be different across the different income groups. Deboosere et al. (2018) [13] also evaluated job accessibility by public transport with the cumulative opportunity measure. The time threshold is calculated as the average commute time, and along with

calculating accessibility to all kinds of jobs, accessibility to low-income jobs is also discerned. Owen et al. (2015) [43] consider job accessibility from a temporal aspect, evaluating transit systems by "continuous accessibility", which is calculated every minute continuously in a period of time. Hu [44] estimates job accessibility for the poor in Los Angles, and the results show poor job seekers do not face spatial mismatch which means the poor who live in the city still have higher job accessibility than poor who live in the suburban. The limitation of this study is that the accessibility is fully based on automobile drivers, while due to insufficient data, access by public transit, which could be the main access for the poor is not calculated.

2.2 Mobile Device Location Data

2.2.1 Summary of Mobile Device Location Data

Table 2-2. Mobile Device Location Data for Mainstream Use

MDLD	**Example**	**Spatial Coverage**	**Temporal coverage**	**Frequency of collection**	**Accuracy**
GPS data	GPS-based travel survey; GPS Data without User Recall	Low	Low	High	High
Cellular Data	Call Detail Record (CDR); sightings	High	High	Low	Low
LBS data	Smartphone activities	High	High	Low	High

The mobile device location data for mainstream use in transportation research can be categorized into three types: Global Positioning System (GPS) data, Cellular Data, and

Location-based Service (LBS) data. Table 2-2 summarizes the comparison between their spatial coverage, temporal coverage, and sample rate.

GPS data has been helping on enhancing travel surveys since the late 1990s [45, 46]. The earliest way of collecting GPS data was by in-vehicle GPS loggers, which record the location data usually at one-second interval when the vehicle is moving and stops recording when the vehicle stops [47]. This method is efficient and convenient for users, but it can only record the in-vehicle part of a trip without any information about the off-vehicle part. To obtain non-vehicle travel activities such as walking or biking, some later travel surveys introduced wearable GPS loggers [48, 49]. In recent years, advanced GPS-based surveys collected GPS data from smartphone applications running background, which is more portable than wearable GPS loggers [50]. Besides those GPS-based travel surveys with user recall as ground truth information, there is also GPS data without user recall which is widely used for estimating the real-time traffic speed and travel time [51, 52].

The cellular data mainly includes Call Detail Record (CDR) and sightings. The CDR data is more frequently used, which records the location of the cellphone tower when users have a call [53-55]. The sightings data is less used, but it has higher spatial resolution than CDR because the location of the cell phone user is imputed by triangulation of multiple towers [53]. The cellular data is also widely used for understanding individual human mobility patterns [54-56].

Location-based Service (LBS) data gets more attention as smartphones become popular. LBS data records location data by smartphone applications from multiple

sources, such as Bluetooth and Wi-Fi, as well as GPS and cellular networks, based on the best accuracy currently available [53]. It provides users' exact location with high accuracy and low location recording interval (LRI). Thus LBS data is the most popular location data in the current stage for both research on human mobility, and commercial data markets [60-62].

The most significant advantage of the GPS-enhanced travel survey data is the frequency of collection, as it keeps collecting data with a low recording interval that neither cellular data nor LBS data are comparable because they can only collect location information when users are using the related services. On the other hand, the shortage of GPS-enhanced travel survey data comes from its low spatial and temporal coverage, as those surveys are usually just conducted in a small region covering a small number of respondents. On the contrary, cellular data and LBS data have much higher spatial and temporal coverage since a much higher percentage of users will call, text or use location-based services on their phones almost every day. In summary, multiple sources of MDLD can provide observed information about the chosen destinations and offer a new perspective in human mobility study.

2.2.2 Home and Work Location Identification based on MDLD

Although MDLD contains more frequent and accurate location records than traditional travel surveys, in most cases it does not have the ground truth information as reported in the surveys, such as travel mode, trip purpose, and users' home and work locations, etc., which are crucial components for many fields of transportation study such as travel demand modeling, traffic assignment, and accessibility study as well. To study the job

accessibility, home and work location is the basis of further analysis [63]. Firstly, the geographic coordinates (latitudes and longitudes) are usually grouped as people's potential activity centers by rule-based methods [64, 65] or clustering methods [66, 67]. Then the classification of the types of those identified activity centers is always based on context-based methods and behavior-based methods [70]. As its name suggests, the context-based methods recognize the types of activity centers based on their surrounding contexts, such as land-use characteristics and types of point-of-interest (POI) nearby [72, 73].

The behavior-based methods are the most widely used for identifying the categories of activity centers, such as home and work locations. As the name suggests, it focuses on behaviors of activity centers, such as time of arrival, dwelling time, and frequency of the observations, etc. [74]. Specifically, for the identification of home and work locations, they are always considered as the where people spend most significant times [68, 69] with the most frequent visits [70, 71] as well during nighttime and daytime, respectively. Pan (2021) [15] proposed a framework that for identifying the home locations, both the overall hours and nighttime hours spent are considered, and for identifying the work locations, a temporal similarity ratio is specifically introduced to prevent identifying places which are close to home locations with frequent visits by people as the work locations. In summary, the behavior-based methods are most applicable for identifying the home and work locations based on MDLD, which rarely contains any personal information.

2.2.3 Accessibility Study based on MDLD

Accessibility studies based on different types of MDLD emerged in the past few years. García-Albertos et al. (2019) [10] utilizes CDR data to calculate the attraction mass of each zone and evaluated the accessibility hourly to consider the temporal changes. Results show that the accessibility is subject to change in peak hours and off-peak hours. Moya-Gómez et al. (2018) [75] uses TomTom, which contains historical speed for the road network based on users' average travel time, to calculate the travel impedance, and geotagged Tweets to calculate the attractiveness of destinations. This study focuses on dynamic accessibility as well. Zhang et al. (2021) [76] utilizes passively collected location data to measure the zonal accessibility by average travel time to opportunities and evaluated the relationships with the median household income. Although the model is relatively simple, it is shown that the mid-income group has the longest travel time to work, and the low-income group shows the longest travel time to food and healthcare.

2.4 Research Gap

As the literature shows, the most widely used state-of-the-practice measure of job accessibility is the cumulative opportunity measure, which suffers from the lack of accurate and detailed individual-level data, in particular the high-quality MDLD data. Moreover, the current research on the socio-demographic effects on job accessibility is mostly limited to dividing the population into several demographic groups and visualizing their average value of accessibility, while there is no literature evaluating the confounding socio-demographic effects on job accessibility via relatively

comprehensive models. This study aims to fill this gap by proposing a job accessibility measure based on home and work locations identified from high-quality Mobile Device Location Data and applying it along with the traditional cumulative accessibility measure on a regional dataset to compare their performances. In addition, generalized addictive models are built on those two results respectively to evaluate the confounding socio-demographic impact, and a comprehensive feature-by-feature comparison is conducted.

Chapter 3: Data

3.1 Mobile Device Location Data

The mobile device location data used in this study is obtained from multiple leading data providers in the United States covering the whole year of 2020. Each raw location sighting from MDLD records an anonymized unique device identifier (ID), the geographic coordinates (latitude and longitude), the timestamp, and accuracy (estimated positioning error, in meters), etc. To better describe the high quality of the MDLD, a set of metrics are employed in terms of sample consistency and population coverage (i.e., monthly active users and regularly active users), temporal consistency, and coverage (i.e., data frequency and active local hours), and location accuracy. The definition and statistics of those quality metrics from the one-month raw sighting data panel in 2020 are listed in Table 3-1.

Table 3-1. Definition and Statistics of Quality Metrics of MDLD

Quality Metrics	Definition	Statistics
Monthly active users (MAU)	The number of devices with at least one sighting in a month.	270,601,232 devices
Regularly active users (RAU)	The number of devices with at least seven days of at least ten daily sightings in a month	68,016,290 devices
Data frequency	The average daily number of sightings for RAUs	234.4 sightings/day
Active local hours	The average daily number of local hours observed for RAUs	6.4 hours/day
Location accuracy	The average positioning accuracy of RAU devices	15 meters

The 270,601,232 monthly active users, which means a national sampling rate of more than 80% on a monthly basis, together with the 68,016,290 regularly active users, which means a sampling rate of more than 20% regarding temporally consistent devices, indicate the great sample consistency and massive population coverage. The average of 234.4 sightings and 6.4 local hours observed per day for the RAUs show high temporal consistency and coverage, which is crucial for the imputation of home and work locations. And the average positioning accuracy of 15 meters ensures the strong reliability of location sightings for RAUs.

3.2 Skim matrix from Maryland Statewide Transportation Model

A skim matrix produced by the Maryland Statewide Transportation Model version 2 (MSTM2) [14], is utilized to measure the zone-to-zone travel time as the impedance. MSTM2 is a statewide activity-based model (ABM) developed for the Maryland State Highway Administration (SHA) for the needs of statewide long-range planning, intercity travel, and planning for rural areas [14], while it covers not only Maryland, but also Washington D. C., Delaware, and parts of Pennsylvania, Virginia, and West Virginia. The off-peak highway skim is selected to best estimate the maximum reachable area within a certain time threshold.

3.3 LEHD Origin Destination Employment Statistics

The Longitudinal Employer Household Dynamics (LEHD) Origin Destination Employment Statistics (LODES) [58], produced by the LEHD program at the U.S. Census Bureau, provides information on workers' residential locations and

employment locations for jobs covered by State unemployment insurance reporting and federal worker earnings records [59].

This study utilizes two sets of LODES data: (1) Workplace Area Characteristic (WAC) data, where jobs are totaled by their employment Census Blocks, for the calculation of the cumulative job opportunities as the traditional job accessibility. Since the locations which can be reached from the Baltimore region within a given time threshold include not only locations from the State of Maryland, but also from the surrounding states, the WAC data of the District of Columbia, Delaware, New Jersey, Pennsylvania, Virginia, and West Virginia are also collected and combined with the WAC data of Maryland, to ensure the accuracy and completeness of the calculation. (2) Residence Area Characteristic (RAC) data, where workers are totaled by their home Census Blocks, for the calculating of the sampling rate for workers, and feature construction of the generalized additive models to evaluate the socio-demographic impact on those two job accessibility measures. The RAC data provides information on the number of workers categorized by age groups, earning groups, industry types, race, ethnicity, educational attainment, and genders, which all contribute to the features of the socio-demographic impact. The RAC data is collected for the Baltimore region, the subject area for the case study, containing 1,945 Block Groups in total.

3.4 American Community Survey

As another essential data source for the sampling rate calculation and feature construction of the generalized additive models, the 2019 American Community Survey (ACS) 5-year estimates [77] conducted by the United States Census Bureau is

utilized as well. The population, proportion of urban areas, and proportion of housing units occupied by renters, are included at the census block group level in the collected ACS data. The ACS data is also collected for the 1,945 Block Groups of the Baltimore region.

Chapter 4: Methodology

4.1 Methodological Framework

Figure 4-1 presents the methodological framework of this study. There are three major components: home and work location identification, job accessibility calculation, and socio-demographic impact analysis. In the home and location identification, the pre-processed Mobile Device Location Data works as the input of the location identification algorithm, and the outputted results are validated against LODES and ACS data introduced in chapter 3. In the job accessibility calculation, the MDLD-based measure and traditional measure are performed separately to obtain two sets of job accessibility results. Finally, in the sociodemographic analysis, generalized additive models are constructed to study the confounding effects of various features on both the MDLD-based and traditional job accessibility.

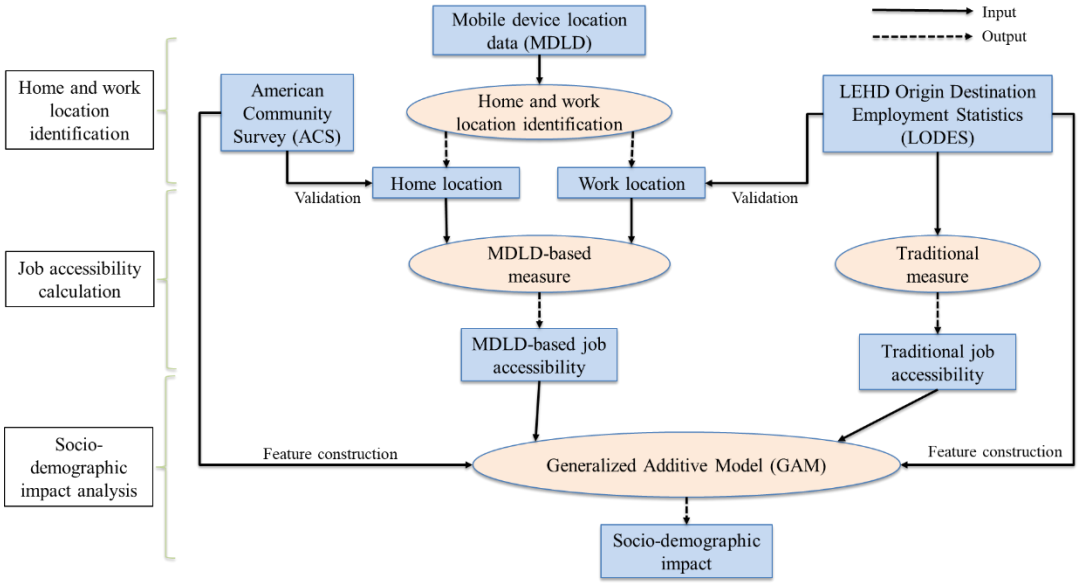


Figure 4-1. Methodological Framework

4.2 Job Accessibility Measure

4.2.1 Traditional Job Accessibility Measure

The traditional job accessibility measure employed in this study, which works as a baseline for comparison, is defined as the cumulative number of jobs within a certain time threshold, which is calculated as:

$$A_i^{Trad} = \sum_j J_j f(t_{ij}), \qquad f(t_{ij}) = \begin{cases} 1 \; if \; t_{ij} \leq t_{threshold} \\ 0 \; if \; t_{ij} > t_{threshold} \end{cases}$$

where A_i^{Trad} denotes the traditional job accessibility of zone i, J_j denotes the number of jobs in zone j, which is collected from the latest 2019 LODES data where jobs are totaled by their employment locations. t_{ij} denotes travel time from zone i to zone j, which is obtained from the highway skim matrix estimated by MSTM2, a statewide activity-based model. $f(\cdot)$ is the binary impedance function, yielding 1 if J_j can be reached within some travel time threshold $t_{threshold}$, and 0 otherwise. The scale of A_i^{Trad} is a positive integer, and the larger A_i^{Trad} indicates the better traditional job accessibility of zone i.

As suggested in literature review, this cumulative opportunity measure is most widely used in the traditional measure of job accessibility, which takes into account all the job opportunities within the reachable zones while does not consider whether workers actually work in those zones, as a result of lacking the person-level information.

4.2.2 MDLD-Based Job Accessibility Measure

The MDLD-based job accessibility measure proposed in this study is defined as the proportion of workers identified working in zones within a certain time threshold, which is calculated as:

$$A_i^{MDLD} = \frac{\sum_j W_{ij} f(t_{ij})}{W_i}, \qquad f(t_{ij}) = \begin{cases} 1 \; if \; t_{ij} \leq t_{threshold} \\ 0 \; if \; t_{ij} > t_{threshold} \end{cases}$$

where A_i^{MDLD} denotes the MDLD-based job accessibility of zone i, W_i denotes the number of workers identified living in zone i, W_{ij} denotes the number of workers identified living in zone i and working in zone j. The workers' home and work locations are identified solely from the MDLD by a behavior-based method, which is introduced in next section. t_{ij} denotes travel time from zone i to zone j, and $f(\cdot)$ is the binary impedance function, yielding 1 if J_j can be reached within some travel time threshold $t_{threshold}$, and 0 otherwise. The t_{ij} is also obtained from the skim matrix by MSTM2. Therefore, when $t_{threshold}$ is set the same, the range of zones considered in the MDLD-based measure are exactly the same as in the traditional measure, which makes the results more comparable. The scale of A_i^{MDLD} is theoretically any value between 0 and 1, which means nobody actually works in considered zones when A_i^{MDLD}=0, and all workers work in considered zones when A_i^{MDLD}=1. The larger A_i^{MDLD} indicates the better MDLD-based job accessibility of zone i.

The MDLD-based measure aggregates observed person-level information into zone-level accessibility. Instead of considering all the job opportunities within the reachable

zones available to people living in each region, as what the traditional measure does, the MDLD-based measure examines how many workers are actually working in those zones, or in other words, how many jobs in the reachable zones are actually accessed by workers living in each regions. Thus, it is argued that the MDLD-based measure is able to better capture people's individual preferences, and reflect more realistic job accessibility.

4.3 Home and Work Location Identification

4.3.1 Data Preprocessing

For the purpose of data quality control, before implementing the actual location identification algorithm, data preprocessing is applied to the raw sightings data to address two dimensions of data quality assessment: consistency and accuracy [78], which includes the following steps: (1) deleting invalid data with unreasonable attributes, for instance, missing or negative values for geographic coordinates; (2) deleting duplicated data with the same timestamp and only keep the sighting with the smallest positioning error; (3) deleting low-accuracy data with the positioning error greater than 150 meters [15] considering the accuracy requirements of home and work location identification; and (4) ordering the data by timestamps to make sure the sighting records for each device are in the time sequence.

In addition to the quality control, all the location sightings are encoded into geohash zones, which is a public domain geocode system, to improve the computational efficiency of the tremendous amount of MDLD. Instead of geographical coordinates, a geohash aggregates locations into rectangular cells with 12 levels of precision.

Specifically, the level-6 geohashes with a cell size of 1.22 km × 0.61 km, and the level-7 geohashes with a cell size of 153 m × 153m (i.e. each level-7 geohash identifies one of 32 sub-cells of a level-6 geohash), are utilized in this study for the home and work location identification.

4.3.2 Home Location Identification

Based on the preprocessed MDLD, this study employs a behavior-based method with a set of heuristic rules to identify the home and work locations of individuals included in the data sample. As suggested in the literature [14], most of the time, especially at nighttime, people stay at their home, and for some regular hours during daytime, people stay at their workplace. Based on 2017, 2018, and 2019 American Time Use Survey (ATUS), the nighttime window is defined as 9:00 p.m.–5:59 a.m., when more than 80% of workers stay at home for some time.

For home location identification, first, for each individual, the level-6 geohash zones are kept as his/her home location candidates only if: (1) they are observed half of the total observed days or more in a month (at least three days), and (2) they are observed average of two hours or more in those observed days. This filter is for excluding individuals with rare observations. Second, the home location candidate with the most observed number of nights is identified as the level-6 geohash zone of the home location. If a tie exists in the most observed number of nights, the tied candidate with the most observed number of hours during the nighttime is identified, and if a tie still exists, the tied candidate with the most number of sightings during the nighttime is identified. Then for better spatial resolution, inside the level-6 geohash zone of the

home location, the same identification procedure is repeated to find the level-7 geohash zone of the home location with the most observed number of nights, the most observed number of hours, and the most observed number of sightings during the nighttime.

The parameter, 2 as the minimum average hours in observed workdays, is set based on the Pearson correlation between the identified number of residents in MDLD and the reported population over 16 in ACS at the county level, which is higher than 0.95.

4.3.3 Work Location Identification

Similar to the home location identification, for each individual, the work location starts with a filter that only keeps the level-6 geohash zones as his/her work location candidates if: (1) they are not the same as his/her home location (2) they are observed half of the total observed workdays or more in a month (at least three workdays), and (3) they are observed average of two hours or more in those observed workdays.

There is a fact that those individuals living close to the boundaries of level-6 geohash zones could be observed spending a significant amount of time in the neighboring level-6 geohash zones of their home location. Thus, in order to avoid identifying such neighboring geohash zones as their work locations, while still being able to identify the actual work locations that are close to home, a temporal similarity is utilized to further identify the work location candidates, defined as:

$$S_i = \frac{H_i^h}{H_i}$$

where S_i denotes the temporal similarity of work location candidate i, H_i^h denotes the number of unique hours with observations in both level-6 geohash zone of the home

location and work location candidate i during the month, and H_i denotes the number of unique hours with observations in work location candidate i during the month.

Since the actual work location is supposed to have a relatively small temporal similarity with the home location, the work location candidates are further filtered by keeping zones with the $S_i \leq 0.6$. Then, the work location candidate with the most observed number of workdays is identified as the temporary level-6 geohash zone of the work location. If a tie exists in the most observed number of workdays, the tied candidate with the most observed number of hours during the workdays is identified, and if a tie still exists, the tied candidate with the most number of sightings during the workdays is identified. Finally, for better spatial resolution, inside the level-6 geohash zone of the work location, the same identification procedure is repeated to find the level-7 geohash zone of the work location with the most observed number of workdays, the most observed number of hours, and the most observed number of sightings during the workdays.

Those parameters, including 2 as the minimum average hours in observed workdays, and 0.6 as the temporal similarity threshold, are set based on the Pearson correlation between the identified number of workers in MDLD and the reported number of workers in LODES at the county level, which is higher than 0.95.

4.4 Model Analysis of Socio-demographic Impact

4.4.1 Generalized Additive Model

The Generalized Additive Model (GAM), developed by Hastie and Tibshirani in 1990 [80], is a popular statistic model used for modeling and analyzing the relationship between the dependent variable and the independent variables. The advantages of GAM over the standard linear regression model estimated from ordinary least squares (OLS) include: (1) the responsible variable Y may follow various distributions other than Gaussian distribution, such as Poisson distribution or Binomial distribution; (2) it is not limited to fitting the linear relationships, while flexible for handling the nonlinear relationships by spline functions [81]. The structure of GAM can be formulated as:

$$g(Y_i) = \beta_0 + \sum_{m=1}^{M} \beta_m * X_{mi} + \sum_{n=1}^{N} f_n(X_{ni}) + \varepsilon_i$$

where Y_i denotes the dependent variable in block group i, as the analysis is at block group-level in this study, $g(.)$ denotes the link function, β_0 denotes the overall intercept of the model, β_m denotes the m^{th} coefficient of X_{mi}, which denotes the m^{th} independent variable with fixed effects in block group i, M equals to the total number of variables with fixed effects, $f_n(.)$ denotes the non-parametric smooth function for X_{ni}, the n^{th} independent variable with nonlinear effects in block group i, N equals to the total number of variables with nonlinear effects, and finally, ε_i denotes the error term.

Based on the testing of various combinations of the distribution of dependent variables, and the smoothing functions of handling the nonlinear effects, the optimal model inputs are selected with the smallest Akaike Information Criterion (AIC). As a result, the distribution of dependent variables in this study are assumed to follow the Gaussian distribution, thus the link function $g(.)$ is the identity links; the independent variable with nonlinear effects is the random effects across different counties, thus the smooth function $f_n(.)$ is smoothers with parametric terms penalized by a ridge penalty [82].

The GAMs are performed with the open-source package "*mgcv*" in R developed by Simon Wood [83]. The smoothness selection method used here is Restricted Maximum Likelihood Estimation (REML), which is supposed to have fewer under-smoothing and overfitting issues [84].

4.4.2 Variables

Table 4-1 lists the variables used in this study. The dependent variables are the job accessibility calculated from these two measures, i.e., the MDLD-based measure A^{MDLD}, and the traditional measure A^{Trad}. The following three aspects are considered to have confounding effects on job accessibility: (1) the basic features of each block group, such as the population density, the job occupation density, and if it is in urban areas, etc.; (2) workers' socio-demographic characteristics, such as workers' race, age, gender, education level, and earnings, etc., which may indicate social inequality issues if the disparities exist; (3) workers' industries, which are highly possible to contribute to the differences in job accessibilities, as for jobs in different industries, either the geographical distribution or the normal working behaviors differ from each other.

Therefore, various independent variables related to these three aspects are included in this study, and summarized in Table 4-1. Several variables are excluded from the models by the multicollinearity check, as their variance inflation factor (VIF) values are greater than 10.0.

Table 4-1. Dependent and Independent Variables Used in the GAM Model

Variables	Description
Dependent Variables	
A^{MDLD}	The MDLD-based job accessibility (the proportion of workers identified working in zones within a certain time threshold)
A^{Trad}	Traditional job accessibility (the cumulative number of jobs within a certain time threshold)
Independent Variables	
Block Group-Related	
Population density	Population density (10^3 people/mile^2)
Worker density	Worker density by residence (10^3 workers/mile^2)
Job density	Job density (10^3 jobs/mile^2)
Urban	The proportion of urban areas
Rent	The proportion of renter occupied housing units
Workers' Socio-Demographic	
White	*The proportion of White workers*
African American	The proportion of African American workers
Asian	The proportion of Asian workers
Hispanic	The proportion of Hispanic workers
Bachelor Degree	The proportion of workers with educational attainment: Bachelor's degree or advanced degree
Male	The proportion of male workers
Earning_0_15k	The proportion of workers with earnings $15000/year or less
Earning_15k_40k	*The proportion of workers with earnings $15001/year to $40000/year*
Earning_40k_more	The proportion of workers with earnings greater than $40000/year

Age_younger_29	The proportion of workers age 29 or younger
Age_30_54	*The proportion of workers age 30 to 54*
Age_55_older	The proportion of workers age 55 or older
Workers' Industry	
Natural Resources and Mining	The proportion of workers in agriculture, forestry, fishing, hunting, mining, quarrying, and oil and gas extraction sectors
Utilities	The proportion of workers in utilities sector
Construction	The proportion of workers in construction sector
Manufacturing	The proportion of workers in manufacturing sector
Trade	The proportion of workers in wholesale trade and retail trade sectors
Transportation and Warehousing	The proportion of workers in transportation and warehousing sector
Information	The proportion of workers in information sector
Financial	The proportion of workers in finance, insurance, real estate, and rental and leasing sectors
Professional and Business Activities	The proportion of workers in professional, scientific, technical services, management of companies and enterprises, administrative and support, and waste management and remediation services sectors
Educational Services	The proportion of workers in educational services sector
Health Care and Social Assistance	The proportion of workers in health care and social assistance sector
Arts, Entertainment, and Recreation	The proportion of workers in arts, entertainment, and recreation sector
Accommodation and Food Services	The proportion of workers in accommodation and food services sector
Other Services	The proportion of workers in other services (except public administration) sector
Public Administration	*The proportion of workers in public administration sector*

a. Italic texts: variables are dropped due to multicollinearity.

Chapter 5: Results

This chapter presents the results of employing the proposed methodological framework by a case study in the Baltimore region, which is the 20th most populated Metropolitan Statistical Area in the United States with over 2.8 million residents [16]. First, the home and work locations are identified based on February 2020 Mobile Device Location Data, considering the outbreak of COVID-19 in March 2020 in the U.S. leads to a dramatic change in human behaviors, especially in people's commuting patterns. The sampling rate for entire residents and workers are calculated to evaluate the spatial consistency of location identification results, and block groups with a higher than 1% sampling rate for both residents and workers are included to reduce the low-sampling biases. Second, the job accessibility calculated from traditional and MDLD-based measures are presented and preliminarily compared in terms of their geographical distribution and basic statistics. Third, the results of GAMs are presented to compare the socio-demographic impact on the general performance of these two job accessibility measures, and specifically, student T-tests are performed to study the socio-demographic characteristics of regions whose job accessibility has major disparities between the two measures.

5.1 Results of Home and Work Location Identification

5.1.1 County-Level Location Identification Results

As the geographic scope of the case study, the Baltimore region includes the jurisdictions of seven counties: Anne Arundel County, Baltimore County, Carroll

County, Harford County, Howard County, Queen Anne's County, and Baltimore City [85]. The home and work identification results are aggregated to the county level, and the sampling rate for residents (number of individuals identified with home locations in each county/ total population reported by 2019 ACS), and for workers (number of workers identified with home locations in each county/number of workers reported by 2019 LODES), are shown in Figure 5-1.

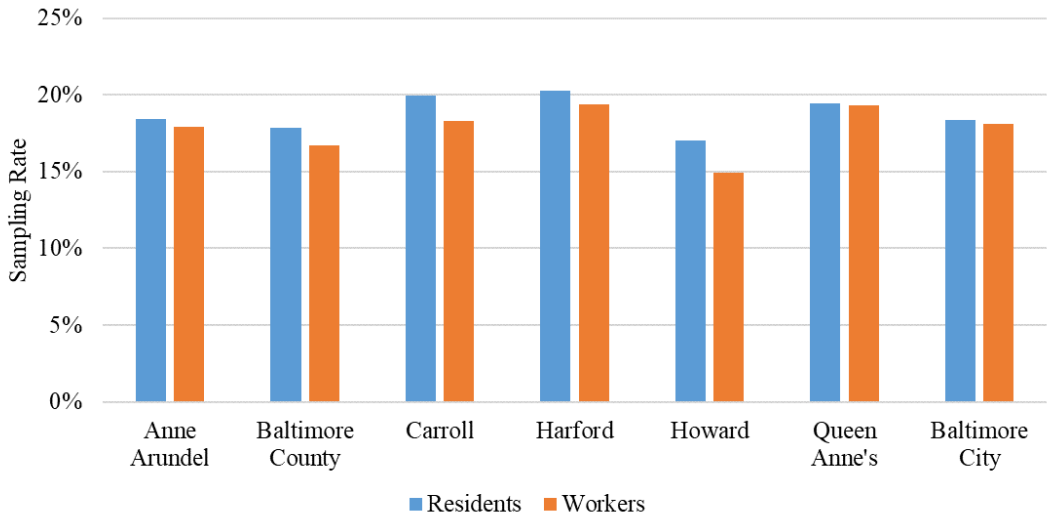


Figure 5-1. County-level sampling rate for residents and workers

As shown, the sampling rates for both residents and workers are all above 15% in seven counties, which represents the high spatial representativeness and consistency at the county level. It can be observed that the sampling rate of residents is always slightly higher than the sampling rate of workers. The reasons might be (1) the total population used to calculate the sampling rate is different, that for residents it is reported by ACS, while for workers it is reported by LODES; (2) the number of workers reported by LODES includes workers who do not have a fixed work location and who work at home, while they are not considered in this study of job accessibility.

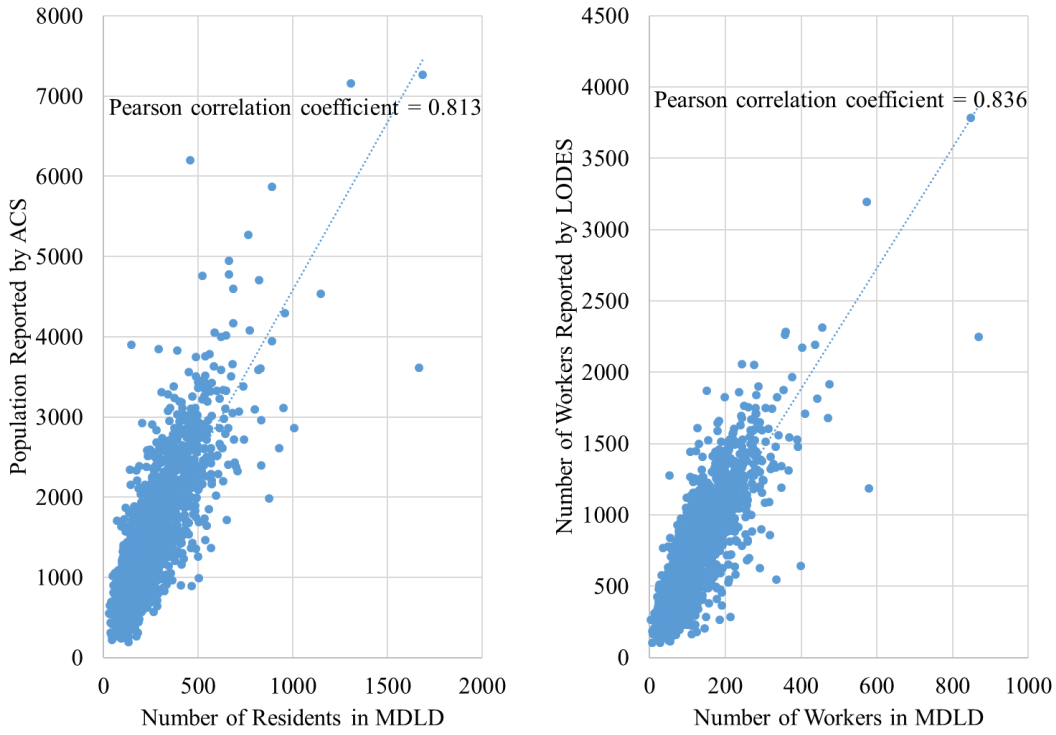
5.1.2 Block Group-Level Location Identification Results

Since both the job accessibility measure and the socio-demographic impact analysis are performed at the census block group-level, similar to the county-level sampling rate, the block group-level sampling rates are also calculated to evaluate the performance of location identification at a finer geographic resolution. The Baltimore region includes 1,945 block groups in total, and 100% of them have workers identified living in there in the MDLD sample, which indicates a wide spatial coverage of MDLD.

Table 5-1 shows the summary statistics of block group-level sampling rates. The average sampling rate for all the residents is 19.36%, and for workers is 17.91%, which indicates high sampling rates for both residents and workers. It also shows a similar pattern to the county-level sampling rate, that the sampling rate of residents is slightly higher than the sampling rate of workers. The 25th percentile of the sampling rates for residents and workers are as high as 14.19% and 13.67% respectively, showing that most of the block groups are identified with a much higher sampling rate than the traditional survey.

Table 5-1. Summary Statistics of Block Group-Level Sampling Rates

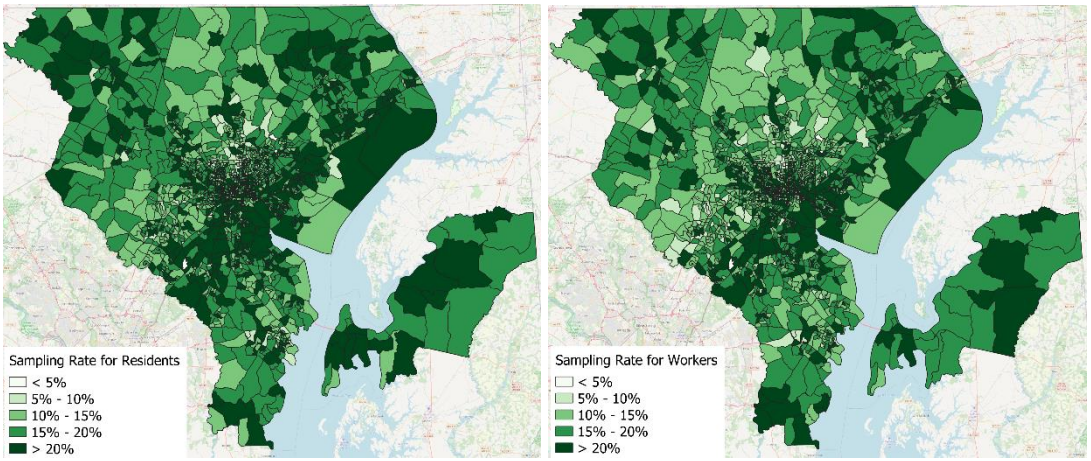
Sampling Rate (%)	Residents	Workers
Mean	19.36	17.91
SD	8.93	8.08
Min.	3.77	1.52
25th percentile	14.19	13.67
50th percentile	17.54	16.63
75th percentile	22.07	20.23
Max.	80.87	98.82



(a) (b)

Figure 5-2. Block Group-Level Distributions (a) Number of Residents from MDLD and ACS, and (b) Number of Workers from MDLD and LODES

The block group-level distributions of residents and workers identified from MDLD are compared with that of the 2019 ACS 5-year population estimates [58] and 2019 LODES7 Residence Area Characteristic data [77] in Figure 5-2. It shows that the residents and worker identification results have similar spatial distributions with both ACS and LODES, which have strong Pearson correlation coefficients of 0.813 and 0.836 respectively.



(a) (b)

Figure 5-3. Block Group-Level Sampling Rate in Baltimore Region for (a) Residents, and (b) Workers

Finally, yet importantly, the spatial distribution of the sampling rate for residents and workers is visualized in Figure 5-3. It can be observed that overall, the block groups in the urban areas, such as Baltimore City, Columbia, and Annapolis, etc., show relatively higher sampling rates for both residents and workers, yet in most of the other areas, the sampling rate is still higher than 10%. The spatial distribution indicates the high geographical coverage and consistency of MDLD.

In summary, the results of home and work location identification are able to verify the robustness of the identification algorithm, and more importantly, the results are sufficient to work as the data foundation for studying the proposed job accessibility measure.

5.2 Results of Job Accessibility

5.2.1 Geographical Distribution of Traditional Job Accessibility

Firstly, the traditional job accessibility, i.e. the cumulative number of jobs in zones within a certain time threshold, is calculated for all the block groups in the Baltimore region. Here, $t_{threshold}$ is set as 15 minutes, 20 minutes, and 30 minutes, respectively, for comparison. Figure 5-4 shows the geographical distribution of the traditional job accessibility results. For a better comparison between different measures with different time thresholds, the results are divided into five groups with every 20^{th} percentile.

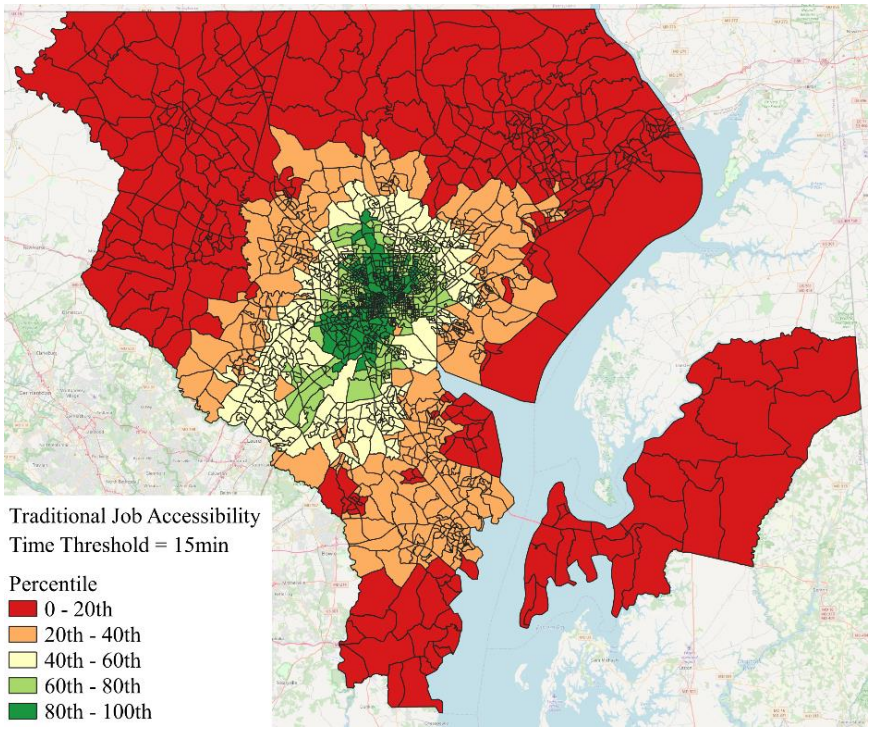
The figures show that with the traditional job accessibility measure, the block groups with high job accessibility are concentrated close to downtown Baltimore City, while almost all of the areas further from downtown Baltimore have low job accessibility. The pattern is simple because the number of jobs in Baltimore City is dominant in this region.

It can also observed that there is a "belt" towards the southwest of Baltimore City, which are block groups along the main highways I-95 and US-1, are showing a relatively higher job accessibility from the traditional measure. It is too hypothetical, since the higher accessibility simply comes from more zones reachable within a certain time threshold from those regions close to highways.

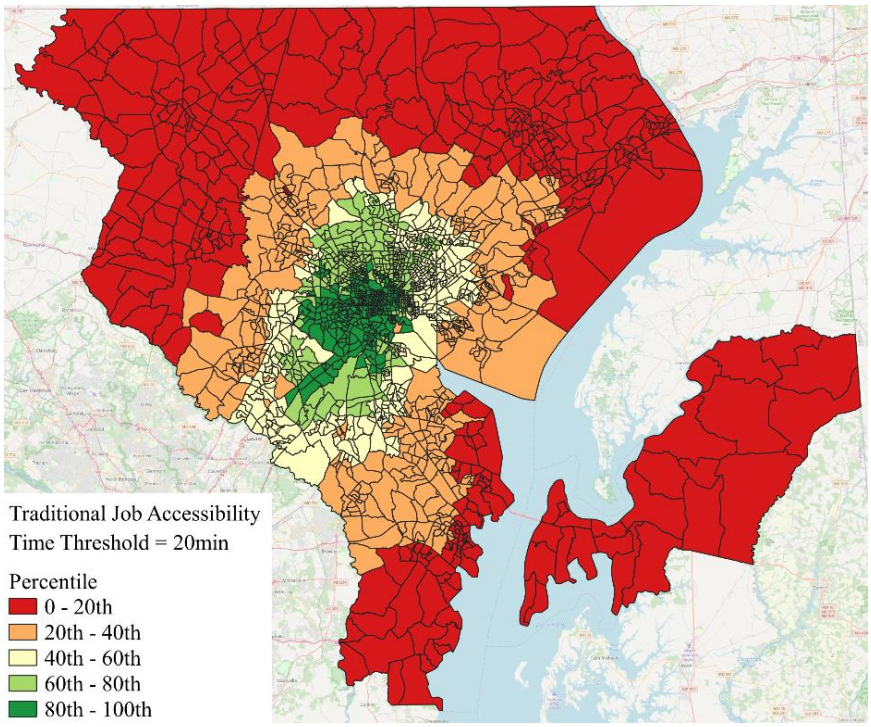
Moreover, based on the horizontal comparison among the three figures, it can be found that with the increase of the time threshold, the high-job-accessibility regions (80^{th} – 100^{th} percentile) move towards the southwest of Baltimore City. The reason is the

greater time threshold makes those regions take into account not only downtown Baltimore City but also more areas in Washington D.C., which is another concentration in terms of the number of jobs, as the reachable zones.

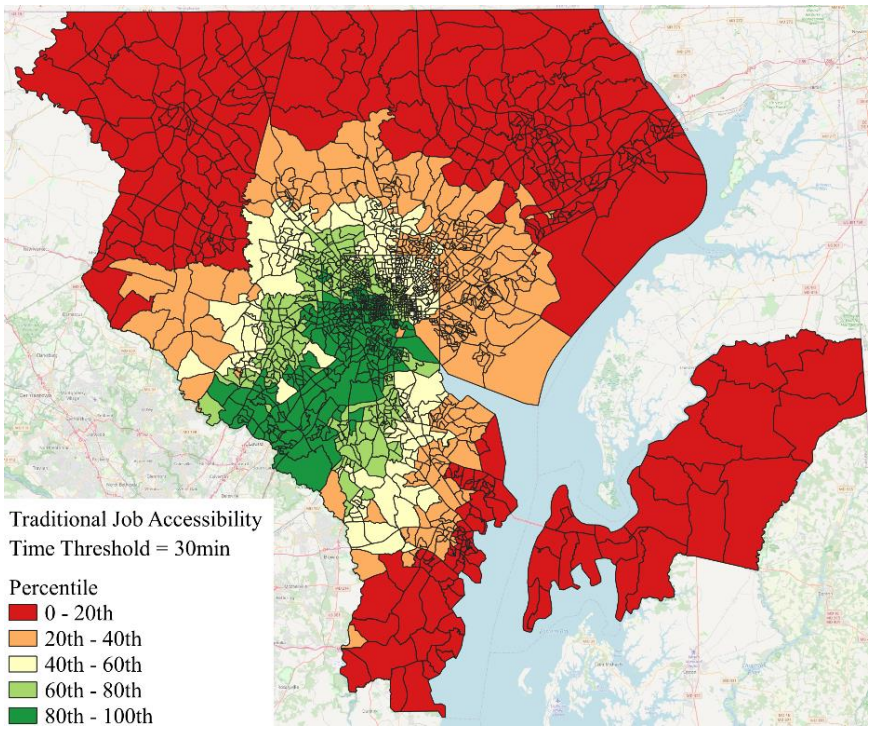
Most importantly, no matter what the time threshold is set to, the traditional job accessibility results always show strong geographical homogeneity, that the zones next to each other always have very similar job accessibility, since the zones reachable within a certain travel impedance are always similar. Thus, those results are hardly able to reveal individual or zonal characteristics, which is the major drawback of the traditional measure.



(a)



(b)



(c)

Figure 5-4. Traditional Job Accessibility for Block Groups in Baltimore Region with time threshold as (a) 15 minutes, (b) 20 minutes, and (c) 30 minutes

5.2.2 Geographical Distribution of MDLD-based Job Accessibility

The MDLD-based job accessibility, i.e. the proportion of workers identified working in zones within a certain time threshold, is also calculated for the same region at block group-level. Similarly, $t_{threshold}$ is also set as 15 minutes, 20 minutes, and 30 minutes, respectively, for comparison. Figure 5-5 shows the results in five groups by every 20th percentile as well. As shown, there are several notable differences compared with the results from the traditional job accessibility measure.

First, block groups close to some smaller well-developed towns with relatively rich job opportunities, such as Westminster and Annapolis, still show relatively high job accessibility, although they are not close to downtown Baltimore. The reason could be these well-developed towns are able to provide sufficient job opportunities for workers living there, thus most of the workers in these regions choose to work nearby.

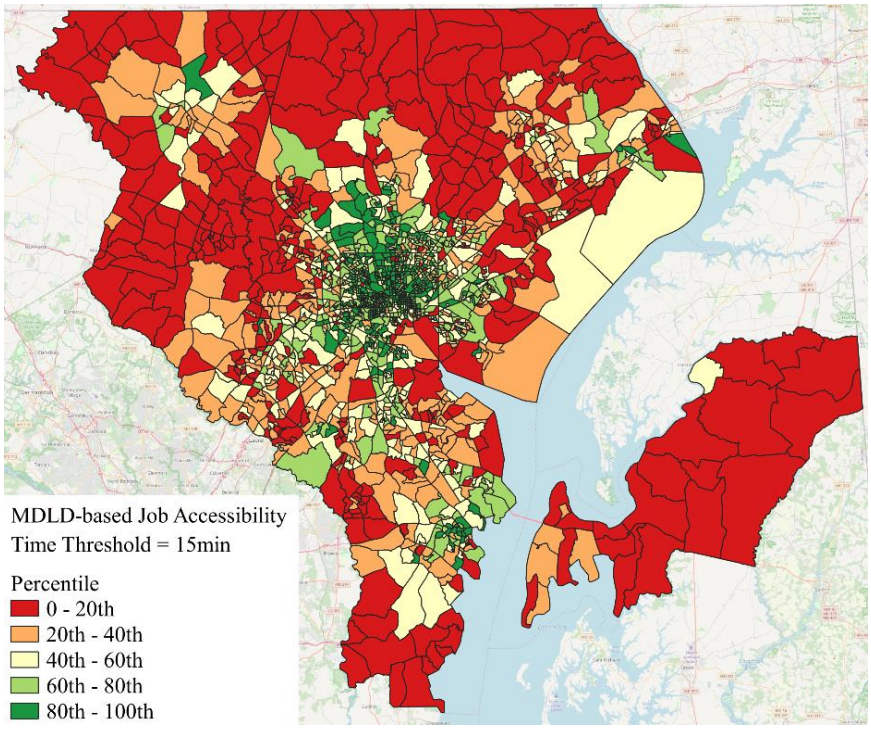
Second, the regions close to main highways are not showing higher job accessibility any more, which implies that there is little correlation between how many zones are reachable within a certain time threshold, and whether people will choose to work in those zones.

Third, with the increase of the time threshold, block groups in the southwest of Baltimore City no longer turn into high-job-accessibility regions, while there is an obvious transformation in the traditional job accessibility results. The reason might be even though the greater time threshold will include more zones in Washington D.C., seldom people who live in the southwest of Baltimore City will actually choose to work

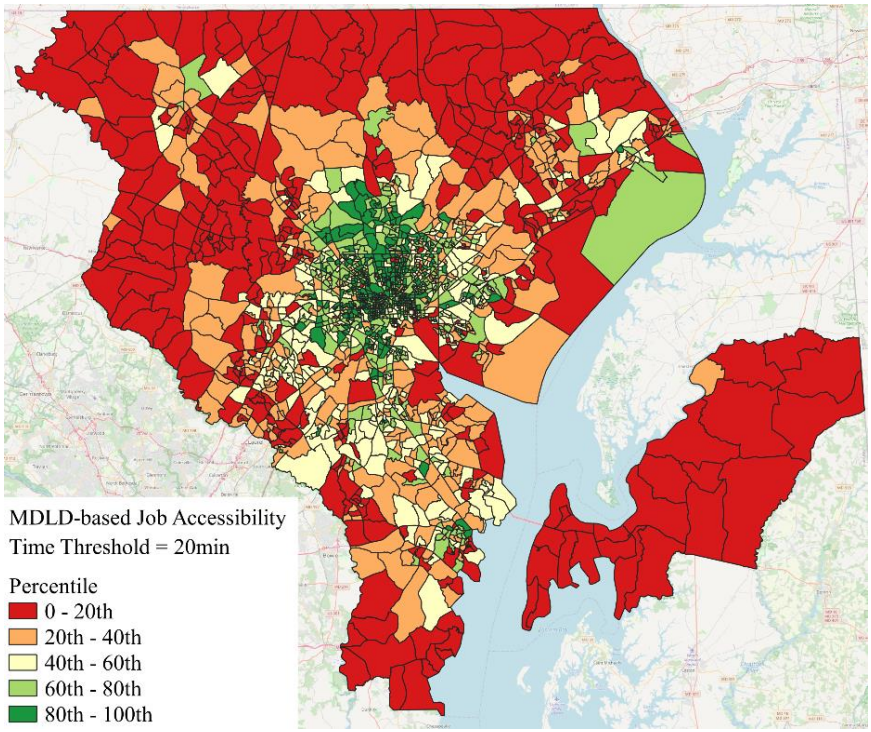
in D.C. Considering the second and the third differences, a further assumption will be that simply expanding the range of the reachable zones will not necessarily improve job accessibility as much as the traditional measure suggest, especially for those areas that are already close to somewhere with a large number of job opportunities, because people naturally tend to choose jobs that are closer, if not considering other complex factors affecting where to work.

Finally, yet importantly, instead of maps with several clear equipotential lines, like what traditional accessibility results show, MDLD-based job accessibility results show geographical heterogeneity in many areas. Specifically, some zones next to the high-job-accessibility regions still show relatively low job accessibility, indicating that social inequity might exist in those areas, which is crucial to the development and evaluation of public policies.

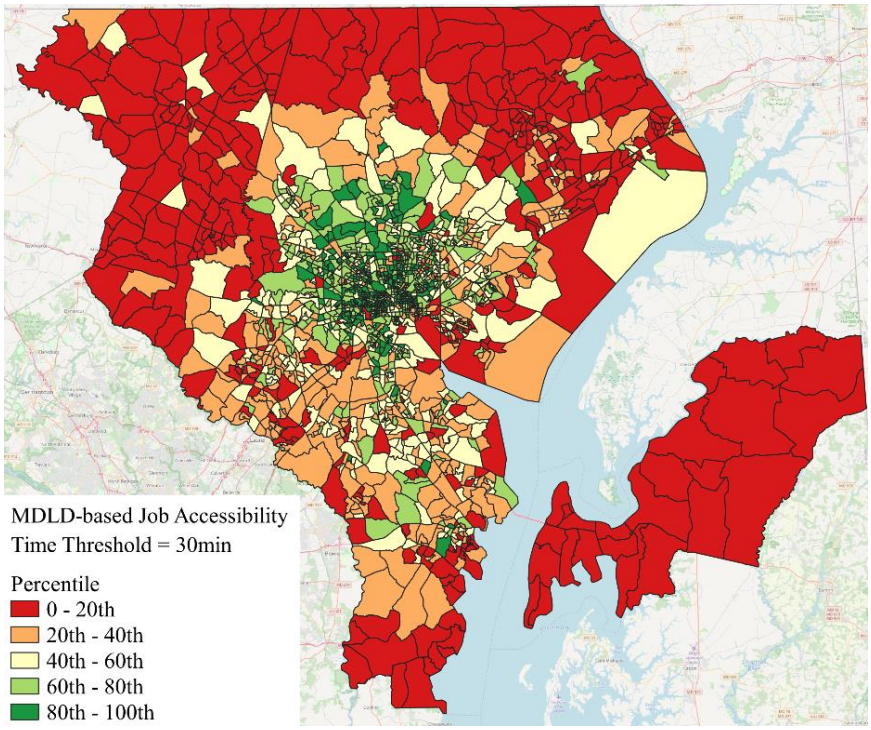
To summarize, compared with traditional measure, the MDLD-based measure provides less hypothetical results which better reveal individual preferences of people living in each zone. The disparities in the results from those two measures is illuminating for people to understand job accessibility from a different perspective. To further investigate the confounding factors affecting the job accessibility calculated from those two measures, results of GAMs and T-tests are demonstrated in section 5.3.


MDLD-based Job Accessibility
Time Threshold = 15min
Percentile
0 - 20th
20th - 40th
40th - 60th
60th - 80th
80th - 100th

(a)


MDLD-based Job Accessibility
Time Threshold = 20min
Percentile
0 - 20th
20th - 40th
40th - 60th
60th - 80th
80th - 100th

(b)



(c)

Figure 5-5. MDLD-based Job Accessibility for Block Groups in Baltimore Region with time threshold as (a) 15 minutes, (b) 20 minutes, and (c) 30 minutes

5.2.3 Statistical Comparison

The job accessibility results are compared statistically as well. Table 5-2 shows the summary statistics of traditional and MDLD-based job accessibility for block groups in the Baltimore region with different time thresholds. As shown, the absolute values of both traditional and MDLD-based job accessibility get higher with the increase of the time threshold, which is intuitive. Another notable statistic is the coefficient of variance (CV), which decreases with the increase of the time threshold, showing a reduction in terms of statistical dispersion for both traditional and MDLD-based job accessibility. Nevertheless, the reasons behind might be different. For the traditional

measure, the overlap of the accessible areas between block groups becomes larger when the time threshold increases. Hence, the proportion of the same job opportunities in those overlapped areas gets higher when calculating the cumulative number of jobs, which makes the variance of the traditional job accessibility decrease. For the MDLD-based measure, there is a fixed upper limit, i.e. when the geographical boundary is farther than the farthest identified work location of the workers living in each zone, the proportion of workers working in zones within the time threshold will be 1. Therefore, the MDLD-based job accessibility will keep approaching 1 as the time threshold increases, and the variance will decrease as well.

Table 5-2. Summary Statistics of Traditional and MDLD-based Job Accessibility for Block Groups in Baltimore Region

Measure	**Traditional**			**MDLD-Based**		
Concept	*Cumulative number of jobs ($\times 10^3$ jobs)*			*Proportion of workers*		
Time Threshold (Minute)	15	20	30	15	20	30
Mean	251.9	427.2	801.6	0.6188	0.7273	0.8554
SD	161.4	238.3	338.2	0.1506	0.1391	0.1026
CV	0.641	0.558	0.421	0.243	0.191	0.120
Min.	1.757	2.394	11.85	0.0278	0.1600	0.3120
25^{th} percentile	88.06	192.0	642.1	0.5223	0.6405	0.8025
50^{th} percentile	251.7	503.1	877.2	0.6226	0.7394	0.8696
75^{th} percentile	417.0	620.2	1063	0.7207	0.8283	0.9314
Max.	554.1	791.6	1330	1.0000	1.0000	1.0000

5.3 Socio-Demographic Impact Analysis

5.3.1 GAM Outputs of Job Accessibility

To reduce the disparities caused by the different choices of time thresholds, the Pearson correlation coefficient between the job accessibility results based on traditional measure and MDLD-based measure are calculated with time thresholds as 15, 20, and 30 minutes, respectively. The results with the time threshold of 20 minutes have the highest correlation coefficient of 0.711, compared with 0.670 when the time threshold is 15 minutes and 0.669 when the time threshold is 30 minutes. Thus, the GAM model analysis will be based on the results with the time threshold of 20 minutes.

Table 5-5 reports the estimation results of the GAMs. The Goodness-of-fit indexes (adjusted R squares) is 0.554 for the MDLD-based job accessibility, and 0.743 for the traditional job accessibility, indicating the data is well fitted by GAMs. The Goodness-of-fit for the MDLD-based job accessibility is lower than that for the traditional job accessibility. The reason might be: the individual preferences for the job will be affected by more complex factors besides the socio-demographics inferable from the dataset, for instance, interpersonal relationships with co-workers, and prospects of his/her future development, which are not able to be captured by the covariates in this study. Thus, the MDLD-based job accessibility is more difficult to predict than the traditional job accessibility by the socio-demographic variables.

For a clearer comparison, the estimation results listed here include the estimated parametric coefficient for the linear fixed effects, effective degrees of freedom for the nonlinear effects, and the significance codes for both linear and nonlinear effects. The

detailed results including the standard error and P-value of each variable are listed in Appendix A. If the P-value is less than 0.1, the variables are considered statistically significant and marked by the significance codes.

Table 5-3. Estimation Results of the Generalized Additive Models for Job Accessibility ($t_{threshold} = 20\ min$)

Dependent Variables	A^{MDLD}		A^{Trad}	
Parametric coefficients:				
	Estimate		**Estimate**	
(Intercept)	3.49E-01	*	-1.12E+05	
Independent Variables				
Block Group-Related				
Population density	9.69E-07		-6.34E-01	
Worker density	3.84E-07		8.47E-02	
Job density	-2.53E-07		2.81E-01	
Urban	9.19E-02	***	1.85E+05	***
Rent	4.70E-02	***	-2.32E+04	
Workers' Socio-Demographic				
African American	-1.33E-01	***	7.59E+04	.
Asian	5.36E-02		4.28E+05	***
Hispanic	-5.41E-02		-3.53E+05	***
Bachelor Degree	3.90E-03		-2.29E+05	
Male	2.51E-01	**	3.13E+05	**
Earning_0_15k	8.14E-02		-2.68E+05	.
Earning_40k_more	-4.53E-01	***	-3.75E+05	***
Age_younger_29	-1.78E-01	*	-1.07E+05	
Age_55_older	2.39E-01	**	-2.61E+05	**
Workers' Industry				
Natural Resources and Mining	-2.31E+00	**	-6.38E+05	
Utilities	9.05E-01		1.65E+06	
Construction	-4.94E-01	*	2.45E+04	
Manufacturing	2.55E-01		7.79E+05	*
Trade	2.95E-01		8.95E+05	***

Transportation and Warehousing	2.88E-01		1.12E+05	
Information	1.10E+00	**	1.88E+06	***
Financial	2.60E-01		-5.67E+04	
Professional and Business Activities	4.45E-01	**	1.11E+06	***
Educational Services	6.53E-01	***	1.05E+06	***
Health Care and Social Assistance	6.03E-01	***	1.34E+05	
Arts, Entertainment, and Recreation	-3.63E-01		7.61E+05	
Accommodation and Food Services	4.89E-01	*	-4.17E+04	
Other Services	-2.70E-02		-8.10E+05	*
Smooth Terms				
	e.d.f.		e.d.f.	
s(County)	5.865	***	5.947	***
Model Fit				
R-sq.(adj)	0.554		0.743	
-REML	-1698.9		23652	

a. Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’0.05 ‘.’ 0.1 ‘ ’ 1.
b. s() refers to a spline function.

Among the block group-related variables, the proportion of urban areas consistently shows a significant positive effect on both MDLD-based and traditional job accessibility, which means the people living in urban areas generally have more job opportunities nearby, and they are able to work closer to home as well. It agrees with what is shown in Figure 5-4 and Figure 5-5, that overall the block groups close to Baltimore City - most of which are urban areas - have higher job accessibility, while most of the rural areas show relatively lower job accessibility for both of the job accessibility results. The disparity occurs in terms of the proportion of renter-occupied housing units, which has a strong positive influence on the MDLD-based job accessibility, while not being significant on the traditional job accessibility. The

MDLD-based measure captures the fact that in this study area, renters are more likely to have shorter commute times, because the convenience of getting to the workplace is one of the most important factors when people choose where to rent, while it is not necessarily the case for where to buy a property. The difference regarding the effectiveness of this variable also shows the advantage of the MDLD-based measure, i.e. the capability of revealing individual preferences.

Two variables related to workers' socio-demographics show consistent influence on both results: the positive effect from the proportion of male workers, which might indicate there is potential gender inequality in terms of job accessibility, and the negative effect from the proportion of workers with earnings greater than 40000$ a year, the reason of which may be that high-income workers prefer to live in the suburban area for the better environment, while jobs with high earnings always concentrated in the urban cities. Most of the other variables related to socio-demographics show different impacts on job accessibility from the two measures, though. The proportion of workers with earnings less than 15000$ a year has a negative impact on the traditional job accessibility, which means workers in the low-income group are less likely to live in the urban cities because of the high cost of living. Considering the negative impact from the high-income group, based on the collinearity, a further inference is mid-income workers are more likely to live in the urban city and have high job accessibility with the traditional measure. The difference in the impact of the race distribution is notable: the proportion of African American workers shows a negative influence on the MDLD-based job accessibility. Alternatively, there is a positive influence on traditional job accessibility, indicating that neighborhoods with

greater portion of African American people have higher traditional job accessibility. A possible explanation may be in this region, a large proportion of the people living in or close to Baltimore City are African American, and as a result, they have high job accessibility based on the traditional measure. However, it is possibly not so easy for them to find jobs in Baltimore City; thus, many of them have to work farther from where they live. Hence, they are more likely to have low job accessibility from the MDLD-based measure. Again, this difference is likely to prove the benefit of the MDLD-based measure: reflecting the actual individual preference. In addition to the African American, the proportion of Asian workers also has a positive influence on the traditional job accessibility, while the proportion of Hispanic workers has a negative influence, indicating they generally live far from Baltimore City in this case study.

Another major disparity occurs in age groups: the proportion of workers age 55 or older has a positive effect on MDLD-based job accessibility, while it has a negative effect on traditional job accessibility. This disparity can be explained by that the elder workers always live far from the urban cities, so they have low job accessibility by the traditional measure. However, they may have the need and the ability to choose a closer job even living in relatively rural areas. Moreover, the proportion of workers age 29 or younger shows a negative effect on MDLD-based job accessibility, indicating that the cost to live close to their workplace might not be affordable to most of the young workers who just started their careers. This difference shown in the modeling results also comes from the individual preference incorporated in the MDLD-based measure.

Regarding the workers' industry, there is no variable showing a completely opposite effect, i.e. significantly positive on one and negative on the other. The proportion of workers in information, professional, scientific, technical services, management of companies and enterprises, administrative and support, waste management and remediation services, and educational services sectors presents a significantly positive impact on both MDLD-based and traditional job accessibility, indicating workers in these industries always live in urban areas where these jobs are also concentrated. Besides, the proportion of workers in the health care and social assistance sectors also has a positive effect on MDLD-based job accessibility. On the other hand, the proportion of workers in agriculture, forestry, fishing, hunting, mining, quarrying, oil and gas extraction, and construction sectors have a negative effect on MDLD-based job accessibility. In other words, workers in these industries always have to travel longer to their work. Finally, The proportion of workers in manufacturing, wholesale trade, and retail trade has a positive impact on the traditional job accessibility, as the density of these industries is always higher in cities.

As for the nonlinear effects across counties, the estimated degrees of freedom (e.d.f.), which is 5.865 and 5.947 respectively for MDLD-based and traditional job accessibility are both largely greater than 1.0, suggesting the substantial unobserved heterogeneity among counties have been well captured by the nonlinear spline functions in GAMs.

5.3.2 T-tests of block groups with major disparities

In addition to the above analysis of the socio-demographic impact on the general performance of both job accessibility measures, researchers and policymakers may also

be interested in the regions whose job accessibility has major disparities between the traditional measure and the MDLD-based measure. For example, what kind of regions have job accessibility significantly overestimated or underestimated by the traditional measure, compared with the job accessibility from the MDLD-based measure? In this study, the traditional and MDLD-based job accessibility results are firstly converted to the percentile in the study area. Then, the "regions with traditionally overestimated job accessibility" is defined as the block groups whose percentile of traditional job accessibility is 33.3% or higher than their percentile of MDLD-based job accessibility, and the "regions with traditionally underestimated job accessibility" is defined as the block groups whose percentile of traditional job accessibility is 33.3% or lower than their percentile of MDLD-based job accessibility.

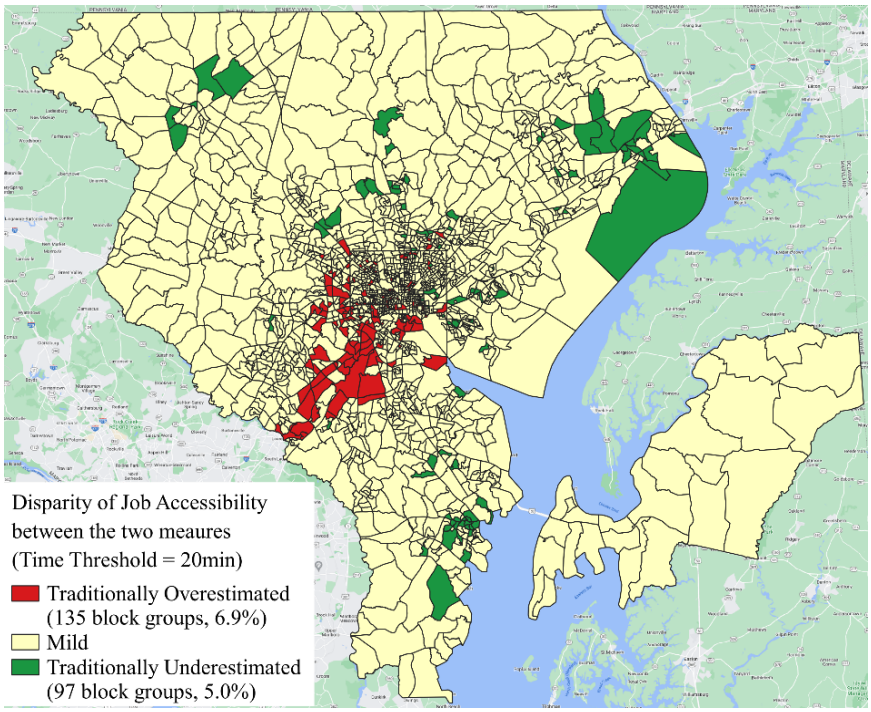


Figure 5-6. Geographical Distribution of Regions with Traditionally Overestimated and Underestimated Job Accessibility

Figure 5-6 shows the geographical distribution of regions with traditionally overestimated and underestimated job accessibility. There are 135 block groups classified into regions with traditionally overestimated job accessibility, which is 6.9% of the total number of block groups; and 97 block groups are classified into regions with traditionally underestimated job accessibility, which is 5.0% of the total number of block groups. Consistent with the comparison between Figure 5-4 and Figure 5-5, the major overestimation by the traditional job accessibility measure happens in those block groups located in the southeast of Baltimore City. Being close to Baltimore City leads to high job accessibility from the traditional measure, while the revealed low proportion of short-commute-time workers leads to low job accessibility from the MDLD-based measure. Also similarly to the findings in section 5.2.2, some block groups close to several well-developed towns, such as Westminster and Annapolis, show major underestimation by the traditional job accessibility measure, as they are far from the big city, but revealing a high proportion of short-commute-time workers.

Table 5-4. T-test results of block groups with major overestimation and underestimation of job accessibility with traditional measure ($t_{threshold} = 20\ min$)

		Traditionally Overestimated		**Traditionally Underestimated**	
	Mean	**Variation**		**Variation**	
Dependent Variables					
A^{MDLD}	7.27E-01	-6.26E-02	***	1.07E-01	***
A^{Trad}	4.27E+05	2.37E+05	***	-1.73E+05	***
Independent Variables					
Block Group-Related					
Population density	7.03E+03	1.21E+03	.	-1.68E+03	**
Worker density	3.16E+03	5.72E+02	.	-8.38E+02	**
Job density	2.27E+03	3.42E+03	.	-3.00E+02	

Urban	8.58E-01	1.13E-01	***	-6.53E-02	.
Rent	3.35E-01	5.29E-02	*	6.65E-02	*
Workers' Socio-Demographic					
White	5.85E-01	-1.25E-01	***	1.14E-01	***
African American	3.45E-01	1.14E-01	***	-1.07E-01	***
Asian	4.87E-02	1.00E-02	.	-8.57E-03	*
Hispanic	4.58E-02	7.81E-05		1.18E-02	**
Bachelor Degree	2.47E-01	-2.09E-02	***	6.53E-03	
Male	4.78E-01	-9.07E-04		7.38E-03	.
Earning_0_15k	2.03E-01	3.41E-04		-1.38E-03	
Earning_15_40k	2.86E-01	2.31E-02	*	-1.03E-02	
Earning_40k_more	5.10E-01	-2.34E-02	.	1.17E-02	
Age_younger_29	2.23E-01	1.63E-02	*	-1.56E-03	
Age_30_54	5.26E-01	4.94E-03		-1.09E-02	**
Age_55_older	2.51E-01	-2.12E-02	***	1.25E-02	*
Workers' Industry					
Natural Resources and Mining	1.93E-03	-9.39E-04	***	2.80E-04	
Utilities	3.62E-03	-1.94E-05		9.20E-06	
Construction	5.45E-02	-3.45E-03		6.85E-03	**
Manufacturing	4.03E-02	-1.32E-03		3.32E-03	.
Trade	1.32E-01	-3.94E-03		7.10E-03	**
Transportation and Warehousing	4.16E-02	5.95E-03	***	-1.11E-03	
Information	1.71E-02	-3.80E-04		1.68E-04	
Financial	5.39E-02	-3.74E-03	*	8.92E-04	
Professional and Business Activities	1.76E-01	7.02E-03	**	-4.33E-03	
Educational Services	1.07E-01	-3.41E-03		-7.56E-03	**
Health Care and Social Assistance	1.63E-01	4.66E-03		-1.13E-02	**
Arts, Entertainment, and Recreation	1.75E-02	8.21E-05		-3.43E-04	
Accommodation and Food Services	8.37E-02	-2.41E-03		6.64E-03	***
Other Services	3.55E-02	-2.51E-03	**	5.28E-03	***

Public Administration	7.26E-02	4.41E-03	.	-5.88E-03	**

a. Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’0.05 ‘.’ 0.1 ‘ ’ 1.

To quantitatively compare the differences in the mean of the socio-demographic variables from the total population in the study area, the one-sample T-tests are conducted respectively for block groups with traditionally overestimated and underestimated job accessibility. The T-test results are presented in Table 5-4. The positive values in the Variance column suggest that the mean of those independent variables in those block groups with traditionally overestimated or underestimated job accessibility are larger than the mean of the entire population in the study area, while the negative values mean smaller. Based on the variance and the significance codes, the profiles of these two groups can be drawn: those regions located in urban areas with high population density and job density, where many renters live, and workers lived there are more likely to be African American and Asian, and younger than 29 years old, but not White and not older than 55, without a Bachelor or advanced degree, earning between 15000-40000 a year, and more likely to work in transportation and warehousing, professional and business activities, and public administration industries, the job accessibility of these regions have higher chance to be overestimated by the traditional job accessibility measure. On the other side, those regions located in rural areas with low population density, where there are also many renters, and workers living there are more likely to be white and Hispanic males, aged older than 55, while less likely to be African American and Asian, and they are more likely to work in industries of construction, wholesale and retail trade, accommodation and food services, but less likely to work in educational services, healthcare, and social

assistance, and public administration industries, the job accessibility of these regions have a higher chance to be underestimated by the traditional measure.

In summary, compared with the traditional job accessibility measure, the MDLD-based measure, which reveals the actual individual preferences, is able to better capture the disparities of job accessibility between different socio-demographic groups, and it may help researchers and policymakers on the awareness and understanding of social inequality.

Chapter 6: Conclusion and Discussion

6.1 Research Summary

This study evaluates the state-of-the-practice accessibility measures as well as the state-of-the-art methods based on MDLD data. The key research gap is identified from the literature review, and an MDLD-based measure is proposed to evaluate the job accessibility based on the identified home and work locations. The socio-demographic impact of job accessibility from the MDLD-based measure is also analyzed, and compared with the results from a traditional job accessibility measure.

First, a behavior-based method is employed to identify home and work locations with only MDLD information, at the geographic scale of the Baltimore Region, which is the subject area of the case study. The location identification results are validated against the reported data from ACS and LODES, which reach a high average sampling rate and high Pearson correlation coefficient with the validation data for both residents and workers at the block group-level.

Second, the block group-level job accessibility is calculated from the proposed MDLD-based measure, as well as a traditional cumulative opportunity measure, and the results are compared at a different set of time thresholds. A noticeable spatial heterogeneity is found in the results from the MDLD-based measure, in comparison to the severe spatial homogeneity shown in the results from the traditional measure, which demonstrated the major advantage of the MDLD-based measure, the capability of revealing individual preference.

Then, a generalized additive model is built, where not only various kinds of features with fixed effects but also the random effects across different counties are considered, to evaluate the socio-demographic effect on job accessibility from the MDLD-based and traditional measures. The data is well fitted by the GAMs, and a feature-by-feature comparison shows that the socio-demographic effects are different between the MDLD-based and traditional job accessibility. Some features especially feature related to racial distribution, show counterintuitive relationships with job accessibility by traditional measure, while the results of MDLD-based job accessibility are more consistent with people's knowledge. In addition, for those regions whose job accessibility has major disparities between the MDLD-based and traditional measure, T-tests are conducted respectively and results indicate that the socio-demographic characteristics of the regions with traditionally overestimated and underestimated job accessibility are almost completely opposite, which indicate that for some zones under certain conditions, the job accessibility from the traditional measure is significantly not in agreement with people's individual preferences. Social inequality problems may exist especially in those regions whose job accessibility is traditionally overestimated.

6.2 Research Contribution

The main contributions of this study can be summarized into three folds: (1) methodology and comprehensive comparison process; (2) uniqueness of the data; (3) application potential.

6.2.1 Methodology and Comparison

This study examines state-of-the-practice accessibility measures as well as the state-of-the-art methods for processing MDLD data. Based on the literature review, this study proposes a novel job accessibility measure based on the imputed home and work locations from MDLD, which is more capable of capturing people's individual preferences than the traditional measure.

The comparison of the geographic distribution of job accessibility in Baltimore region show that the results from the MDLD-based measure is more granular, which present a strong spatial heterogeneity with distinctions within neighboring communities, better reflecting worker's actual destination for work than the results from traditional measure.

The comparison of the GAM results shows that there are significant differences in the socio-demographic effects on the job accessibility measured by the MDLD-based and traditional measures, in terms of the proportion of renters, race, age and earning distributions, and workers' industries, suggesting that people who live in areas with large number of job occupations nearby, and people who actually work in those areas, could be two different subgroups of the population. Moreover, the T-tests results present clear socio-demographic characteristics of regions with considerable discrepancies in job accessibility between the traditional measure and the MDLD-based measure, suggesting that the job accessibility of people with some certain socio-demographic features have higher chance to be strongly overestimated by the traditional measure, and with some other socio-demographic features, people's job

accessibility is more likely to be underestimated by the traditional measure, indicating that there could be underlining social inequality problems on these population groups.

In summary, the proposed MDLD-based job accessibility measure can better reveal people's individual preferences, and the comparison against the results from traditional measure indicates that for many regions and population groups, the job accessibility from the traditional measure is not consistent with people's actual destination for work revealed by the MDLD-based measure.

6.2.2 Uniqueness of the Data

This thesis presents the first study that utilizes high-quality mobile device location data with wide spatial and temporal coverages to systematically study accessibility. The MDLD used in this study is obtained from multiple leading providers of passively collected passenger travel data, covering over 15% of the U.S. population. The case study is based on data from Feb 2020, while the original dataset has a temporal coverage of the entire 2019, 2020, and 2021, which provides great potential for the longitudinal comparison.

In addition to the MDLD, this study also utilized skim matrices produced by a statewide activity-based travel demand model MSTM2 to measure the zone-to-zone travel time as the impedance, for the higher estimation accuracy.

6.2.3 Application Potential

The new measure proposed in this study is developed using a real-world mobile device location dataset and applied to the Baltimore region for case study purposes. Although

a tremendous amount of MDLD is processed, the methodological framework achieves high computational efficiency and the proposed measure keeps high interpretability, which is important to both researchers and policymakers. The MDLD and the publicly available data used to generalize the traditional measure can both be expanded to the national population.

6.3 Discussion and Future Research Directions

With the individual preference information included, the MDLD-based accessibility measure shows the capability of more precisely identifying the accessibility across regions with different characteristics, compared with the state-of-the-practice accessibility measure. It is worth mentioning that the goal of the proposed MDLD-based measure is not to fully deny the performance of the traditional measure. The traditional measure is still valuable on a larger scale, as it is still true that overall, the places with a greater number of opportunities have higher accessibility, which is supported by the positive correlation between the results from the MDLD-based measure and the traditional measure. The value of the MDLD-based measure is allowing researchers and policymakers to evaluate the accessibility from a different point of view, especially for those regions where some social inequity problems are covered by the high accessibility from the traditional measure, and neglected by people as a result.

The limitation of this study and the correlated future research directions may come from the following aspects: (1) the mobile device location data used in the study might not be able to precisely represent the population-level behavior, since the sampling rate

varies at finer geographic resolution such as block group-level, also the natural sample bias exists in MDLD. Additional weighting and validation can be further conducted to better address this issue; (2) the traditional measure compared in this study, although widely used, is relatively simple. The other more complex measures could be evaluated and compared since they leverage richer information and might have better performance than the cumulative opportunity; (3) the study is conducted on a regional scale, while the accessibility is affected by geographic locations and regional social, humanity, and economic situation. The proposed methodology and the comparison process can be applied on a larger scale such as statewide and nationwide levels, to evaluate the heterogeneity across different regions.

Appendices

Appendix A. Detailed Results of the Generalized Additive Models for MDLD-based Job Accessibility ($t_{threshold} = 20\ min$)

Dependent Variables	A^{MB}			
Parametric coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	3.49E-01	1.71E-01	2.0370	4.18E-02
Independent Variables				
Block Group-Related				
Population density	9.69E-07	7.93E-07	1.2226	2.22E-01
Worker density	3.84E-07	1.84E-06	0.2087	8.35E-01
Job density	-2.53E-07	2.34E-07	-1.0815	2.80E-01
Urban	9.19E-02	1.02E-02	9.0397	3.96E-19
Rent	4.70E-02	1.15E-02	4.0810	4.68E-05
Workers' Socio-Demographic				
African American	-1.33E-01	3.10E-02	-4.2914	1.87E-05
Asian	5.36E-02	6.44E-02	0.8323	4.05E-01
Hispanic	-5.41E-02	8.02E-02	-0.6741	5.00E-01
Bachelor Degree	3.90E-03	1.19E-01	0.0329	9.74E-01
Male	2.51E-01	8.63E-02	2.9095	3.67E-03
Earning_0_15k	8.14E-02	1.22E-01	0.6686	5.04E-01
Earning_40k_more	-4.53E-01	7.58E-02	-5.9798	2.69E-09
Age_younger_29	-1.78E-01	9.65E-02	-1.8427	6.55E-02
Age_55_older	2.39E-01	7.40E-02	3.2215	1.30E-03
Workers' Industry				
Natural Resources and Mining	7.50E-01	-3.0793	2.11E-03	7.50E-01
Utilities	7.84E-01	1.1544	2.49E-01	7.84E-01
Construction	2.32E-01	-2.1316	3.32E-02	2.32E-01
Manufacturing	2.44E-01	1.0434	2.97E-01	2.44E-01
Trade	1.84E-01	1.6035	1.09E-01	1.84E-01
Transportation and Warehousing	2.26E-01	1.2718	2.04E-01	2.26E-01
Information	3.93E-01	2.7983	5.19E-03	3.93E-01
Financial	2.27E-01	1.1419	2.54E-01	2.27E-01
Professional and Business Activities	1.64E-01	2.7110	6.77E-03	1.64E-01
Educational Services	1.75E-01	3.7420	1.88E-04	1.75E-01
Health Care and Social Assistance	1.70E-01	3.5432	4.05E-04	1.70E-01
Arts, Entertainment, and Recreation	3.91E-01	-0.9296	3.53E-01	3.91E-01
Accommodation and Food Services	2.17E-01	2.2500	2.46E-02	2.17E-01
Other Services	2.90E-01	-0.0932	9.26E-01	2.90E-01
Smooth Terms				
	e.d.f.	**Ref.df**	**F**	**p-value**
s(County)	5.865	6	45.35	<2e-16
Model Fit				
R-sq.(adj) = 0.554		Deviance explained = 56.2%		
-REML = -1698.9		Scale est. = 0.0083763		

Appendix B. Detailed Results of the Generalized Additive Models for Traditional Job Accessibility ($t_{threshold} = 20\ min$)

Dependent Variables		A^{CJ}		
Parametric coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-1.12E+05	2.30E+05	-0.4874	6.26E-01
Independent Variables				
Block Group-Related				
Population density	-6.34E-01	1.04E+00	-0.6079	5.43E-01
Worker density	8.47E-02	2.42E+00	0.0350	9.72E-01
Job density	2.81E-01	3.08E-01	0.9128	3.61E-01
Urban	1.85E+05	1.34E+04	13.7867	3.57E-41
Rent	-2.32E+04	1.52E+04	-1.5279	1.27E-01
Workers' Socio-Demographic				
African American	7.59E+04	4.08E+04	1.8630	6.26E-02
Asian	4.28E+05	8.49E+04	5.0456	4.98E-07
Hispanic	-3.53E+05	1.06E+05	-3.3481	8.30E-04
Bachelor Degree	-2.29E+05	1.56E+05	-1.4677	1.42E-01
Male	3.13E+05	1.14E+05	2.7526	5.97E-03
Earning_0_15k	-2.68E+05	1.60E+05	-1.6763	9.39E-02
Earning_40k_more	-3.75E+05	9.98E+04	-3.7560	1.78E-04
Age_younger_29	-1.07E+05	1.27E+05	-0.8397	4.01E-01
Age_55_older	-2.61E+05	9.74E+04	-2.6804	7.42E-03
Workers' Industry				
Natural Resources and Mining	-6.38E+05	9.88E+05	-0.6455	5.19E-01
Utilities	1.65E+06	1.03E+06	1.5959	1.11E-01
Construction	2.45E+04	3.05E+05	0.0803	9.36E-01
Manufacturing	7.79E+05	3.21E+05	2.4253	1.54E-02
Trade	8.95E+05	2.42E+05	3.6913	2.30E-04
Transportation and Warehousing	1.12E+05	2.98E+05	0.3751	7.08E-01
Information	1.88E+06	5.17E+05	3.6449	2.75E-04
Financial	-5.67E+04	3.00E+05	-0.1891	8.50E-01
Professional and Business Activities	1.11E+06	2.16E+05	5.1337	3.15E-07
Educational Services	1.05E+06	2.30E+05	4.5887	4.77E-06
Health Care and Social Assistance	1.34E+05	2.24E+05	0.6002	5.48E-01
Arts, Entertainment, and Recreation	7.61E+05	5.14E+05	1.4792	1.39E-01
Accommodation and Food Services	-4.17E+04	2.86E+05	-0.1458	8.84E-01
Other Services	-8.10E+05	3.81E+05	-2.1251	3.37E-02
Smooth Terms				
	e.d.f.	**Ref.df**	**F**	**p-value**
s(County)	5.947	6	160.5	<2e-16
Model Fit				
R-sq.(adj) = 0.743		Deviance explained = 74.8%		
-REML = 23652		Scale est. = 1.4501e+10		

References

[1] Litman, T. (2017). Evaluating accessibility for transport planning. Victoria, BC, Canada: Victoria Transport Policy Institute.

[2] Guzman, L. A., Oviedo, D., & Rivera, C. (2017). Assessing equity in transport accessibility to work and study: The Bogotá region. Journal of Transport Geography, 58, 236-246.

[3] Ohnmacht, T., Maksim, H., & Bergman, M. M. (Eds.). (2009). Mobilities and inequality. Ashgate Publishing, Ltd..

[4] Kain, J. F. (1968). Housing segregation, negro employment, and metropolitan decentralization. The quarterly journal of economics, 82(2), 175-197.

[5] Östh, J. (2011). Introducing a method for the computation of doubly constrained accessibility models in larger datasets. Networks and Spatial Economics, 11(4), 581-620.

[6] Shen, Q. (1998). Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers. Environment and planning B: Planning and Design, 25(3), 345-365.

[7] Hansen, W. G. (1959). How accessibility shapes land use. Journal of the American Institute of planners, 25(2), 73-76.

[8] Van Wee, B. (2016). Accessible accessibility research challenges. Journal of transport geography, 51, 9-16.

[9] Niemeier, D. A. (1997). Accessibility: an evaluation using consumer welfare. Transportation, 24(4), 377-396.

[10] García-Albertos, P., Picornell, M., Salas-Olmedo, M. H., & Gutiérrez, J. (2019). Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. Transportation Research Part A: Policy and Practice, 125, 294-307.

[11] Boisjoly, G., & El-Geneidy, A. M. (2017). The insider: A planners' perspective on accessibility. Journal of Transport Geography, 64, 33-43.

[12] Geurs, K. T., & Van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport geography, 12(2), 127-140.

[13] Deboosere, R., & El-Geneidy, A. (2018). Evaluating equity and accessibility to jobs by public transport across Canada. Journal of Transport Geography, 73, 54-63.

[14] Ghader, S., Yang, D., Darzi, A., Assadabadi, A., Carrion, C., Rossi, T. F., ... & Zhang, L. (2019). Integrating Micro-Simulation Models of Short-Distance and Long-Distance Trips for Statewide Applications (No. 19-05637).

[15] Pan, Y. (2021). National-Level Origin-Destination Estimation Based on Passively Collected Location Data and Machine Learning Methods (Doctoral dissertation).

[16] "2020 Population and Housing State Data". United States Census Bureau, Population Division. August 12, 2021.

[17] Lucas, K., Van Wee, B., & Maat, K. (2016). A method to evaluate equitable accessibility: combining ethical theories and accessibility-based approaches. Transportation, 43(3), 473-490.
[18] Litman, T. (2017). Evaluating transportation equity. Victoria Transport Policy Institute.
[19] Titheridge, H., Mackett, R. L., Christie, N., Oviedo Hernández, D., & Ye, R. (2014). Transport and poverty: a review of the evidence.
[20] Burns, L. D., & Golob, T. F. (1976). The role of accessibility in basic transportation choice behavior. Transportation, 5(2), 175-198.
[21] Bocarejo S, J. P., & Oviedo H, D. R. (2012). Transport accessibility and social inequities: a tool for identification of mobility needs and evaluation of transport investments. Journal of transport geography, 24, 142-154.
[22] Halden, D., Mcguigan, D., Nisbet, A., & Mckinnon, A. (2000). Accessibility: Review of measuring techniques and their application. Edinburgh, UK: Great Britain, Scottish Executive, Central Research Unit.
[23] Black, J., & Conroy, M. (1977). Accessibility measures and the social evaluation of urban structure. Environment and Planning A, 9(9), 1013-1031.
[24] Cohen, D. S., & Basner, C. (1972). Accessibility--its Use as an Evaluation Criterion in Testing and Evaluating Alternative Transportation Systems. US Department of Transportation, Federal Highway Administration, Office of Highway Planning.
[25] Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. Journal of Transport Geography, 25, 141-153.
[26] Linneker B J, Spence N A. Accessibility measures compared in an analysis of the impact of the M25 London Orbital Motorway on Britain[J]. Environment and Planning A, 1992, 24(8): 1137-1154.
[27] Farber, S., & Fu, L. (2017). Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis) investments over time. Computers, Environment and Urban Systems, 62, 30-40.
[28] Wachs, M., & Kumagai, T. G. (1973). Physical accessibility as a social indicator. Socio-Economic Planning Sciences, 7(5), 437-456.
[29] Wickstrom, G. V. (1971). Defining balanced transportation-a question of opportunity. Traffic quarterly, 25(3).
[30] Järv, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. Applied geography, 95, 101-110.
[31] Horner, M. W., & Mascarenhas, A. K. (2007). Analyzing location‐based accessibility to dental services: an Ohio case study. Journal of public health dentistry, 67(2), 113-118.
[32] Recker, W. W., Chen, C., & McNally, M. G. (2001). Measuring the impact of efficient household travel decisions on potential travel time savings and accessibility gains. Transportation Research Part A: Policy and Practice, 35(4), 339-369.
[33] Charleux, L. (2015). A GIS Toolbox for Measuring and Mapping Person‐Based Space‐Time Accessibility. Transactions in GIS, 19(2), 262-278.

[34] Fransen, K., & Farber, S. (2019). Using person-based accessibility measures to assess the equity of transport systems. In Measuring transport equity (pp. 57-72). Elsevier.
[35] Gulhan, G., Ceylan, H., Özuysal, M., & Ceylan, H. (2013). Impact of utility-based accessibility measures on urban public transportation planning: A case study of Denizli, Turkey. Cities, 32, 102-112.
[36] Nassir, N., Hickman, M., Malekzadeh, A., & Irannezhad, E. (2016). A utility-based travel impedance measure for public transit network accessibility. Transportation Research Part A: Policy and Practice, 88, 26-39.
[37] Vickerman RW. Accessibility, attraction, and potential: a review of some concepts and their 22 use in determining mobility. Environment and Planning A.
[38] Handy, S. L. (1992). Regional versus local accessibility: neo-traditional development and its implications for non-work travel. Built Environment (1978-), 253-267.
[39] Hanson, S., & Schwab, M. (1987). Accessibility and intraurban travel. Environment and planning A, 19(6), 735-748.
[40] Cascetta, E., Cartenì, A., & Montanino, M. (2013). A new measure of accessibility based on perceived opportunities. Procedia-Social and Behavioral Sciences, 87, 117-132.
[41] Levinson, D. M. (1998). Accessibility and the journey to work. Journal of transport geography, 6(1), 11-21.
[42] Hernandez, D. (2018). Uneven mobilities, uneven opportunities: Social distribution of public transport accessibility to jobs and education in Montevideo. Journal of Transport Geography, 67, 119-125.
[43] Owen, A., & Levinson, D. M. (2015). Modeling the commute mode share of transit using continuous accessibility to jobs. Transportation research part A: policy and practice, 74, 110-122.
[44] Hu L. Job accessibility of the poor in Los Angeles: Has suburbanization affected spatial mismatch?. Journal of the American Planning Association. 2015 Jan 2;81(1):30-45.
[45] Wagner, D., Neumeister, D., & Murakami, E. (1997). Global Positioning Systems for Personal Travel Surveys: Lexington Area Travel Data Collection Test: Appendixes.
[46] Ojah, M., & Pearson, D. (2008). Austin/San Antonio GPS-Enhanced Household Travel Survey,". Texas Transportation Institute.
[47] Wolf, J., & Lee, M. (2008, May). Synthesis of and statistics for recent GPS-enhanced travel surveys. In Paper submitted to the Eighth Int. Conf. Survey Methods in Transport: Harmonization and Data Comparability, Annecy, France.
[48] Frank, P. (2010). Chicago regional household travel inventory. Chicago Metropolitan Agency for Planning.
[49] NuStats, P. T. V. (2011). Regional travel survey: final report. Atlanta Regional Commission, Atlanta.
[50] Safi, H., Assemi, B., Mesbah, M., Ferreira, L., & Hickman, M. (2015). Design and implementation of a smartphone-based travel survey. Transportation Research Record, 2526(1), 99-107.

[51] Haghani, A., Hamedi, M., & Sadabadi, K. F. (2009). I-95 Corridor coalition vehicle probe project: Validation of INRIX data. I-95 Corridor Coalition, 9.
[52] Schrank, D., Eisele, B., & Lomax, T. (2015). 2014 Urban mobility report: powered by Inrix Traffic Data (No. SWUTC/15/161302-1).
[53] Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies, 68, 285-299.
[54] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. nature, 453(7196), 779-782.
[55] Kang, C., Liu, Y., Ma, X., & Wu, L. (2012). Towards estimating urban population distributions from mobile call data. Journal of Urban Technology, 19(4), 3-21.
[56] Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies, 26, 301-313.
[57] Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J. (2019). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. Transportation Research Part C: Emerging Technologies, 105, 183-202.
[58] U.S. Census Bureau. (2022). LEHD Origin-Destination Employment Statistics Data (2019). Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program, accessed on February 2022 at https://lehd.ces.census.gov/data/#lodes. LODES 7.5.
[59] Graham, M. R., Kutzbach, M. J., & McKenzie, B. (2014). Design comparison of LODES and ACS commuting data products (No. 14-38).
[60] Leber, J. (2013). How wireless carriers are monetizing your movements. MIT Technology Rev.
[61] Riederer, C. J., Zimmeck, S., Phanord, C., Chaintreau, A., & Bellovin, S. M. (2015, November). " I don't have a photograph, but you can have my footprints." Revealing the Demographics of Location Data. In Proceedings of the 2015 ACM on Conference on Online Social Networks (pp. 185-195).
[62] Barros, C. P., & Prieto-Rodriguez, J. (2008). A revenue-neutral tax reform to increase demand for public transport services. Transportation Research Part A: Policy and Practice, 42(4), 659-672.
[63] Flamm, M., & Kaufmann, V. (2007, June). The concept of Network of Usual Places as a tool for analysing human activity spaces. In 11th World conference on transport research Berkeley (pp. 1-28).
[64] Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. Procedia-Social and Behavioral Sciences, 138, 557-565.
[65] Stopher, P. R., Jiang, Q., & FitzGerald, C. (2005). Processing GPS data from travel surveys. 2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications, Toronto.

[66] Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. ACM Transactions on Information Systems (TOIS), 25(3), 12-es.
[67] Chen, W., Ji, M. H., & Wang, J. M. (2014). T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation. International Journal of Online Engineering, 10(6).
[68] Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of urban technology, 17(1), 3-27.
[69] Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area.
[70] Phithakkitnukoon, S., Horanont, T., Lorenzo, G. D., Shibasaki, R., & Ratti, C. (2010, August). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In International workshop on human behavior understanding (pp. 14-25). Springer, Berlin, Heidelberg.
[71] Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation research part c: emerging technologies, 58, 240-250.
[72] Xie, K., Deng, K., & Zhou, X. (2009, November). From trajectories to activities: a spatio-temporal join approach. In Proceedings of the 2009 International Workshop on Location Based Social Networks (pp. 25-32).
[73] Huang, L., Li, Q., & Yue, Y. (2010, November). Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks (pp. 27-30).
[74] Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., & Bravo, L. (2021). Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. EPJ data science, 10(1), 29.
[75] Moya-Gómez, B., Salas-Olmedo, M. H., García-Palomares, J. C., & Gutiérrez, J. (2018). Dynamic accessibility using big data: the role of the changing conditions of network congestion and destination attractiveness. Networks and Spatial Economics, 18(2), 273-290.
[76] Zhang, L., Shin, H. S., Ghader, S., Darzi, A., Zhao, G., & Kabiri, A. (2021). Equity in Accessibility to Opportunities: Insights, Measures, and Solutions based on Mobile Device Location Data.
[77] U.S. Census Bureau. (2020). 2019 American Community Survey (ACS) 5-Year Estimates, accessed on February 2022 at https://data.census.gov/cedsci.
[78] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 1-52.
[79] U.S. Bureau of Labor Statistics. (2020). 2017, 2018, and 2019 American Time Use Survey (ATUS). https://www.bls.gov/tus/.
[80] Hastie, T. J., & Tibshirani, R. J. (2017). Generalized additive models. Routledge.

[81] Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 95-114.
[82] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 99(467), 673-686.
[83] Wood, S. N. (2006). Generalized additive models: an introduction with R. chapman and hall/CRC.
[84] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1), 3-36.
[85] "About The Region". Baltimore Metropolitan Council, Population Division. August 12, 2021. https://baltometro.org/about-us/about-bmc/about-the-region