

Unobserved heterogeneity in transportation equity analysis: Evidence from a bike-sharing system in southern Tampa

Zhiwei Chen, Xiaopeng Li^{*}

Department of Civil and Environmental Engineering, University of South Florida, United States

ARTICLE INFO

Keywords:

Transportation equity
Unobserved heterogeneity
Random parameters model
Disaggregated data
Bike-sharing

ABSTRACT

Assessing the equity impacts of transportation systems/policies has become a crucial component in transportation planning. Existing statistical modeling approaches for transportation equity analysis have typically assumed that parameter estimates are constant across all observations and used data aggregated to certain geographic units for the analysis. Such methods cannot capture unobserved factors that are not contained in the dataset, i.e., *unobserved heterogeneity*, which is likely to be present in the increasingly popular disaggregated datasets. To investigate whether there is unobserved heterogeneity in transportation equity impacts, this study carries out an empirical study focusing on the distribution of individual accessibility to activity locations via bike-sharing in southern Tampa. A disaggregated dataset containing information on individual bike-sharing accessibility and socio-economic factors is modeled with a random parameters logit model that allows for the investigation of possible unobserved heterogeneity. Further, models are estimated using data aggregated to parcel- and TAZ-levels to explore the impacts of data aggregation on model estimation results. The models unveil the unobserved heterogeneity in bike-sharing accessibility among populations in different groups defined by different sociodemographic factors in southern Tampa. These results shed insights into how the inconsistent disparity direction of transportation outcomes across individuals in a population group can be measured from the heterogeneity effects. Finally, a comparison between different models show that to capture such inconsistency, the use of disaggregated data with heterogeneity models is highly recommended for transportation equity analysis.

1. Introduction

Title VI of the Civil Rights Act of 1964 requires programs, transportation programs included, that receive federal funds to bring outcomes (e.g., costs, benefits) to society non-discriminatively (Karer and Niemeier, 2013). Achieving an equitable distribution of the outcomes of transportation systems across space and different sociodemographic groups has then become a primary challenge facing urban transportation planning (Gao and Klein, 2010; Carrier et al., 2014). Regardless of enormous efforts on ensuring equity from transportation planners, assertions of inequity have been witnessed in traditional and emerging transportation systems (Noland and Lem, 2002; Guo et al., 2020). Many traditional transportation practices were known to favor automobile travel rather than other modes of transportation that are heavily relied on by socially- or economically-disadvantaged groups (Litman and Brennan, 2012). Further, research has also identified inequities in access to the emerging bike-sharing systems, with white, college-educated,

and affluent people being overrepresented among the registered users of several bike-sharing systems in the United States (Ursaki and Aultman-Hall, 2015). Therefore, analyzing the equity impacts of transportation systems/policies has become an important part of transportation project design.

In the challenging mission of analyzing the equity impacts for transportation systems, a critical step is to assess the distribution of their outcomes across space and/or different demographic groups. One typical approach is to apply a mismatch analysis of simple descriptive statistics (e.g., mean) of the transportation outcome measures. This method presents the distribution of the outcome measures in maps or tables and then compares the distributions (Currie, 2004; Karn Kaplan et al., 2014; Golub and Martens, 2014; El-Geneidy et al., 2016). Mismatch analysis is quite simple and intuitive, but it cannot offer much quantitative information on equity performance. Therefore, transportation researchers have adopted inequality indexes from social science to obtain a quantitative evaluation of the overall degree of

^{*} Corresponding author.

E-mail addresses: zhiweic@usf.edu (Z. Chen), xiaopengli@usf.edu (X. Li).

<https://doi.org/10.1016/j.jtrangeo.2021.102956>

Received 31 March 2020; Received in revised form 4 January 2021; Accepted 5 January 2021

Available online 18 January 2021

0966-6923/© 2021 Elsevier Ltd. All rights reserved.

inequality. Popular inequality indicators include Gini index (Delbosc and Currie, 2011; Karn Kaplan et al., 2014; Guzman et al., 2017; Pritchard et al., 2019), Atkinson index (Levy et al., 2009), Theil's entropy index (Hamidi et al., 2019), comparative environmental risk index (Kocak, 2019), and subgroup inequality index (Gurram et al., 2015; Gurram et al., 2019; Chen et al., 2019). Another approach to quantitatively evaluating inequity, which is the focus of this paper, is statistical modeling such as the ordinary least squares regression model (Ogilvie and Goodman, 2012), the negative binomial model (Wang and Akar, 2019), and the multinomial logit model (Wang and Akar, 2019; Raux et al., 2017). In these models, the sign of the estimated coefficient on each explanatory variable indicates whether a population group with that factor is over- or under-represented in terms of the outcome they receive, while the magnitude of the coefficient is often used as a measure of the strength of the disparity (or level of inequality).

Due to their exploratory capability in describing the relationship between transportation outcomes and sociodemographic attributes, statistical models have been popularly applied for transportation equity analysis such as disparities among populations in terms of accessibility (Ogilvie and Goodman, 2012; Wang and Akar, 2019; Raux et al., 2017), exposure to emissions (Buzzelli and Jerrett, 2007; Havard et al., 2009), and safety outcomes (Harper et al., 2015; Kravetz and Noland, 2012; Morency et al., 2012). In these studies, data were usually aggregated to certain geographic scales for analysis, e.g., census tracts (Morency et al., 2012; Buzzelli and Jerrett, 2007), census blocks (Havard et al., 2009; Wang and Akar, 2019), and census block groups (Kravetz and Noland, 2012). However, data aggregation has been demonstrated to absorb individual heterogeneity in the distribution of transportation outcomes (Chen et al., 2019). Therefore, researchers have promoted the use of individual-level disaggregated data to unveil individual disparities in transportation equity analysis (Bills and Walker, 2017; Chen et al., 2019; Gurram et al., 2019). Using disaggregated data unveils individual heterogeneity that tends to be hindered by aggregated approaches and therefore offers a better interpretation of the equity impacts. Although disaggregated data contain rich information on individual sociodemographic attributes, these databases usually only cover a small fraction of the large number of elements that define individual sociodemographic status. Many other factors indeed remain unobserved to analysts. For example, the availability of credit cards has been shown to be a significant factor contributing to individual access to bike-sharing services (Shaheen et al., 2017). However, datasets providing sociodemographic information (e.g., those from the American Community Survey) do not contain such information.

These unobserved factors can introduce variations in the distribution of transportation outcomes among individuals that belong to a socio-demographic group. For instance, consider race as an unobserved factor that correlates with the distribution of transportation outcomes. Although there has been ample evidence that different racial groups benefit differently from some transportation systems in terms of their accessibility (Chen et al., 2019), there are also great variations across people of the same race, including differences in transportation needs, availability of credit cards, English proficiency and other factors that are generally unavailable to the analysts. This type of heterogeneity essentially are not captured in the dataset, and thus is called *unobserved heterogeneity* in the literature. It has been frequently observed in highway accident data and traffic flow data. In highway accident data, unobserved heterogeneity is found in a range of explanatory variables such as gender, age, vehicle type, traffic volume, and so on (Mannering et al., 2016). In traffic flow data, unobserved heterogeneity exists in driving styles, vehicle types, and leader-follower pair compositions (Ossen and Hoogendoorn, 2011).

However, few studies have investigated unobserved heterogeneity for transportation equity analysis. Existing studies applying exploratory statistical models for transportation equity analysis have generally assumed that parameters in the estimated model are constant across all observations; i.e., there is no unobserved heterogeneity in the collected

data. However, ignoring the unobserved heterogeneity may result in misspecified models, and thus the estimated parameters will be biased and inefficient (Mannering et al., 2016). As a result, inference and policy implications based on the estimated models will be erroneous. An instance is the relationship between race and the accessibility an individual receives from a bike-sharing system. As mentioned previously, there are multiple reasons to believe that the accessibility people receive from a bike-sharing system varies from individual to individual even if they are the same race owing to unobserved individual heterogeneity. Nevertheless, if possible unobserved heterogeneity across individuals were ignored, incorrect conclusions may have been drawn from erroneous parameter estimates such as believing that all individuals in one racial group have higher accessibility than the general population. Indeed, there are possibly some individuals of that race who do not receive accessibility from the system at all because the lack of credit cards prevents them from using the services (which is not captured by the analysis dataset). This phenomenon is particularly likely when a disaggregated dataset containing a large number of samples is used for analysis. Without explicitly considering the unobserved heterogeneity, it will be difficult to determine whether this statement is true or not. Therefore, approaches to accounting for unobserved heterogeneity for analyzing transportation equity impacts are needed.

Against this background, this paper aims to apply an exploratory modeling approach to investigate whether there is unobserved heterogeneity in transportation equity impacts. To this end, we carry out a case study focusing on the distribution of individual accessibility to activity locations via bike-sharing in southern Tampa using a disaggregated dataset. The dataset includes individual bike-sharing accessibility and sociodemographic information. To allow for the investigation of possible unobserved heterogeneity in the data, three random parameters logit models are estimated using the individual-level data and data aggregated to land parcel and traffic analysis zone (TAZ) levels, respectively. The novelty of this paper lies in allowing for possible unobserved heterogeneity of random parameters and the use of heterogeneity effects in equity impact measurement for transportation equity analysis. The main contributions of this paper are threefold. First, we apply a heterogeneity modeling approach to study the equity impacts of a case study transportation system in southern Tampa using a disaggregated dataset. The use of a heterogeneity model captures the potential unobserved heterogeneity hidden in the large amount of individual-level data while traditional fixed parameters models cannot serve the same purpose. We also investigate how heterogeneity effects can be used to measure the equity impacts in transportation systems. Incorporating this unobserved heterogeneity ensures that the model is correctly specified, and thus the resultant equity interpretations are unbiased and efficient. Second, we offer a comparison between the model estimation results using disaggregated data, data aggregated to parcels, and data aggregated to TAZs. This comparison enables an in-depth understanding of the impacts of data aggregation on parameter estimates in the model, which are of special importance to transportation planners and equity analysts. It also provides insights into the presence of unobserved heterogeneity in aggregated data. Finally, the case study reveals how bike-sharing accessibility is distributed among the population and population subgroups defined by various sociodemographic attributes in southern Tampa. It also confirms the existence of unobserved heterogeneity in the investigated data, the necessity of using a heterogeneity model, and the advantage of using disaggregated data. Overall, this study offers empirical evidence to transportation equity analysts of the existence and importance of incorporating unobserved heterogeneity given the emerging use of disaggregated data in transportation equity analysis. Particularly, various equity implications and recommendations of the studied bike-sharing system provide insights that can assist bike-sharing operators in designing unobserved heterogeneity-aware equitable expansion plans in southern Tampa and beyond.

The remainder of this paper is organized as follows. [Section 2](#)

presents the materials (including the bike-sharing system, analysis area, and data collection), and the analytical method used for modeling the data. Section 3 presents and discusses the analysis results, including descriptive statistics, model estimation results using the individual-level, parcel-level, TAZ-level data, and the corresponding equity implications. Finally, Section 4 provides conclusions and potential future research directions.

2. Materials and methods

The primary goal of this study is to investigate unobserved heterogeneity in the equity impacts of transportation outcomes via a case study bike-sharing system. To this end, we first collected data on individual sociodemographic and accessibility to activity locations via a bike-sharing system in the study area. Then, a statistical modeling approach that captures unobserved heterogeneity was applied to model the data. This section presents the bike-sharing system, the analysis area, the investigated dataset and the statistical modeling approach successively.

2.1. Bike-sharing system

We selected the Coast Bike Share system in Downtown Tampa, the central business district of Tampa (see Fig. 1), as a case study for exploring possible unobserved heterogeneity in the equity impacts of a transportation system. The selected system is an interesting case study for a number of reasons. Bike-sharing is a fast-growing mode of transportation in many countries with increasing equity concerns (Lei and Ouyang, 2018; Chen et al., 2019; Chen et al., 2020; Qian and Niemeier, 2019). Starting in late 2014, the Coast Bike Share system has been serving both tourists and residents (Mckenna, 2016), and therefore we can use it as a case to analyze residents' accessibility to activity locations via bike-sharing. Note that although the Coast Bike Share system in Downtown Tampa is relatively small, it was listed as one of the 29th largest bike-sharing systems in the US (Barajas and Drive, 2018). Its small service scale also makes equity a more sensitive issue. Additionally, the area where the bike-sharing system is located features high diversity in terms of residents' sociodemographic profiles, making it appropriate for studying the accessibility distribution across different population groups. Furthermore, although a previous study (Chen et al.,

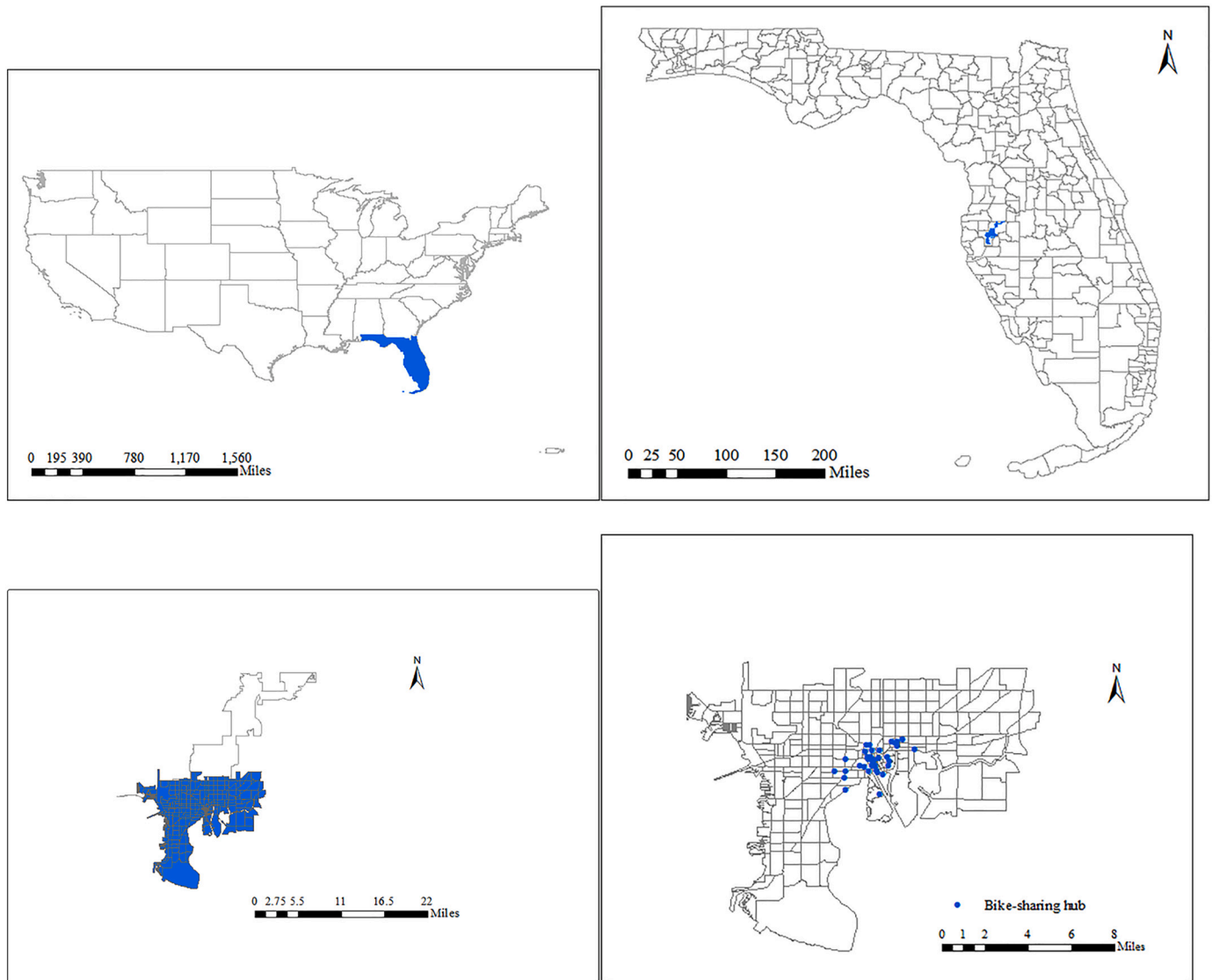


Fig. 1. The analysis area. The top left sub-figure presents the location of Florida within the United States. The top right sub-figure presents the location of Tampa within Florida. The bottom left sub-figure presents the area selected for analyzing the equity impacts in terms of individual accessibility via bike-sharing, i.e., southern Tampa. The bottom right sub-figure shows the distribution of bike-share hubs (represented by dots) of the Coast Bike Share system in Downtown Tampa. Source: Created by the authors.

2019) has revealed inequality of the accessibility distribution of the Coast Bike Share system among different sociodemographic groups of southern Tampa, their method does not capture any unobserved heterogeneity. Certain groups (e.g., whites, Asians, non-Hispanics) have been shown to be overrepresented in terms of the accessibility they receive from the system but whether the overrepresentation is consistent across the entire population subgroup is still not clear. Therefore, a case study using an approach that capture unobserved heterogeneity would be helpful for policymakers and bike-sharing operators to better understand the equity impacts and to design unobserved heterogeneity-aware equitable expansion plans of this system and other bike-sharing systems more broadly.

2.2. Analysis area

Following Chen et al. (2019), we selected an area that surrounds the Coast Bike Share system for the analysis. This area corresponds to the Tampa City Public Use Microdata Area (PUMA) in Florida, which is the most detailed geographic unit with available individual records from the US Census Bureau (Gliebe et al., 2014; US Census Bureau, 2010). It is located in the southern part of Tampa as shown in Fig. 1, and thus we call it southern Tampa. According to the US Census Bureau (2019), this area had a population of 168,835 as of 2018 and the sociodemographic profiles of the residents in this area were highly diverse. Specifically, the population distribution for white, black, Asian and other racial groups was 58%, 10%, 2%, and 30%, respectively. Twenty six percent of the population identified themselves as Hispanic or Latino origin. Male accounted for 50% of the population and the age distribution was 20% under 18 years, 65% between 18 and 64, and 15% over 65. In terms of the annual household income, 11.5% of the population lived below the poverty line while 34% of the households earned more than \$100,000 per year. Further, the selected analysis area covers residents within and outside the service area of the bike-sharing system, and thus an analysis based on this area likely paints a more complete picture of the equity landscape of the system.

2.3. Data collection

The data used for this study were collected from a synthetic population of southern Tampa simulated with the Person Day Activity and Travel Simulator (Daysim, Bradley et al., 2010). These include information on individual sociodemographic and accessibility to activity locations in their daily travel itineraries over a typical weekday via the Coast Bike Share system. The synthetic population was generated by an iterative proportional fitting approach (Beckman et al., 1996) from the PUMS of the American Community Survey (US Census Bureau, 2010) by Gurram et al. (2019). To ensure that the synthetic population is representative of the real-world data, multiple variables are used to calibrate the model, such as land use, demographic information, and trip characteristics. For example, the statistics of the synthetic population are consistent with those of the original census data, indicating the validity of the synthetic population. Note that since using real individual information will involve privacy concerns and a huge amount of time and money to collect data from a full population, here we adopt a synthetic population that well represents the real population in southern Tampa. There are some individual records with missing sociodemographic information in the dataset due to the presence of incomplete individual records in the PUMA data for generating the synthetic population. Further, kids under the age of 18 are not allowed to use the bike-sharing service due to the unavailability of credit cards and/or incapability to ride (Coast Bike Share, 2019). Therefore, we excluded them from the analysis. Finally, we removed all these problematic records and obtained a sample size of 75,697.

The sociodemographic information contains the individuals' residence locations and sociodemographic attributes. Both the land parcel and TAZ of each hypothetical individual were identified. We determined

the centroid coordinate locations of land parcels and TAZ's from Daysim and PlanHillsborough (2017), respectively. These were used to aggregate individual-level data to different geographic scales for comparative equity analyses. For sociodemographic attributes, we considered the number of tours in one's daily travel itinerary, age (18–35, 35–45, 45–65, above 65), gender (male, female), race (white, black, Asian, other), ethnicity (Hispanic and non-Hispanic), nativity (native American and non-native), marital status (married, widowed, divorced, separated, never married or under 15 years old), household income level (below poverty, middle income defined as above the 2009 poverty level but with an annual household income below \$75,000, and upper income with an annual household income above \$75,000), person type (workers, university students), commute time (i.e., travel time from home to work) by car in minutes, household size (i.e., number of people in the household), number of workers in the household, number of retired adults in the household, number of kids aged less than 16 in the household, and number of university students in the household. The attributes are divided into different subgroups following Chen et al. (2019) and American Community Survey.

Individual accessibility to activity locations via the bike-sharing system was estimated with a tour-based approach proposed in Chen et al. (2019) using information regarding the locations of bike-sharing stations and individual daily travel itineraries generated from Daysim. This approach takes into account the intermodal characteristic of bike-sharing trips and the interdependence of individual mode choice decisions for neighboring trips. The resulting accessibility of each person is a continuous value ranging from 0 to 1, with 0 representing the lowest value and 1 representing the highest. This accessibility measure is essentially a person-based potential accessibility measure (Geurs and Van Wee, 2004). It quantifies the potential of an individual in reaching all activity locations in her daily travel itinerary via the bike-sharing system. The potential accessibility measure has been shown to be an appropriate social indicator for analyzing the level of access to social and economic opportunities related to the activity locations of one's daily travel itinerary for different sociodemographic groups. Additional details of the tour-based accessibility modeling approach can be found in Chen et al. (2019).

However, the accessibility measure selected does not involve an individual's actual usage of the bike-sharing system. As a result, the proposed method cannot answer questions such as how the visitors' (the other group of target customers for the studied bike-sharing system) usage on the bike-sharing system affects city residents' bike-sharing accessibility. The actual usage of shared bikes depends on the spatio-temporal distribution of the shared bikes, which is a function of a range of complicated factors including the repositioning activities, the bicycle pick-up and drop-off dynamics at each station regardless of the user type (i.e., residents and tourists), etc. The availability of data on the travel behavior of visitors poses another challenge to the consideration of visitors in the analysis. Thus, to investigate the bike-sharing accessibility based on the users' actual usage of the system would amount to a new problem that is out of the scope of this study.

2.4. Analysis methodology

We aim to estimate a heterogeneity model to capture unobserved heterogeneity in the distribution of bike-sharing accessibility among different sociodemographic groups of southern Tampa. There are various heterogeneity models in statistical modeling, such as random parameters models (Anastasopoulos and Mannering, 2009, 2011, latent class models (Behnood et al., 2014), and their combinations (Xiong and Mannering, 2013), as summarized in Table 1. Yet, no consensus has been reached on which approach is always superior in addressing unobserved heterogeneity. As an empirical study to demonstrate the presence of unobserved heterogeneity in transportation equity impacts, this study adopts the most widely used random parameters approach (Mannering et al., 2016) to model the collected data.

Table 1

Comparisons of popular heterogeneity models. Source: After [Mannering et al. \(2016\)](#).

Models	Concepts	Advantages	Disadvantages
Random-parameters models	Allow parameter estimates to follow a continuous statistical distribution	Capture heterogeneity across individual observations	Require an assumption on the parameter distribution
Latent class models	Separate the sample into groups (or classes) within which parameters are homogeneous	Allow for fixed parameters within classes; do not require an assumption on the parameter distribution	Are difficult to estimate if there are many latent classes; cannot capture possible heterogeneity within a class
Latent class models with random-parameters	Separate the samples into groups and allow parameter variations within each group	Combine the advantages of random-parameters and latent class models	Are very cumbersome to estimate, especially for large-scale disaggregated datasets

2.4.1. Statistical modeling approach

To convert the continuous bike-sharing accessibility values to a set of binary variables that can be modeled by the mixed logit model, we define a set of bike-sharing accessibility levels indexed as $n \in \mathcal{N} := \{0, 1\}$, where $n = 0$ represents an accessibility value of 0 and $n = 1$ represents an accessibility value greater than 0. Besides, we define binary variable Y_{ni} , which equals 1 if the bike-sharing accessibility level of individual i is n and 0 otherwise. Specifically, $Y_{0i} = 1$ and $Y_{1i} = 0$ if the accessibility value of individual i is 0, while $Y_{0i} = 0$ and $Y_{1i} = 1$ if the accessibility value of individual i is greater than 0. Thus, dependent variables $Y_{ni} := \{Y_{0i}, Y_{1i}\}$ represent a set of discrete (binary in this case) outcomes, to model which mixed (or random parameters) logit model is appropriate. To arrive at a mixed logit model that allows for heterogeneity in the estimated parameters across observations (i.e., individuals in this paper), a function that determines the probability of individual i having outcome n can be defined as ([Washington et al., 2010](#)):

$$Y_{ni} = \beta_n X_{ni} + \epsilon_{ni} \quad (1)$$

where β_n is a vector of estimated parameters for discrete outcome n and X_{ni} is a vector of observable explanatory sociodemographic attributes that correlate with discrete outcome n for individual i , and ϵ_{ni} is a disturbance term that is assumed to be extreme value distributed. Note that all notation in bold is vector while the remaining is not. The basic idea of random parameters models is to account for the unobserved heterogeneity by allowing estimated parameters in the model to vary across observations following a predefined probability distribution with probability function $f(\beta|\varphi)$, where φ refers to a vector of parameters of the density function (i.e., mean and variance). Then, the outcome probability of a mixed logit model, which accounts for unobserved heterogeneity in the collected data, can be derived as ([Train, 2009](#)):

$$P_i(n|\varphi) = \int \frac{e^{\beta_n X_{ni}}}{\sum_{n \in \mathcal{N}} e^{\beta_n X_{ni}}} f(\beta_n|\varphi) d\beta_n \quad (2)$$

where $P_i(n|\varphi)$ represents the probability that the accessibility outcome of individual i is n given a vector of parameters φ of the density function. Obviously, the probability shown in Eq. (2) is a weighted average of the probabilities of a set of standard multinomial logit models corresponding to different values of β .

For model estimation, we used a popular statistical analysis software, NLOGIT5. We considered various probability distributions in the model estimation, including normal, lognormal, triangular, and uniform distributions. The model estimation was realized using the simulated likelihood method with 200 Halton draws, which has been shown to

result in sufficiently accurate parameter estimation ([McFadden and Train, 2000](#)). Marginal effects were computed as $(\partial E[Y_{ni}|X_{ni}]) / (\partial X_{ni})$ by NLOGIT5 ([Greene, 2007](#)) to quantify the influence that one-unit increase in an explanatory variable X_{ni} has on the accessibility outcome probabilities Y_{ni} to further interpret the model estimation results.

Apart from analyzing the accessibility distribution of the Coast Bike Share system using the individual data, we also aggregated the data to both the parcel- and TAZ-level to investigate the impacts of data aggregation on the unobserved heterogeneity. For example, we computed the mean for all numerical values (e.g., accessibility, age) for all individuals residing in a land parcel. For sociodemographic attributes with factorial values (e.g., race, gender), we computed the proportion of the population that belongs to a group in each land parcel. Similar aggregations were also applied to TAZs. Then, we estimated mixed logit models to analyze the accessibility distribution with the aggregated data also using NLOGIT 5.

2.4.2. Statistical evaluation criteria

To make sure that the included variables in the models are statistically significant and representative, we applied four statistical evaluation criteria during the model estimation process. First, the p -values of all included variables should be less than 0.05, indicating that the variables are statistically significant given a confidence level of 95%. Second, the sample size of the included variables should account for at least 10% and at most 90% of the population because we cannot determine causality if the number of observations in a subgroup is too high or too small. By applying this rule, the sample size of each subgroup represented by an indicator variable will be between 7,569 and 68,127, which would provide enough observations for causality inference. Third, we computed the McFadden ρ^2 statistic to assess the overall model fitness ([McFadden, 1973](#)). The formulation of the ρ^2 statistic is

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)} \quad (3)$$

where $LL(\beta)$ is the log-likelihood at convergence with parameter vector β and $LL(0)$ is the log-likelihood at zero (i.e., with all parameters set to 0). This statistic is similar to R^2 in ordinary least square regression models, whose value ranges from 0 to 1. A ρ^2 statistic value close to 1 indicates that the model can predict the outcomes with near certainty. Finally, to statistically determine whether we need a mixed logit model rather than a fixed parameters logit model for the investigated dataset, we carried out a likelihood ratio test. Specifically, we computed the likelihood ratio test statistic as ([Washington et al., 2010](#))

$$X^2 = -2[LL(\beta_{\text{fixed}}) - LL(\beta_{\text{mixed}})] \quad (4)$$

where $LL(\beta_{\text{mixed}})$ is the log-likelihood at convergence of the mixed logit model and $LL(\beta_{\text{fixed}})$ is the log-likelihood at convergence of the fixed parameters logit model with the same explanatory variables but all their estimated coefficients remain constant across observations. The resulting X^2 is χ^2 distributed, with degrees of freedom equaling to the difference between the numbers of parameters in the mixed and fixed parameters logit models. Then we used the Chi-Square Calculator ([Stat Trek, 2019](#)) to determine whether the null hypothesis stating that the parameters in the logit model are constant across observations (i.e., there is no unobserved heterogeneity in the data) can be rejected.

3. Results and discussion

This section discusses the results of our analyses. [Section 3.1](#) presents the variables used for model estimation and summary statistics for bike-sharing accessibility computed by each variable that represents a population subgroup. Results from the disaggregated (i.e., individual-level) data analysis are presented in [Section 3.2](#). Finally, [Section 3.3](#) compares the results from the disaggregated model with those from the aggregated (i.e., parcel-level and TAZ-level) data analyses.

3.1. Descriptive statistics

Table 2 summarizes the variables used for model estimation and summary statistics for bike-sharing accessibility by each variable (or population subgroup) in the analysis area.

As seen in Table 2, there are 75,697 individuals in the final dataset used for the model estimation. The sampled population is divided into different subgroups, each of which corresponds to an indicator variable defined by a sociodemographic attribute. We see that from the last row in Table 2 that the mean bike-sharing accessibility for the population in southern Tampa is 0.0027. A 75th percentile value of 0 indicates that at least 75% of the population in southern Tampa does not have bike-sharing accessibility at all. This finding is consistent with those from Chen et al. (2019), which also reveals relatively low bike-sharing accessibility and an extremely skewed distribution of the bike-sharing accessibility in southern Tampa overall.

The bike-sharing accessibility is distributed among the individuals in each subgroup following a similar skewed distribution, with the 75th percentiles of all subgroups being 0. However, there is an obvious disparity of the mean bike-sharing. Specifically, the mean bike-sharing accessibility for certain subgroups (i.e., individuals with the commute time by car less than 29 min, people aged between 19 and 34, those with an annual household income higher than \$74,999, males, Asians, non-Hispanics, native American, people who never married, university students, individuals living in household without kids under 16 years old, and those who live in a household without university students) are higher than the population mean. However, certain groups receive lower bike-sharing accessibility than the entire population on average, including individuals with the commute time by car more than 29 min, people aged over 34, those with an annual household income from \$25,000 to \$74,999, females, black people, Hispanics, non-Native American, individuals who have married at least once, workers, those living in a household with at least 2 workers, and those living in a household with retired adults. For the rest of the subgroups, the mean bike-sharing accessibility is the same as the population mean. These results describe a seemingly uneven distribution of the bike-sharing accessibility among different population subgroups in southern Tampa. However, the difference of the mean bike-sharing accessibility values between different subgroups is relatively small (i.e., less than 0.01). Thus, it is not clear whether the disparity in the mean bike-sharing accessibility between different subgroups is statistically significant or not; probably this disparity can be ignored if it is statistically insignificant. An in-depth analysis using statistical methods (e.g., analysis of variance, regression modeling) is needed to answer this question. We will present results from the estimated mixed logit model in the following subsection for this purpose.

3.2. Model estimation results using individual-level data

The mixed logit model estimation results using individual-level data are presented in Table 3. Parameter distribution of the random parameters in the mixed logit model is reported in Table 4. Results of the likelihood ratio tests are presented in Table 5.

We first analyze the statistical performance of the mixed logit model. As shown in Table 3, 12 variables (including a constant) are included in the estimated mixed logit model and their absolute values of the *t*-statistics are all greater than 1.96 (i.e., the *p*-values are less than 0.05), indicating that the variables in the estimated mixed logit model are statistically different from zero when conducting a two-tailed *t*-test. Note that not all variables with a mean accessibility unequal to the population mean is found to be statistically significant in the estimated model (e.g., age group 3). This observation highlights an advantage of using statistical modeling for assessing the equity impacts of transportation outcomes over comparisons of simple summary statistics: they can filter out population subgroups that are slightly over-/under-represented so that focuses can be placed on groups facing substantial

Table 2

Variable definition and summary statistics for the bike-sharing accessibility by each variable (or population subgroup). Source: Created by the authors.

Variable	Sample size (%)	Descriptive statistics of accessibility					
		Min	25th % ile	Median	Mean	75th % ile	Max
Indicator variables							
Very short trip: 1 if the commute time by car is less than 10 min; 0 otherwise	8096 (10.70)	0	0	0	0.0031	0	0.32
Short trip: 1 if the commute time by car is between 10 and 19 min and 0 otherwise	21,875 (28.90)	0	0	0	0.0029	0	0.35
Medium trip: 1 if the commute time by car is between 20 and 29 min and 0 otherwise	17,923 (23.68)	0	0	0	0.0029	0	0.35
Long trip: 1 if the commute time by car is more than 29 min and 0 otherwise	27,803 (36.73)	0	0	0	0.0023	0	0.41
Age group 1: 1 if the age ranges from 19 to 34; 0 otherwise	26,015 (34.37)	0	0	0	0.0038	0	0.41
Age group 2: 1 if the age ranges from 35 to 44; 0 otherwise	18,152 (23.98)	0	0	0	0.0025	0	0.28
Age group 3: 1 if the age ranges from 45 to 64; 0 otherwise	27,393 (36.19)	0	0	0	0.0018	0	0.34
Age group 4: 1 if the age is greater than 64; 0 otherwise	4137 (5.47)	0	0	0	0.0020	0	0.32
Low income: 1 if the annual household income is less than \$25,000; 0 otherwise	8804 (11.63)	0	0	0	0.0027	0	0.32
Middle income: 1 if the annual household income ranges from \$25,000 to \$74,999; 0 otherwise	34,783 (45.95)	0	0	0	0.0022	0	0.35
High income:1 if the annual household income is	32,110 (42.42)	0	0	0	0.0031	0	0.41

(continued on next page)

Table 2 (continued)

Variable	Sample size (%)	Descriptive statistics of accessibility					
		Min	25th % ile	Median	Mean	75th % ile	Max
higher than \$74,999; 0 otherwise							
Male: 1 if the gender is male; 0 otherwise	39,000 (51.52)	0	0	0	0.0028	0	0.41
Female: 1 if the gender is female; 0 otherwise	36,697 (48.48)	0	0	0	0.0025	0	0.35
White: 1 if the race is white; 0 otherwise	59,505 (78.61)	0	0	0	0.0027	0	0.41
Black: 1 if the race is black; 0 otherwise	10,256 (13.55)	0	0	0	0.0022	0	0.30
Asian: 1 if the race is Asian; 0 otherwise	2182 (2.88)	0	0	0	0.0028	0	0.20
Other: 1 if the race is not white, black or Asian; 0 otherwise	3754 (4.96)	0	0	0	0.0027	0	0.29
Hispanic: 1 if the ethnicity is Hispanic; 0 otherwise	15,625 (20.64)	0	0	0	0.0023	0	0.35
Non-Hispanic: 1 if the ethnicity is not Hispanic; 0 otherwise	60,072 (79.36)	0	0	0	0.0028	0	0.41
Native: 1 if an individual is native American; 0 otherwise	57,976 (76.59)	0	0	0	0.0028	0	0.41
Non-native: 1 if an individual is not native American; 0 otherwise	17,721 (23.41)	0	0	0	0.0021	0	0.35
Married: 1 if the marital status is married; 0 otherwise	23,699 (31.31)	0	0	0	0.0015	0	0.29
Widowed: 1 if the marital status is widowed; 0 otherwise	2031 (2.68)	0	0	0	0.0017	0	0.28
Divorced: 1 if the marital status is divorced; 0 otherwise	15,228 (20.12)	0	0	0	0.0024	0	0.32
Separated: 1 if the marital status is separated; 0 otherwise	3057 (4.04)	0	0	0	0.0025	0	0.26
Never married: 1 if the marital status is never	31,682 (41.85)	0	0	0	0.0037	0	0.41

Table 2 (continued)

Variable	Sample size (%)	Descriptive statistics of accessibility					
		Min	25th % ile	Median	Mean	75th % ile	Max
married or under 15 years old; 0 otherwise							
Worker: 1 if an individual is a worker; 0 otherwise	73,806 (97.50)	0	0	0	0.0026	0	0.41
University student: 1 if an individual is a university student; 0 otherwise	1800 (2.38)	0	0	0	0.0053	0	0.27
Household with workers: 1 if an individual lives in an household with at least 2 workers; 0 otherwise	46,465 (61.38)	0	0	0	0.0022	0	0.41
Household with retired adults: 1 if an individual lives in household with at least 1 retired adult; 0 otherwise	16,011 (21.15)	0	0	0	0.0013	0	0.30
Household without kids under 16 years old: 1 if an individual lives in a household with no kid under 16 years old; 0 otherwise	53,715 (70.96)	0	0	0	0.0033	0	0.41
Household without university students: 1 if an individual lives in an household without university students; 0 otherwise	70,891 (93.65)	0	0	0	0.0026	0	0.41
The entire population	75,697 (100)	0	0	0	0.0027	0	0.41

disparities. Also, the estimated model shows a good overall statistical fit with a ρ^2 value of 0.70 (Anastasopoulos and Mannering, 2009, 2011).

The results also reveal the superiority to use a heterogeneity model. We see that the t-statistics of the standard deviations (i.e., the t-statistics in parenthesis) of 8 variables are greater than 1.96, indicating the standard deviations of the density functions of these parameters are statistically different from zero in a two-tailed t-test (at a confidence level of 95%). Thus, these random parameters are statistically

Table 3

Mixed logit model results of the individual-level bike-sharing accessibility in southern Tampa. (Standard deviation of parameter estimate, in parentheses). Source: Created by the authors.

Variable	Estimated parameter	t-Stat.	Marginal effects
Constant	-1.67 (0.82)	-8.95 (45.28)	–
Number of tours	0.31 (0.12)	14.71 (15.35)	0.018
Age group 1: 1 if the age ranges from 19 to 34; 0 otherwise	0.55	16.69	0.032
Age group 2: 1 if the age ranges from 35 to 44; 0 otherwise	0.13 (0.80)	3.65 (21.37)	0.008
Female: 1 if the gender is female; 0 otherwise	-0.29 (0.83)	-10.24 (31.53)	-0.017
Never married: 1 if the marital status never married or under 15 years old; 0 otherwise	0.12	4.31	0.007
High income: 1 if the annual household income is higher than \$74,999; 0 otherwise	0.31 (0.13)	12.61 (5.18)	0.018
Long trip: 1 if the commute time by car is more than 29 min and 0 otherwise	-0.08 (0.28)	-3.13 (9.47)	-0.005
Number of workers in the household	-0.44 (0.33)	-23.96 (34.46)	-0.025
Number of retired adults in the household	-0.64 (0.92)	-16.15 (23.22)	-0.037
Household without kids under 16 years old: 1 if an individual lives in a household with no kid under 16 years old; 0 otherwise	0.60	17.71	0.035
Mean distance between activity locations to the nearest bike-sharing hubs	-0.16	-8.37	-0.009
Model statistics			
Number of observations	75,697		
Log-likelihood at zero, $LL(0)$	-52,449.53		
Log-likelihood at convergence, $LL(\beta)$	-15,734.86		
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.70		

Table 4

Parameter distribution of random parameters of the mixed logit model using individual-level bike-sharing accessibility in southern Tampa. Source: Created by the authors.

Variable	Above 0 (%)	Below 0 (%)
Constant	2.08	97.92
Number of tours	0.49	99.51
Age group 2: 1 if the age ranges from 35 to 44; 0 otherwise	56.45	43.55
Female: 1 if the gender is female; 0 otherwise	36.34	63.66
High income: 1 if the annual household income is higher than \$74,999; 0 otherwise	99.15	0.85
Long trip: 1 if the commute time by car is more than 29 min and 0 otherwise	38.75	61.25
Number of workers in the household	9.12	90.88
Number of retired adults in household	24.33	75.67

significant (Anastasopoulos et al., 2011); the effects of these independent variables on the dependent variable varies across the observations following a normal distribution. Note that we also tried other distributions, including lognormal, triangular, and uniform distributions, but the empirical analysis shows that no distribution was statistically superior to the normal distribution. Further, the log-likelihood ratio test results in Table 5 indicate that the null hypothesis can be rejected at a confidence level of 100%, and thus not all coefficients of the included variables in the model are constant across the observations. This result confirms the existence of unobserved heterogeneity in the investigated individual-level dataset. It also confirms that the mixed logit model

Table 5

Results of the likelihood ratio tests. Source: Created by the authors.

Item	Random parameters	Fixed parameters
Model using individual-level data		
Number of parameters	20	12
Log-likelihood at convergence, $LL(\beta)$	-15,734.86	-15,752.87
$X^2 = -2[LL(\beta_{\text{fixed}}) - LL(\beta_{\text{mixed}})]$	36.02	
Degree of freedom	8	
Probability of rejecting the null hypothesis	100%	
Model using parcel-level data		
Number of parameters	17	13
Log-likelihood at convergence, $LL(\beta)$	-10,231.11	-10,256.13
$X^2 = -2[LL(\beta_{\text{fixed}}) - LL(\beta_{\text{mixed}})]$	50.04	
Degree of freedom	4	
Probability of rejecting the null hypothesis	100%	
Model using TAZ-level data		
Number of parameters	7	6
Log-likelihood at convergence, $LL(\beta)$	-77.84	-82.27
$X^2 = -2[LL(\beta_{\text{fixed}}) - LL(\beta_{\text{mixed}})]$	8.86	
Degree of freedom	1	
Probability of rejecting the null hypothesis	99.70%	

provides a more statistically superior model fit compared with the traditional fixed parameters logit model. Therefore, we should adopt a mixed logit model that can capture possible unobserved heterogeneity in the dataset to achieve a better model fit.

Turning to the explanatory variables with fixed parameters, we see from Table 3 that four of the included parameters are fixed across the individual observations. Specifically, the marginal effects of variable *age group 1* and variable *never married* indicate that if an individual is aged between 19 and 34 or never married, the probability of them having bike-sharing accessibility will increase by 0.032 and 0.007, respectively. Likewise, the marginal effect of variable *household without kids under 16 years old* reveals that if an individual lives in a household without kids under 16 years old, his/her probability of having bike-sharing accessibility is higher than the general population by 0.035. Further, the marginal effect of variable *mean distance between activity locations to the nearest bike-sharing hub* indicates that a 1 unit increase in the mean distance between an individual's activity locations to the nearest bike-sharing hub decreases the probability of having bike-sharing accessibility by 0.009. Note that to account for possible endogeneity between the distance-related variable and the accessibility metric, an instrumental variable approach was employed: the distance-related variable was regressed against all exogenous variables and their instruments were used as independent variables.

These results reveal that population subgroups represented by these variables are entirely overrepresented (i.e., subgroups with positive marginal effects) or underrepresented (i.e., subgroups with negative marginal effects) in terms of the bike-sharing accessibility distribution. When policymakers intend to decrease/eliminate the disparity between subgroups defined by these sociodemographic attributes, they should formulate policies to improve the bike-sharing accessibility for all individuals in these subgroups that are underrepresented as a whole. In the case of southern Tampa, people having longer distance from their daily activity locations to the bike-sharing hubs are those who consistently experience lower likelihoods of having bike-sharing accessibility than the general population. If the Coast Bike Share operator intends to bridge the bike-sharing accessibility gap between this group and the general population, measures can be adopted such as deploying more bike-sharing facilities to communities with few bike-sharing hubs, offering a discounted membership or lower payment option, and promoting bike-sharing services among individuals in this subgroup. Individuals need not be treated differently in these policies since there is

no unobserved heterogeneity between individuals.

Next, we focus on the explanatory variables with random parameters, i.e., parameters with t-statistics in parenthesis. The disparity direction related to these variables are not consistent across the population; not all individuals in a related subgroup follow the same disparity direction as the subgroup mean reveals. For instance, the constant produces a negative mean parameter, indicating that the population in southern Tampa is more likely to not have bike-sharing accessibility on average. However, as shown in Table 4, the specified normal distribution for the constant suggests that 97.92% of the population in southern Tampa has a negative constant term while 2.08% has a positive one. This result indicates that despite the low likelihood for individuals to have bike-sharing accessibility in southern Tampa, a small portion (2.08%) is more likely to receive bike-sharing accessibility than others. Since the constant term does not involve any sociodemographic attributes (i.e., individuals are treated as similar), this observation indeed implies that the bike-sharing accessibility is not evenly distributed among the population in southern Tampa from the horizontal equity perspective. This result is consistent with those in Chen et al. (2019), indicating the validity and correctness of the estimated model. Likewise, variable *age group 2* produces a normally distributed positive parameter with a mean of 0.13 and a standard deviation of 0.80. This result reveals that 56.45% of individuals aged between 35 and 64 have a higher likelihood of having bike-sharing accessibility (with estimated parameters above 0) while 43.55% of individuals in this subgroup have a lower probability (with estimated parameters below 0). Other variables resulting in random parameters can be interpreted in a similar way using the parameter distributions summarized in Table 4. These results indicate that there are discrepancies between the accessibility received by different population subgroups, i.e.; the accessibility distribution in southern Tampa is not equal from the vertical equity perspective. These findings on parameters with unobserved heterogeneity offer important managerial insights for policy makers as we discuss below.

The measures that can be applied for subgroups without unobserved heterogeneity may not be as effective in addressing the inequality issues between population subgroups defined by variables with random parameters. For subgroups with negative mean parameters (meaning that these subgroups are underrepresented on average), there might be an evident portion of individuals resulting in positive estimated parameters or vice versa. For example, 43.55% of the individuals aged between 35 and 44 produce a negative parameter despite a positive mean parameter for this subgroup. In contrast, 36.34% of females lead to a positive parameter although this subgroup results in a negative mean parameter. Therefore, it may be a waste of investment to enact a policy simply targeting all individuals in underrepresented subgroups with unobserved effects. Worse still, this kind of policy may exasperate the inequality issue between different subgroups. For instance, if more bike-sharing stations are sited in areas with a higher proportion of females in southern Tampa, it is likely that the 36.34% of females who already produce a positive parameter will receive more bike-sharing accessibility owing to the improvement. As a result, this increase will drive the disparity further to the point that females will eventually be overrepresented on average.

Therefore, to design effective policies for subgroups with unobserved heterogeneity, policymakers should spend efforts on identifying people who are truly underrepresented in these subgroups, e.g., the 63.66% of females with a negative parameter. With this, equitable system improvement plans can be designed to target these individuals only rather than all individuals in an underrepresented subgroup. However, differentiating these individuals from others in the same subgroups is a challenging task that needs more sophisticated models; it involves an analysis of the distributions of individual sociodemographic and accessibility outcomes. A relevant study (Qian and Niemeier, 2019) can shed insights into the development of such methods. It proposes a new index-based method to determine the priority to site bike-sharing stations at different geographic units of analysis so that the expanded

system can better serve low income and minority households. Additionally, even though such sophisticated methods have not been available yet, results from the mixed logit model can be integrated to better define disadvantaged subgroups at which the equitable system improvement measures should target. One way is to simply prioritize subgroups with higher proportions of individuals producing negative parameter estimates. Equitable improvement investments can then be allocated to address the accessibility gaps between subgroups with high priority and those with low priority. Also, several sociodemographic attributes with higher proportions of individuals having negative parameter estimates can be weighted to define a composite index that reflects to what extent the population in a geographic unit of analysis is deprived of bike-sharing accessibility. The index can then be used to determine the priority of equitable improvement for different geographic units of analysis with their sociodemographic profiles.

3.3. Model estimation results using parcel-level and TAZ-level data

To investigate the impacts of data aggregation on revealing unobserved heterogeneity in transportation equity effects, we present and discuss the results from estimated models using the parcel-level and TAZ-level data. We also present the spatial distribution of the accessibility results and three representative sociodemographics to visualize the unobserved heterogeneity. The estimated parcel-level and TAZ-level models are summarized in Table 6 and Table 7, respectively. Results of the likelihood ratio test for these models are shown in Table 5. Spatial distributions of the accessibility and population subgroups are shown in Figs. 2 and 3.

As seen from Table 6, when data are aggregated to land parcels,

Table 6

Mixed logit model results of the parcel-level bike-sharing accessibility in southern Tampa. (Standard deviation of parameter estimate, in parentheses). Source: Created by the authors.

Variable description	Estimated parameter	t-Stat.	Marginal effects
Constant	-2.773 (2.573)	-7.71 (59.02)	/
Average number of tours	0.491 (0.068)	12.52 (5.60)	0.007
Proportion of people aged from 19 to 34 years old	0.014	13.12	2.00E-4
Proportion of people aged from 35 to 44 years old	0.008	7.95	1.20E-4
Proportion of people aged from 44 to 65 years old	0.003	2.82	0.40E-4
Proportion of people who never married	0.002	4.01	0.27E-4
Proportion of people whose annual household income is higher than \$74,999	0.006	15.30	0.87E-4
Proportion of people with the commute time greater than 29 min	-0.002	-4.77	-0.30E-4
Proportion of Native Americans	-0.002 (0.001)	-3.41 (3.29)	-0.24E-4
Average number of people in a household	-0.881 (0.782)	-25.73 (49.27)	-0.013
Average number of workers in a household	0.460	11.56	0.007
Proportion of people whose household has no kid under 16 years old	-0.004	-4.90	-0.52E-4
Mean distance between activity locations to the nearest bike-sharing hubs	-0.183	-5.31	-0.003
Model statistics			
Number of observations	41,380		
Log-likelihood at zero, $LL(0)$	-28,419.75		
Log-likelihood at convergence, $LL(\beta)$	-10,231.11		
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.64		

Table 7

Mixed logit model results of the TAZ-level bike-sharing accessibility in southern Tampa. (Standard deviation of parameter estimate, in parentheses). Source: Created by the authors.

Variable description	Estimated parameter	t-Stat.	Marginal effects
Constant	-10.46	-3.31	/
Proportion of people with the commute time ranging from 20 to 29 min	0.18	2.91	0.0037
Proportion of people with the commute time more than 29 min	0.13	2.50	0.0028
Proportion of people whose annual household income is higher than \$74,999	0.04	2.82	0.0008
Proportion of people that is in the combined other racial group	1.09 (1.11)	4.75 (4.67)	0.0230
Proportion of households without kids under 16 years old	-0.06	-2.31	-0.0012
Model statistics			
Number of observations	178		
Log-likelihood at zero, $LL(0)$	-123.56		
Log-likelihood at convergence, $LL(\beta)$	-77.84		
$\rho^2 = 1 - LL(\beta)/LL(0)$	0.37		

many of the variables that are significant in the estimated individual-level model are still statistically significant, with the absolute values of their t -statistics greater than 1.96 (i.e., p -values less than 0.05). Most of the estimated parameters show the same signs and thus indicate the same disparity direction as the individual-level model does. However, several subgroups (e.g., females) are not statistically significant in the parcel-level model, revealing that data aggregation dilutes the disparity among individuals in groups defined by these sociodemographic attributes. Interestingly, variable *proportion of people whose household has no kid under 16 years old* produces a negative sign, meaning that people living in households with higher proportions of kids aged 16 are less likely to have bike-sharing accessibility, which contradicts the finding

from the individual-level data. However, this result is not very surprising because it has been well-known in geography that analytical results from multivariate analysis using aggregated data are dependent on the geographic units used for the analysis; different geographic scales may even result in contradictory results. This so-called modifiable unit areal problem (MUAP) poses a major challenge to using multivariate analysis of aggregated spatial data to formulate policies (Fotheringham and Wong, 1991).

Further, although results from the log-likelihood ratio test confirm the necessity of using a mixed logit model for analyzing the parcel-level data at a confidence level of 100% (see Table 5), most individual unobserved heterogeneity revealed by the random parameters are absorbed. Specifically, there are only four statistically significant random parameters in the parcel-level model. For example, the constant follows a normal distribution with a mean of -2.773 and a standard deviation of 2.573, meaning that the estimated coefficients are greater than 0 for 14.06% of the land parcels and less than 0 for 85.94% of them. This result indicates the distribution of bike-sharing accessibility is not horizontally equitable at the land parcel levels in southern Tampa, either. Additionally, variable *average number of people in a household* follows a normal distribution with a mean of -0.881 and a standard deviation of 0.782, revealing that the corresponding parameter is positive for 13% of the parcels and negative for 87% of them. This result indicates that when data are aggregated, the unobserved heterogeneity is also likely absorbed along with the observable individual heterogeneity.

When data are further aggregated to the TAZ level, the number of variables that are included in the estimated model (with p -values less than 0.05) is much fewer, meaning that more heterogeneity that could have been captured in the disaggregated dataset tends to be diluted. Specifically, only three variables statistically significant in the individual-level model are also significant in the TAZ-level model, including the *proportion of people with the commute time more than 29 min*, the *proportion of people whose annual household income is higher than \$74,999*, and the *proportion of households without kids under 16 years old*. Again, the number of university students in a household shows an

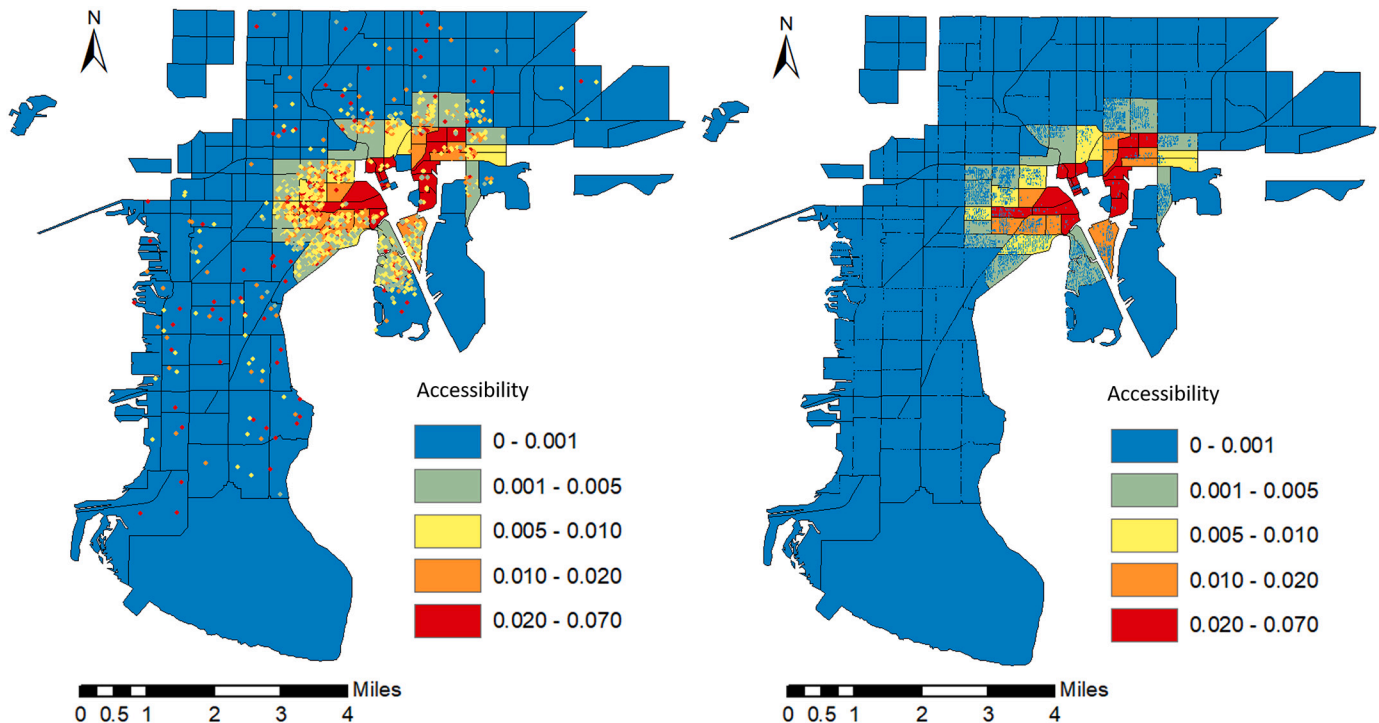


Fig. 2. Spatial distribution of bike-sharing accessibility in southern Tampa. Dots represent land parcels. The left figure contains all TAZs and parcels with bike-sharing accessibility greater than 0.001. The right figure contains all TAZs and parcels with bike-sharing accessibility less than and equal to 0.001. Source: Created by the authors.

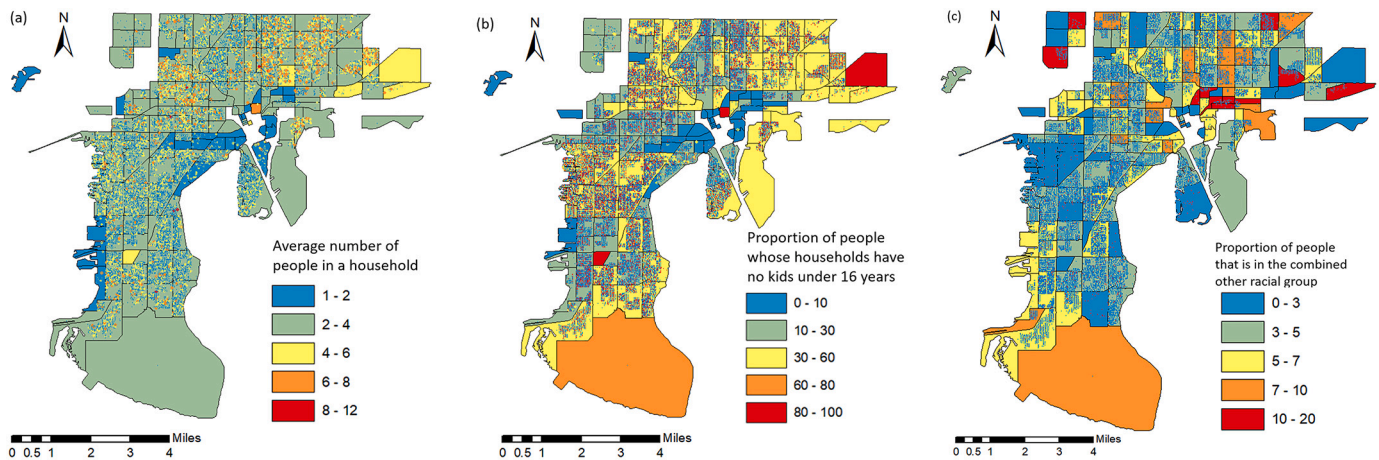


Fig. 3. Spatial distribution of sociodemographic attributes in southern Tampa. Dots represent land parcels. Three attributes are selected: (a) average number of people in a household, which is statistically significant in the parcel-level model but not in the TAZ-level model; (b) proportion of people whose households have no kids under 16 years old, which is statistically significant in both models; and (c) proportion of people that are in the combined other racial group, which is statistically significant in the TAZ-level model but not in the parcel-level model. Source: Created by the authors.

opposite impact on the distribution of bike-sharing accessibility as that in the individual-level model. Further, several variables that do not show significant impacts on the distribution of bike-sharing accessibility in the individual-level model are found to be statistically different from zero in a two-tailed *t*-test in the TAZ-level model. These variables include the proportion of people with commute time ranging from 20 to 29 min and the proportion of people that is in the combined other racial group.

The model estimated with TAZ-level data also produces different results in terms of unobserved heterogeneity. Only the parameter estimate for variable *proportion of people that is in the combined other racial group* is found to follow the normal distribution across TAZs. However, the variable pertinent to individuals with high annual household income, whose parameter estimate follows the normal distribution in the individual-level data, is found to be fixed across TAZs. These inconsistent results are, again, rooted in the fact that data aggregation absorbs the heterogeneity between different individuals. Note that in this case, we observe that as the aggregation level increases (i.e., the geographic unit of analysis becomes larger), the results tend to be more deviant from those in the individual-level model. However, this finding may not be a universal rule that holds for other transportation systems. As pointed out by Fotheringham and Wong (1991), the impact of the modifiable unit areal problem in multivariate analysis is unpredictable since the interactions between changes in the variance and covariance of different independent variables cannot be anticipated.

The spatial distributions of the accessibility metric and sociodemographic offer a better understanding of the above statistical analysis results. Fig. 2 confirms that the bike-sharing accessibility is concentrated in only a small portion of parcels and TAZs, mainly located in Downtown Tampa. Fig. 2 also reveals the smoothing effect of data aggregation. We see that within a large portion of TAZs, there are parcels with bike-sharing accessibility beyond the accessibility levels of these TAZs (e.g., red dots in TAZs shaded blue). Fig. 3(a) shows that parcels in the north side of the study area have higher average numbers of people in the households. These include parcels in the downtown where the bike-sharing accessibility is generally higher. Thus, the parcel-level model produces a positive parameter estimate for this attribute. Meanwhile, not all parcels in the north side are in the downtown and therefore have a bike-sharing accessibility less than 0.001. Thus, variable *average number of people in a household* produces a random parameter in the parcel-level model. However, at the TAZ level, the average number of people in a household is relatively homogenous (i.e., most areas are shaded green). As a result, this variable is not statistically significant in the TAZ-level model. From Fig. 3(b), we see that both TAZs and parcels beyond the downtown area have a higher proportion of people living in

households without kids under 16 years old. These areas are where the bike-sharing accessibility is low. Therefore, this attribute has a negative parameter estimate in both models. Finally, while the proportion of people that is in the combined other racial group is almost evenly distributed across different parcels, TAZs in the north side generally have higher proportions, as shown in Fig. 3(c). As a result, this attribute only produces a statistically significant parameter in the TAZ-level model. Again, because the north side covers areas with high and low bike-sharing accessibility, variable *proportion of people that is in the combined other racial group* generates a random parameter in the TAZ-level model.

These results offer strong evidence of the modifiable unit areal problem and unobserved heterogeneity in using aggregated data for transportation equity analysis. Data aggregation involves a smoothing effect so that the variation in a variable that could have been captured by the disaggregated data (observed heterogeneity) or could be modeled by heterogeneity models (unobserved heterogeneity) decreases as the aggregation level increases. Thus, different aggregation levels usually lead to different or even contradictory analytical results. This inconsistency raises a question whether it is reliable to use multivariate analysis results from data at a particular aggregation level for decision making and policy formulation. Fortunately, as evidenced in other studies (Bills and Walker, 2017; Hu and Wang, 2015), using individual-level disaggregated data can help mitigate the impacts of the MUAP. Thus, the disaggregated analysis approach presented in this paper shall be promising in addressing the MUAP as well. Thus, the use of disaggregated data along with unobserved heterogeneity modeling techniques, when available, is beneficial for transportation equity analysis. However, here we do not intend to claim that using aggregated data in transportation equity analysis is unreasonable or not useful. Instead, for planning agencies that do not have access to disaggregated data, the only feasible solution is aggregated data. However, equity analysts must bear in mind that analyses based on aggregated data may miss significant factors and/or unobserved heterogeneity. Thus, they should exert cautions while interpreting the analysis results from aggregated data. The purpose of this study is to provide empirical evidence of the existence of unobserved heterogeneity in transportation equity analysis.

4. Conclusions

Using individual data on bike-sharing accessibility and sociodemographic in southern Tampa, this paper carries out a case study on the existence of unobserved heterogeneity in analyzing the equity impacts of the Coast Bike Share system in terms of the accessibility it brings to

society. By grouping the individual accessibility values into two outcomes of having bike-sharing accessibility and not having bike-sharing accessibility, a mixed logit model was estimated to investigate the relationship between the accessibility outcomes and various socio-demographic attributes. Furthermore, models were also estimated using data aggregated to the parcel- and TAZ- levels to explore the impacts of data aggregation on model estimation results. The main findings are:

- Bike-sharing accessibility is relatively low in southern Tampa, resulting in a higher likelihood for individuals to not have bike-sharing accessibility in general. The bike-sharing accessibility is not evenly distributed in southern Tampa from both the horizontal and vertical equity perspectives. Specifically, the bike-sharing accessibility is concentrated in 2.08% of the individuals. Bike-sharing accessibility is higher for people with a higher number of tours in their daily travel itineraries, people aged between 19 and 44, individuals who have never married, those with an annual household income greater than \$74,999, and those living in a household with no kids under 16 years old on average. In contrast, females, married people, individuals with commute time greater than 29 min by car, those living in households with higher numbers of workers and retired adults, and those have longer mean distances between activity locations to the nearest bike-sharing hubs are underrepresented on average.
- A number of parameters follow normal distributions with different values of mean and standard deviation in the model estimated using individual-level data, demonstrating the presence of unobserved heterogeneity in the disaggregated dataset. Unobserved heterogeneity is also found in the parcel-level and TAZ-level models but with a fewer number of random parameters. The existence of unobserved heterogeneity reveals that relevant parameter estimates do not produce the same sign across the observations, therefore revealing different disparity directions for individuals in a population subgroup. Ignoring this effect would likely lead to incorrect parameter estimates and thus ineffective policy formulation. Thus, this finding demonstrates the necessity and importance of addressing unobserved heterogeneity in transportation equity analysis.
- Data aggregation absorbs the variation of a variable that could have been captured by the disaggregated data (observed heterogeneity) or been modeled by heterogeneity models (unobserved heterogeneity). As a result, as the aggregation level increases, fewer variables are statistically significant or produce a random parameter in the estimated models. Worse still, due to the modifiable unit areal problem, the use of aggregated data likely leads to different analytical results at different aggregation levels. This inconsistency makes it unreliable to apply modeling results from a particular level of aggregation for decision making and policy formulation. Instead, incorporating disaggregated data into transportation equity analysis when available can avoid this issue. Thus, it is preferable to use disaggregated data with heterogeneity methods for transportation equity analysis if possible.

This study provides empirical evidence for the existence of unobserved heterogeneity in transportation equity analysis and offers a timely alert for transportation equity analysts to this intriguing issue. However, only the most popular heterogeneity modeling approach (i.e., random parameters model) is investigated. Applying other relevant approaches such as latent class models, random parameters models with heterogeneity in mean and standard deviation and comparing the results would offer important methodological and planning implications for transportation planners. Further, for variables with random parameters, how to differentiate individuals truly underrepresented from others in the same subgroup is a challenging future research direction. Such research would be insightful for policymakers but requires more sophisticated modeling techniques that allow an analysis of the distributions of individual sociodemographic and transportation outcomes.

Also, the bike-sharing system investigated in this paper serves a limited geographic scale. Heterogeneity models are a sophisticated statistical modeling technique that can be applied to data sets of various sizes. We were not able to apply the analysis methods to other bike-sharing systems due to the lack of data. It would be an interesting and meaningful future research direction to apply the heterogeneity models to other transportation systems of different sizes (e.g., Citi Bike in New York) and structures (e.g., systems without station clusters). Finally, the estimation of a random parameters model using a huge amount of disaggregated data takes expensive computational and time resources. The development of efficient model estimation methods would also be an interesting and meaningful future research direction.

CRedit authorship contribution statement

Zhiwei Chen: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Xiaopeng Li:** Supervision, Funding acquisition, Writing - review & editing.

Acknowledgements

This research was supported by the U.S. National Science Foundation through Grants CMMI #1634738 and CMMI #1638355. We thank Dr. Sashikanth Gurram and Dr. Amy L. Stuart for providing the disaggregated activity and demographic data necessary for this analysis. We also thank Dr. Fred Mannering for his comments on heterogeneity modeling.

References

- Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accid. Anal. Prev.* 41 (1), 153–159.
- Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accid. Anal. Prev.* 43 (3), 1140–1147.
- Anastasopoulos, P.C., Labi, S., Karlaftis, M.G., Mannering, F.L., 2011. Exploratory state-level empirical assessment of pavement performance. *J. Infrastruct. Syst.* 17 (4), 200–215.
- Barajas, J.M., Drive, L.T., 2018. How equitable is bikesharing? Exploring population characteristics and access to employment. In: Paper Presented at the 97th Annual Meeting of the Transportation Research Board, Washington, D.C., January 7–11. Retrieved from: <https://trid.trb.org/view/1497044>.
- Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. *Transp. Res. A Policy Pract.* 30 (6), 415–429.
- Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. In: *Analytic Methods in Accident Research*, 3, pp. 56–91.
- Bills, T.S., Walker, J.L., 2017. Looking beyond the mean for equity analysis: examining distributional impacts of transportation improvements. *Transp. Policy* 54, 61–69.
- Bradley, M., Bowman, J.L., Griesenbeck, B., 2010. SACSIM: an applied activity-based model system with fine-level spatial and temporal resolution. *J. Choice Model.* 3 (1), 5–31. [https://doi.org/10.1016/S1755-5345\(13\)70027-7](https://doi.org/10.1016/S1755-5345(13)70027-7).
- Buzzelli, M., Jerrett, M., 2007. Geographies of susceptibility and exposure in the city: environmental inequity of traffic-related air pollution in Toronto. *Can. J. Reg. Sci.* 30 (2), 195–210.
- Carrier, M., Apparicio, P., Séguin, A.M., Crouse, D., 2014. The application of three methods to measure the statistical association between different social groups and the concentration of air pollutants in Montreal: a case of environmental equity. *Transp. Res. Part D: Transp. Environ.* 30, 38–52.
- Chen, Z., Guo, Y., Stuart, A.L., Zhang, Y., Li, X., 2019. Exploring the equity performance of bike-sharing systems with disaggregated data: a story of southern Tampa. *Transp. Res. A Policy Pract.* 130, 529–545.
- Chen, Z., Hu, Y., Li, J., Wu, X., 2020. Optimal deployment of electric bicycle sharing stations: model formulation and solution technique. *Netw. Spat. Econ.* 20 (1), 99–136.
- Coast Bike Share, 2019. Coast Bike Share Rental Agreement and Waiver. Coast Bike Share, Tampa, FL. Retrieved from: <http://coastbikeshare.com/terms/>.
- Currie, G., 2004. Gap analysis of public transport needs: measuring spatial distribution of public transport needs and identifying gaps in the quality of public transport provision. *Transp. Res. Rec.* 1895, 137–146.
- Delbosc, A., Currie, G., 2011. Using Lorenz curves to assess public transport equity. *J. Transp. Geogr.* 19 (6), 1252–1259.
- El-Geneidy, A., Levinson, D., Diab, E., Boisjoly, G., Verbich, D., Loong, C., 2016. The cost of equity: assessing transit accessibility and social disparity using total travel cost. *Transp. Res. A Policy Pract.* 91, 302–316.

- Fotheringham, A.S., Wong, D.W., 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plan. A* 23 (7), 1025–1044.
- Gao, O.H., Klein, R.A., 2010. Environmental equity in participation of the clean air school bus program: the case of New York state. *Transp. Res. Part D: Transp. Environ.* 15 (4), 220–227.
- Geurs, K.T., Van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *J. Transp. Geogr.* 12 (2), 127–140.
- Gliebe, J., Bradley, M., Ferdous, N., Outwater, M., Lin, H., Chen, J., 2014. Transferability of activity-based model parameters. *Transportation Research Board, Washington, D. C. SHRP2 Report*. Retrieved from: <https://trid.trb.org/view/1330766>.
- Golub, A., Martens, K., 2014. Using principles of justice to assess the modal equity of regional transportation plans. *J. Transp. Geogr.* 41, 10–20.
- Greene, W.H., 2007. *Nlogit Version 4.0 Reference Guide*. Econometric Software. Inc., Castle Hill, Australia.
- Guo, Y., Chen, Z., Stuart, A., Li, X., Zhang, Y., 2020. A systematic overview of transportation equity in terms of accessibility, traffic emissions, and safety outcomes: from conventional to emerging technologies. *Transp. Res. Interdiscip. Perspect.* 100091, 1–14.
- Gurram, S., Stuart, A.L., Pinjari, A.R., 2015. Impacts of travel activity and urbanicity on exposures to ambient oxides of nitrogen and on exposure disparities. *Air Qual. Atmos. Health* 8 (1), 97–114.
- Gurram, S., Stuart, A.L., Pinjari, A.R., 2019. Agent-based modeling to estimate exposures to urban air pollution from transportation: exposure disparities and impacts of high-resolution data. *Comput. Environ. Urban. Syst.* 75, 22–34.
- Guzman, L.A., Oviedo, D., Rivera, C., 2017. Assessing equity in transport accessibility to work and study: the Bogotá region. *J. Transp. Geogr.* 58, 236–246.
- Hamidi, Z., Camporeale, R., Caggiani, L., 2019. Inequalities in access to bike-and-ride opportunities: findings for the city of Malmö. *Transp. Res. A Policy Pract.* 130, 673–688.
- Harper, S., Charters, T.J., Strumpf, E.C., 2015. Trends in socioeconomic inequalities in motor vehicle accident deaths in the United States, 1995–2010. *Am. J. Epidemiol.* 182 (7), 606–614.
- Havard, S., Deguen, S., Zmirou-Navier, D., Schillinger, C., Bard, D., 2009. Traffic-related air pollution and socioeconomic status: a spatial autocorrelation study to assess environmental equity on a small-area scale. *Epidemiology* 20, 223–230.
- Hu, Y., Wang, F., 2015. Decomposing excess commuting: a Monte Carlo simulation approach. *J. Transp. Geogr.* 44, 43–52.
- Karer, A., Niemeier, D., 2013. Civil rights guidance and equity analysis methods for regional transportation plans: a critical review of literature and practice. *J. Transp. Geogr.* 33, 126–134.
- Karn Kaplan, S., Popoks, D., Prato, C.G., Ceder, A.A., 2014. Using connectivity for measuring equity in transit provision. *J. Transp. Geogr.* 37, 82–92.
- Kocak, T.K., 2019. *Investigating Air Pollution and Equity Impacts of a Proposed Transportation Improvement Program for Tampa*. M.S.P.H. Thesis. Department of Public Health, University of South Florida, Tampa, FL. Online at: <https://scholarcommons.usf.edu/etd/7832/>.
- Kravetz, D., Noland, R.B., 2012. Spatial analysis of income disparities in pedestrian safety in northern New Jersey: is there an environmental justice issue? *Transp. Res. Rec.* 2320, 10–17.
- Lei, C., Ouyang, Y., 2018. Continuous approximation for demand balancing in solving large-scale one-commodity pickup and delivery problems. *Transp. Res. B Methodol.* 109, 90–109.
- Levy, J.I., Greco, S.L., Melly, S.J., Mukhi, N., 2009. Evaluating efficiency-equality tradeoffs for mobile source control strategies in an urban area. *Risk Anal.* 29 (1), 34–47.
- Litman, T., Brenman, M., 2012. *A New Social Equity Agenda for Sustainable Transportation*. Victoria Transport Policy Institute, Victoria, Canada. Retrieved from: <https://www.vtpi.org/equityagenda.pdf>.
- Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Anal. Methods Accid. Res.* 11, 1–16.
- McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic Press, New York, pp. 105–142.
- McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *J. Appl. Econ.* 15 (5), 447–470.
- Mckenna, K., 2016. *Coasting around Downtown*. Tampa Magazine, Tampa, FL. Retrieved from: <https://tampamagazines.com/coasting-around-downtown/>.
- Morency, P., Gauvin, L., Plante, C., Fournier, M., Morency, C., 2012. Neighborhood social inequalities in road traffic injuries: the influence of traffic volume and road design. *Am. J. Public Health* 102 (6), 1112–1119.
- Noland, R.B., Lem, L.L., 2002. A review of the evidence for induced travel and changes in transportation and environmental policy in the US and the UK. *Transp. Res. Part D: Transp. Environ.* 7 (1), 1–26.
- Ogilvie, F., Goodman, A., 2012. Inequalities in usage of a public bicycle sharing scheme: socio-demographic predictors of uptake and usage of the London (UK) cycle hire scheme. *Prev. Med.* 55 (1), 40–45.
- Ossen, S., Hoogendoorn, S.P., 2011. Heterogeneity in car-following behavior: theory and empirics. *Transp. Res. C: Emerg. Technol.* 19 (2), 182–195.
- PlanHillsborough, 2017. *HC 2040 LRTP TAZ Population and Dwelling Units Shapefile*. PlanHillsborough, Tampa, FL. Retrieved from: http://www.planhillsborough.org/wp-content/uploads/2017/08/TAZ2010_2040_Pop_DU.zip.
- Pritchard, J.P., Tomasiello, D.B., Giannotti, M., Geurs, K., 2019. Potential impacts of bike-and-ride on job accessibility and spatial equity in São Paulo, Brazil. *Transp. Res. A Policy Pract.* 121, 386–400.
- Qian, X., Niemeier, D., 2019. High impact prioritization of bikeshare program investment to improve disadvantaged communities' access to jobs and essential services. *J. Transp. Geogr.* 76, 52–70.
- Raux, C., Zoubir, A., Geyik, M., 2017. Who are bike sharing schemes members and do they travel differently? The case of Lyon's "Velo'v" scheme. *Transp. Res. A Policy Pract.* 106, 350–363.
- Shaheen, S., Bell, C., Cohen, A., Yelchuru, B., 2017. *Travel Behavior: Shared Mobility and Transportation Equity*. US Department of Transportation, Washington D.C. Retrieved from: fhwa.dot.gov/policy/otps/shared_use_mobility_equity_final.pdf.
- Stat Trek, 2019. *Chi Square Calculator*. National Council of Teachers of Mathematics, Reston, VA. Retrieved from: <https://stattrek.com/online-calculator/chi-square.aspx>.
- Train, K.E., 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, U.K.
- Ursaki, J., Aultman-Hall, L., 2015. *Quantifying the Equity of Bikeshare Access in US Cities (No. TRC Report 15-011)*. University of Vermont, Transportation Research Center, Burlington, VT. Retrieved from: <https://rosap.nhtl.bts.gov/view/dot/36739>.
- US Census Bureau, 2010. *American Community Survey (ACS): PUMS Data*. Census Bureau, Washington, D.C.. Retrieved from: www.census.gov/programs-surveys/acs/data/pums.html.
- US Census Bureau, 2019. *Census Reporter: Hillsborough County-Tampa City (South)*. Census Bureau, Washington, D.C.. Retrieved from: <https://censusreporter.org/profiles/05000US12057-hillsborough-county-fl/>.
- Wang, K., Akar, G., 2019. Gender gap generators for bike share ridership: evidence from Citi Bike system in New York City. *J. Transp. Geogr.* 76, 1–9.
- Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
- Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. *Transp. Res. B Methodol.* 49, 39–54.