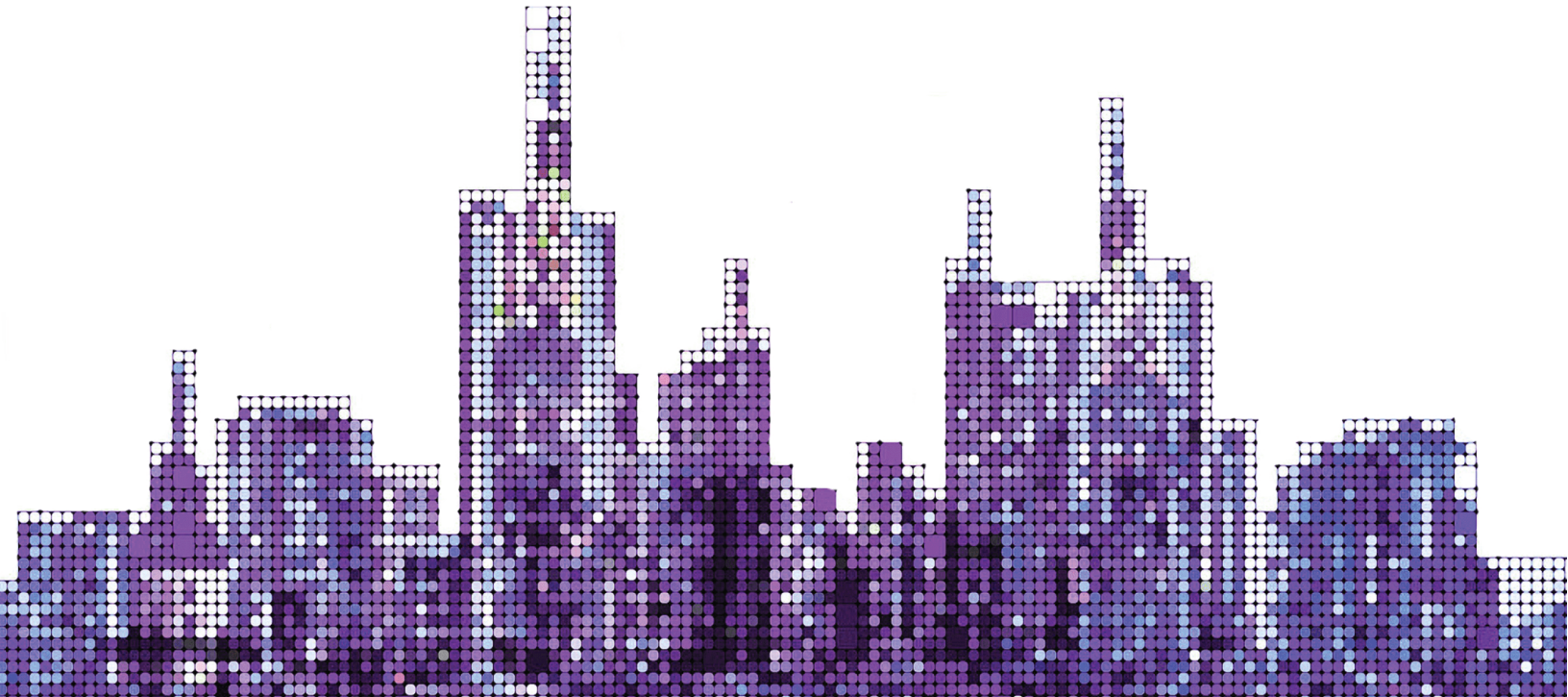
C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION
A USDOT University Transportation Center

New York University
Rutgers University
University of Washington
University of Texas at El Paso
The City College of New York

NY STATEWIDE BEHAVIORAL EQUITY IMPACT DECISION SUPPORT TOOL WITH REPLICA

July 2023



TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. NY Statewide Behavioral Equity Impact Decision Support Tool with Replica		5. Report Date July 2023
		6. Performing Organization Code:
7. Author(s) Joseph Y. J. Chow, Xiyuan Ren, Ngoc Hoang		8. Performing Organization Report No.
9. Performing Organization Name and Address Connected Cities for Smart Mobility towards Accessible and Resilient Transportation Center (C2SMART), 6 Metrotech Center, 4th Floor, NYU Tandon School of Engineering, Brooklyn, NY, 11201, United States		10. Work Unit No. 11. Contract or Grant No. 69A3551747119
12. Sponsoring Agency Name and Address Office of Research, Development, and Technology Federal Highway Administration 6300 Georgetown Pike McLean, VA 22101-2296		13. Type of Report and Period Final report, 3 /1/22-7/1/23 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

A NY statewide model choice model is developed to deterministically fit heterogeneous coefficients for trips along each census block-group OD pair conducted by each population segment within a random-utility-consistent framework. The proposed approach is to use inverse optimization (IO) to derive coefficients for each OD pair times population segment as an agent. This is only possible with ubiquitous population data. We call this a group-level agent-based mixed logit (g-AMXL) model, which is an extension of the AMXL model proposed by Ren and Chow (2022). The significance of g-AMXL is as follows. First, g-AMXL takes OD level (instead of individual level) trip data as inputs, which is efficient in dealing with ubiquitous datasets containing millions of observations. Second, preference heterogeneities are based on non-parametric aggregation of coefficients per agent instead of having to assume a distributional fit. Third, since each agent's representative utility function is fully specified, g-AMXL can be directly integrated into system design optimization models as constraints instead of dealing with simulation-based approaches required by mixed logit (MXL) models.

17. Key Words	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. http://www.ntis.gov		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 97	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page author

NY Statewide Behavioral Equity Impact Decision Support Tool with Replica

C2SMART Center is a USDOT Tier 1 University Transportation Center taking on some of today's most pressing urban mobility challenges. Using cities as living laboratories, the center examines transportation problems and field tests novel solutions that draw on unprecedented recent advances in communication and smart technologies. Its research activities are focused on three key areas: Urban Mobility and Connected Citizens; Urban Analytics for Smart Cities; and Resilient, Secure and Smart Transportation Infrastructure.

Some of the key areas C2SMART is focusing on include:

Disruptive Technologies

We are developing innovative solutions that focus on emerging disruptive technologies and their impacts on transportation systems. Our aim is to accelerate technology transfer from the research phase to the real world.

Unconventional Big Data Applications

C2SMART is working to make it possible to safely share data from field tests and non-traditional sensing technologies so that decision-makers can address a wide range of urban mobility problems with the best information available to them.

Impactful Engagement

The center aims to overcome institutional barriers to innovation and hear and meet the needs of city and state stakeholders, including government agencies, policy makers, the private sector, non-profit organizations, and entrepreneurs.

Forward-thinking Training and Development

As an academic institution, we are dedicated to training the workforce of tomorrow to deal with new mobility problems in ways that are not covered in existing transportation curricula.

Led by the New York University Tandon School of Engineering, C2SMART is a consortium of five leading research universities, including Rutgers University, University of Washington, the University of Texas at El Paso, and The City College of New York.

c2smart.engineering.nyu.edu

PI: Joseph Y. J. Chow
New York University
0000-0002-6471-3419

Xiyuan Ren
New York University
0000-0002-7719-7695

Ngoc Hoang
NYU Abu Dhabi

Disclaimer

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

Acknowledgements

In addition to the funding support from C2SMART, some of the researchers were supported by NYU's Summer Undergraduate Research Programs. Data shared by Replica are gratefully acknowledged.

Executive Summary

Transportation behavioral equity describes a normative condition in which no person or group is disadvantaged by a lack of access to transportation services they need to lead a dignified and meaningful life. One of the enduring challenges in statewide transportation planning is that consistent population travel data remains scarce, particularly for underserved and rural communities. Planning models are often only estimated using survey data collected by Metropolitan Planning Organizations for the urban areas, not for rural communities, which are at risk of increasing inequities.

This is changing with the availability of large-scale ICT data. Replica Inc. (2021) developed a nationwide synthetic population dataset that includes both sociodemographic information and trip/activity details. With this unique data opportunity, it is now possible to develop behavioral models for a range of different population segments such that new mobility scenarios can be analyzed to determine the forecasted ridership, revenue to those services, and change in consumer surplus. This unprecedented level of detail in equity analysis can provide support for statewide policymaking and programming.

However, conventional discrete choice models (DCMs) are not designed for ubiquitous datasets. Simple multinomial logit (MNL) models assume the coefficients are fixed for everyone and thus cannot capture the heterogeneity within a population segment. Stochastic models like mixed logit (MXL) and latent class models result in probabilistic parameters for each individual, which fits poorly within large-scale optimization models. Therefore, innovative models dealing with these limitations are theoretically essential and empirically critical to the development of an equity-based decision support tool for statewide transportation projects.

We initiated a one-year project from April 2022 to March 2023 to develop a behavioral equity impact decision support tool based on NY statewide synthetic population data provided by Replica Inc.:

- April 2022 to June 2022, Phase I: Replica Data Ingestion and Data Validation
- July 2022 to December 2022, Phase II: NY Statewide Mode Choice Model (g-AMXL)
- January 2023 to March 2023, Phase III: Equity-based Service Region Design

In this collaborative project with Replica Inc., we first checked the quality of Replica's datasets, which include the synthetic population and their trips on a typical Thursday and Saturday in the Fall 2019 season (September 2019 – November 2019). On the national scale, we

generated a data quality report through Replica's pipeline that compared the synthetic data and the census ground truth data. The report showed that at the census tract level, Replica's data are generally 99% accurate in population size and 95% accurate in sociodemographic information. We then conducted descriptive analysis on the statewide and city scale. We identified urban areas versus rural areas and disadvantaged communities versus non-disadvantaged communities in NY state. The results are consistent with our empirical knowledge and 70% overlap with the disadvantaged communities identified by NYSERDA. Trip details of four major cities (New York City, Buffalo, Rochester, and Syracuse) were visualized. The synthetic trip rate, mode share, and OD distribution revealed obvious differences between weekdays and weekends as well as among cities.

A NY statewide model choice model is developed to deterministically fit heterogeneous coefficients for trips along each census block-group OD pair conducted by each population segment within a random-utility-consistent framework. The proposed approach is to use inverse optimization (IO) to derive coefficients for each OD pair times population segment as an agent. This is only possible with ubiquitous population data. We call this a group-level agent-based mixed logit (g-AMXL) model, which is an extension of the AMXL model proposed by Ren and Chow (2022). The significance of g-AMXL is as follows. First, g-AMXL takes OD level (instead of individual level) trip data as inputs, which is efficient in dealing with ubiquitous datasets containing millions of observations. Second, preference heterogeneities are based on non-parametric aggregation of coefficients per agent instead of having to assume a distributional fit. Third, since each agent's representative utility function is fully specified, g-AMXL can be directly integrated into system design optimization models as constraints instead of dealing with simulation-based approaches required by mixed logit (MXL) models.

Synthetic trips on a typical weekday were used to calibrate the model. We considered six modes enabled by Replica's datasets, including private auto, public transit, on-demand auto, biking, walking, and carpool. Ten attributes were included into the utility function of mode choice, including travel time for auto, in-vehicle-time for public transit, access time for public transit, egress time for public transit, number of transfers for public transit, travel time for non-vehicle, trip monetary cost, mode constant for auto, mode constant for auto, mode constant for public transit, and mode constant for non-vehicle. Four mutually exclusive population segments were separated, including low-income working population, non-low-income working population, senior population, and students. The g-AMXL model with 120,740 agents took 2.79 hours to converge at the 26th iteration, with a rho value of 0.523. The average value of time (VOT) of LowIncome population is \$3.05/hour in NY state and \$14.59/hour in NYC. The average VOT of

NotLowIncome population is $13.78/hour in NY state and $18.74/hour in NYC. The average VOT of the Senior population is $9.98/hour in NY state and $5.31/hour in NYC. The average VOT of the Student population is $10.78/hour in NY state and $17.07/hour in NYC.

The estimated coefficients can be integrated directly into system optimization, which is the essential part of our equity-based decision support tool. In the proposed tool, we assume that there will be new mobility services entering into the market. Each service selects counties in NY state as the service zones, in which a few vehicles should provide bi-direction trip services to meet the demands on the selected OD links. Each vehicle has a maximum service distance and a maximum number of trips per day. The use cases of our tool include single-service region design and multi-service region assortment. Each use case outputs the optimal strategies and equity impact metrics given a budget level and one of the three objectives including: (1) maximizing the total revenue; (2) maximizing the total welfare (change of consumer surplus); (3) minimizing the welfare disparity between disadvantaged and non-disadvantaged communities. Equity metrics of the optimal solutions indicate that maximizing total revenue and maximizing total welfare result in similar service regions, where both might increase transportation inequities by increasing the welfare disparity by up to 0.59%. On the other hand, minimizing welfare disparity between disadvantaged communities and other communities can effectively decrease the welfare disparity by up to 7.37%, though this is at the cost of losing revenue. This suggests guidance by public policymakers is necessary if equity goals are desired.

The project has resulted in a database of NY statewide mode choice coefficients, a python package for a g-AMXL model, and a python package of equity-based service region design tool. One paper related to the modeling part was published. Two papers were prepared directly from the model building and system design. The work has supported a PhD student for portions of his dissertations and one undergraduate summer project at C2SMART.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Project background

Transportation policies, plans, and projects flow through state institutions due to the substantial cost of infrastructure and the need to assess transportation system performance, including equity implications (Karner et al., 2020). Over the last decade, there has been concentrated activity among Departments of Transportation (DOTs) at all levels considering transportation behavioral equity to provide a variety of groups with the access to transportation services they need to lead a dignified and meaningful life (Barajas, Natekal and Abrams, 2022; New York State Department of Transportation, 2023). For example, U.S. DOT Volpe Center team has built a Tableau-based prototype geospatial tool to visualize percentile ranks of a composite social vulnerability metric (see Figure 1.1). However, one of the enduring challenges is that consistent population travel data remains scarce, particularly for underserved and rural communities. Planning models are often only estimated using survey data collected by Metropolitan Planning Organizations (MPOs) for urban areas, not for rural communities. The lack of representative travel data at the state level can lead to exacerbations of inequity issues and the failure of new transportation projects.

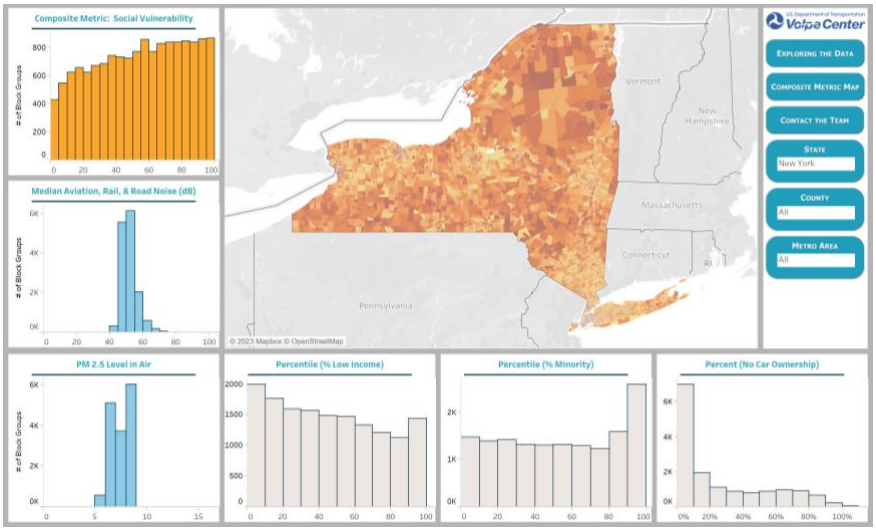


Figure 1.1. Transportation for Social Equity (TransportSE) tool developed by U.S. DOT Volpe Center (U.S. DOT Volpe Center, 2022)

This is changing with the availability of large-scale ICT data that have been used by a growing number of research institutions and companies to synthesize trip details of the population (Lee

et al., 2017; He et al., 2020; Hörl and Balac, 2021). Replica Inc. (2021) has developed a nationwide synthetic population dataset that includes both sociodemographic information (e.g., age, gender, income level, education, etc.) and trip details (e.g., trip length, trip duration, mode choice, route choice, etc.). With this unique data opportunity, it is now possible to develop behavioral models that can account for individuals in underserved communities. In other words, it is potentially possible to estimate discrete choice models (DCMs) for a range of different population segments living in different regions such that new mobility scenarios can be analyzed to determine the forecasted ridership, revenue to those services, and change in consumer surplus of targeted groups. For instance, with a statewide mode choice model we would be able to measure the change in consumer surplus to seniors living in Utica, NY, given a new mobility service operating there. An equity-based decision support tool with this unprecedented level of detail is essential to statewide policymaking and programming.

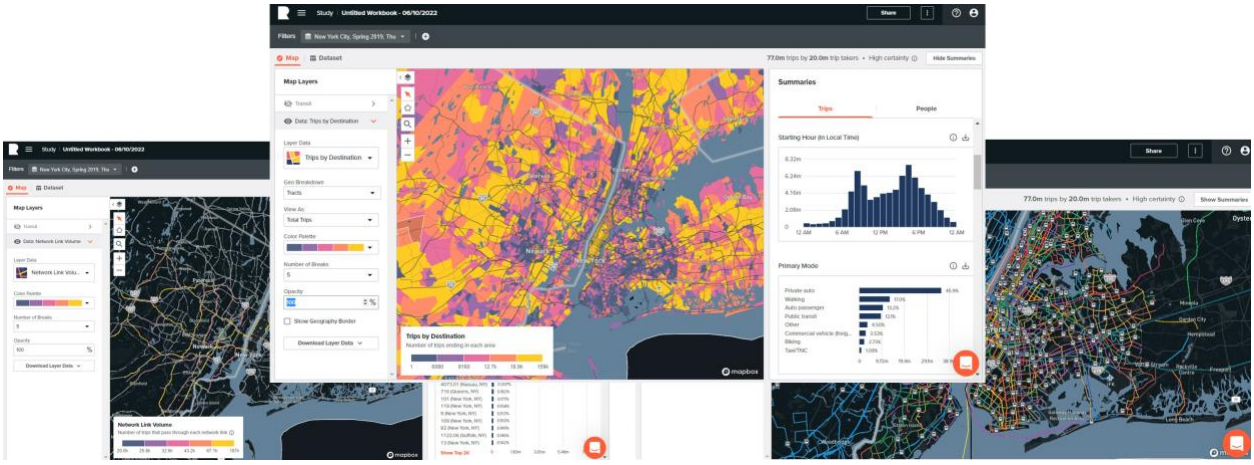


Figure 1.2. User interface of synthetic population data developed by Replica Inc. (Replica Inc., 2023)

However, challenges in calibrating a statewide choice model remain since conventional DCMs are not designed for the ubiquitous datasets (Ren and Chow, 2022). Simple multinomial logit (MNL) models cannot capture the heterogeneity in a population segment, assuming the parameters are fixed for everyone and that the decisions differ only because of unobservable utility and not due to differences in taste (Bowman and Ben-Akiva, 2001). This is not realistic because two individuals belonging to the same population segment (e.g., seniors) may have very different values of time if one is living in Utica and the other lives in Manhattan. On the other hand, heterogeneous taste preference models like mixed logit and latent class models (see Train, 2009) result in probabilistic parameters for each individual, which fits poorly within large-scale

optimization models. The demand response from a mixed logit model is possible but would require simulation to predict the response of travelers to determine the optimal solution (Pacheco et al., 2021). This lacks efficiency and consistency since simulation is time consuming and an optimization run would produce different results. Therefore, innovative models dealing with these limitations are theoretically essential and empirically critical to statewide transportation equity analysis.

1.2. Research objectives

In cooperation with Replica Inc., we initiated a one-year project from April 2022 to March 2023 to develop a behavioral equity impact decision support tool based on NY statewide synthetic population data and their trip details. The project contains three phases according to our research objectives.

Phase I: Replica Data Ingestion and Data Validation (April 2022 to June 2022)

The objective in this phase is to ingest Replica's data and check the quality. Datasets provided by Replica Inc. include the synthetic population and their trips on a typical Thursday and Saturday in the Fall 2019 season (September 2019 – November 2019). On the national scale, we generated a data quality report, to validate the information on population size, sociodemographic information, and trip volume. Descriptive analysis of trip details was conducted on the statewide and city scale. We used the data to identify urban versus rural areas and disadvantaged versus non-disadvantaged communities. Also, we picked four cities (New York City, Buffalo, Rochester, and Syracuse) to visualize their trip distribution, trip purpose, and mode share.

Phase II: NY Statewide Mode Choice Model (July 2022 to December 2022)

The objective in this phase is to develop an advanced model that deterministically fits heterogeneous coefficients for trips along each census block-group OD pair conducted by each population segment within a random-utility-consistent framework. We propose to use inverse optimization (IO) to derive coefficients for each OD pair times population segment as an agent. The proposed approach is to use inverse optimization (IO) to derive coefficients for each OD pair times population segment as an agent. This is only possible with ubiquitous population data. We call this a group-level agent-based mixed logit (g-AMXL) model, which is an extension of AMXL model proposed by Ren and Chow (2022). The significance of g-AMXL includes: (1) taking OD level (instead of individual level) trip data as inputs, which is efficient in dealing with ubiquitous

datasets containing millions of observations; (2) imputing preference heterogeneities that are based on non-parametric aggregation of coefficients per agent instead of having to assume a distributional fit; (3) fully specified utility function for each agent, which can be directly integrated into system design optimization models as constraints instead of dealing with simulation-based approaches required by mixed logit (MXL) models.

Synthetic trips on Thursday in 2019 Q4 are used to calibrate the model. We consider six modes enabled by Replica's datasets, including private auto, public transit, on-demand auto, biking, walking, and carpool. Ten attributes are included into the utility function of mode choice, including travel time for auto, in-vehicle-time for public transit, access time for public transit, egress time for public transit, number of transfers for public transit, travel time for non-vehicle, trip monetary cost, mode constant for auto, mode constant for auto, mode constant for public transit, and mode constant for non-vehicle. Four population segments are separated, including low-income population, not low-income population, senior population, and students.

Phase III: Equity-based Service Region Decision (January 2023 to March 2023)

The objective in this phase is to integrate coefficients estimated by g-AMXL into optimization models to build an equity-based decision support tool. We assume that there will be new mobility services selecting counties in NY state as operating zones, in which a few vehicles should provide bi-direction trip services to meet the demands on the selected OD links. Each vehicle has a maximum service distance and a maximum number of trips per day. Three objectives are considered: (1) maximizing the total revenue of the new service; (2) maximizing the total welfare (change of consumer surplus); (3) minimizing the pairwise welfare difference between disadvantaged and non-disadvantaged communities. We formulate linear programming (LP) and quadratic programming (QP) problems to get the optimal strategies given one of the three objectives and a budget level. Also, we construct six equity metrics to measure equity impacts of the new mobility services.

1.3. Organization of report

The project report is organized to firstly provide an overview of synthetic population, behavioral models for discrete choice analysis, and equity-based optimization in operational research. It is followed by a discussion on the proposed methodology including the group-level agent-based mixed logit (g-AMXL) model and the system optimization model to support equity-

based service region design. After that, descriptive analysis of Replica's data is presented at the national, statewide, and city level, along with the identification of urban versus rural areas and disadvantaged versus non-disadvantaged communities. The following section shows the g-AMXL model results, including performance metrics, distribution of coefficients, and value of time (VOT) analysis for each population segment. Finally, a section is devoted for integrating estimated coefficients into system optimization models to illustrate how our tool can support the design of equity-based service region of new mobility services.

- Section 2: overview of synthetic population, behavioral models, and equity-based optimization
- Section 3: proposed methodology
- Section 4: data validation and descriptive analysis
- Section 5: NY statewide mode choice modeling
- Section 6: equity-based service region optimization tool
- Section 7: conclusion
- Section 8: summary of research outputs and technology transfer

2. Overview of Synthetic Population, Behavioral Models, and Equity-based Optimization

2.1. Synthetic population and travel demand

2.1.1. General set up for the synthesis pipeline

Synthetic populations of travelers and their behavior details are an important basis for agent-based transport simulations, which are increasingly used in transport planning with the availability of information and communications technology (ICT) data (Hörl and Balac, 2021). Typical synthetic data include a synthetic population of households and persons with sociodemographic attributes and their daily activity patterns in time and space. The process of travel demand synthesis is usually composed of multiple steps (see Figure 2.1).

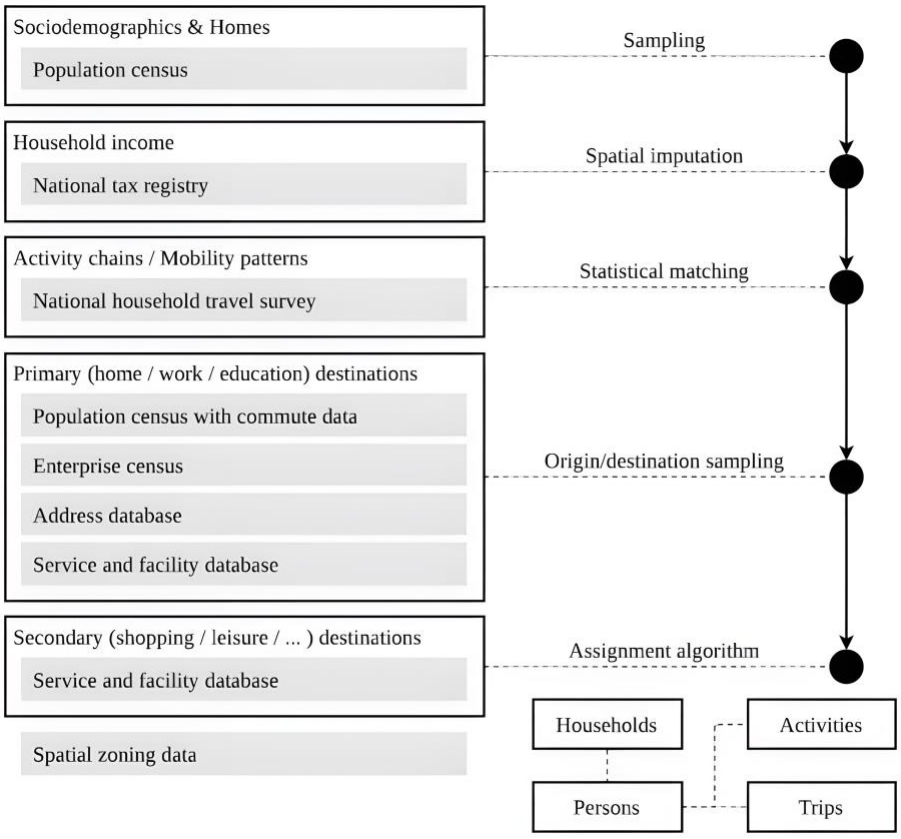


Figure 2.1. General set up for the synthesis pipeline (Hörl and Balac, 2021)

First, a population needs to be created with synthesis algorithms that process a small sample of the population (e.g., from a household travel survey) to create a model from which the full

population can be generated under certain assumptions. Related approaches include Iterative Proportional Fitting (Durán-Heras et al., 2018), Bayesian Networks (Saadi et al., 2018), and Deep Generative Modeling (Garrido et al., 2020). Second, statistical matching approaches (Namazi-Rad et al., 2017) are used to define daily activity patterns for synthesized persons, which make use of a household survey and attach whole activity chains based on sociodemographic attributes. Another choice in this step is to use activity-based models where a sequence of statistical models is applied to construct activity chains step by step (Joubert and Waal, 2020; Anda et al., 2021). Finally, locations of different activity types are synthesized in the targeted area. For commuting patterns, gravity models that quantify origin-destination flows by production and attraction of each zone are widely used (Ahrens and Lyons, 2021). For other activities, most approaches first determine all potential locations for certain activity purposes with constraints on opening times or daily time budgets (Esztergár-Kiss et al., 2020). After that, discrete choice models are applied to choose one alternative from the choice set (Yoon et al., 2012).

A major limitation is that synthetic travel demands are rarely replicable, reusable, and verifiable, which is due to the unavailability of data sources to public or the use of proprietary software (Hörl and Balac, 2021). To this end, data validation is an indispensable step before applying synthetic datasets to transportation planning and research.

2.1.2. Replica Methodology

Replica Inc. started as a research and development project inside of Alphabet's Sidewalk Labs in 2017. Its technical ambition is to support research and real projects across the public and private sectors by running large scale simulations of all movements for all places (Replica Inc., 2023). Replica's nationwide synthetic datasets are generated from three primary sources: (1) public use population census data; (2) proprietary location data from telecommunications and other IT infrastructure in the region; (3) field observation data from customer public agencies (ground truth). The synthesis pipeline is composed of three steps: population synthesis, persona training, and activity generation & calibration (see Figure 2.2).

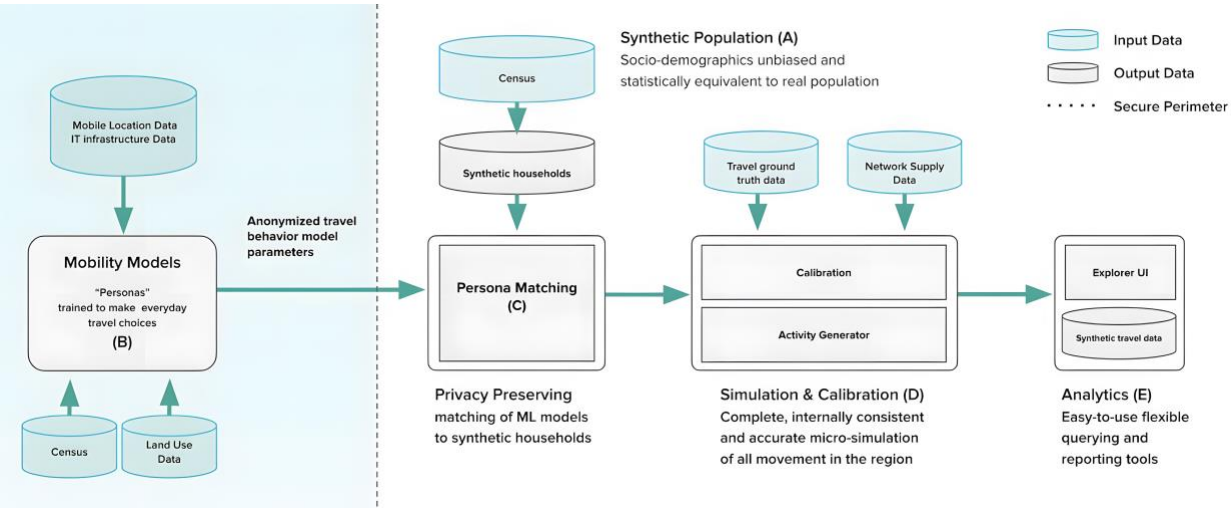


Figure 2.2. Replica synthesis pipeline diagram (Replica Inc., 2020)

The Population Generator (stage (A) in Figure 2.2) uses census data as well as other sources that contain sociodemographic and locational information, regional housing, and employment availability. It applies modeling and optimization algorithms to generate a synthetic population that is statistically similar to the census population in person-level attributes, individual household composition, as well as in aggregate. These synthetic people and households are assigned housing units and locations of workplaces and schools. All place assignments at the population synthesis stage represent outcomes of long-term choices that do not vary during the one week simulated time frame for each Replica. The result of stage (A) is a synthetic population with a home location and usual places of work (and/or school).

Persona Trainer (stage (B) in Figure 2.2) extracts behavioral patterns from de-identified mobile location data collected from mobile device(s) of real people. Persona training is informed by the region's spatial and sociodemographic data as well as other auxiliary sources of human activity in the region such as POI visits and credit transactions data. The purpose of creating personas from observed travel behaviors of real people are three-fold: preserve privacy of source data, provide explanation for travel behavior, and enable model sensitivity to urban design and policy changes. The personas can be used to reproduce complete daily activity sequences with each activity annotated with planned start and end times. At the Persona Matching (C) stage, each synthetic person is matched to a persona based on housing and working information.

The Activity Generation and Calibration (stage (D) in Figure 2.2) simulates a given day of the representative week to recreate the travel patterns for the entire metropolitan area. Simulation

includes feedback loops where calibration adjustments are applied to the values of the behavioral methods (e.g., mode choice, departure times, route choice) to better match the ground truth (e.g., transit counts). The final output is the Replica synthetic population with activities which contains a synthetically generated travel itinerary for each synthetic person on a given day of the season and serves as a single consistent source for further Analytics stage (E). It is not possible to re-identify any person from the sample locational data from the Replica synthetic population.

2.2. Behavioral models for discrete choice analysis

2.2.1. Discrete choice models (DCMs) for behavioral choice

As econometric models, DCMs assume individuals make choices by maximizing the overall utility they can expect to gain (Bowman and Ben-Akiva, 2001). Typically, decisions related to travel behavior include the choice of activity pattern (staying at home, working, or shopping), the destination, time-of-day, and mode & route choice between activities (Ettema et al., 2007; Ding et al., 2017; Nurul Habib, 2018). Advanced DCMs for behavioral choice can be divided into two categories. The first category treats choice dimensions with a nested structure, in the sequence from time frames to travel modes and from mode choice to route choice (Horni et al., 2016; Bowman and Ben-Akiva, 2001). A basic form is the nested logit model (NL) while a more advanced one follows a Markov decision process (MDP) and models choices as dynamic DCMs (Aguirregabiria and Mira, 2010; Västberg et al., 2020). Dynamic DCMs assume that individual $i \in P$ acts to maximizes the utility function defined by Eq. (2.1).

$$U_{ijt} = x_{ijt}\beta_{jt} + \varepsilon_{ijt} + \mu_t EV(i,j,t)\,, \qquad \forall i \in P, \forall j \in J, \forall t \in T \tag{2.1}$$

where t denotes the choice situation or time period. x_{ijt} denotes a set of observed variables of individual i choosing alternative j in situation t. β_{jt} is a set of coefficients reflecting preferences. $x_{ijt}\beta_{jt}$ and ε_{ijt} denotes the deterministic and random utility, which is aligned with conventional DCMs. $EV(i,j,t)$ is the expected utility of all possible alternatives in the remainder of the day, usually calculated using multi-dimensional integrals or backward induction with a relatively high computational cost (Västberg et al., 2020). μ_t is a coefficient defining the weight of expected

utility in choice situation t. Accordingly, the probability of individual i choosing alternative j in situation t is defined as Eq. (2.2).

$$P_{ijt} = \frac{e^{x_{ijt}\beta_{jt} + \varepsilon_{ijt} + \mu EV(i,j,t)}}{\sum_{j' \in J} e^{x_{ij't}\beta_{jt} + \varepsilon_{ij't} + \mu EV(i,j',t)}}, \quad \forall i \in P, \forall j \in J, \forall t \in T \tag{2.2}$$

The second category focuses on stochastic heterogeneity models, considering that preference may vary across different choice situations of different individuals. Up to this point, logit mixtures incorporating inter- and intra-individual heterogeneity are estimated with a maximum likelihood procedure (Becker et al., 2018; Krueger et al., 2021). For example, a mixed logit model (MXL) assumes that each individual i faces a choice among J alternatives. Then, the utility associated with each alternative $j = 1, \dots, J$ for individual i is defined as Eq. (2.3).

$$U_{ij} = x_{ij}\beta + \varepsilon_{ij}, \qquad \forall i \in P, \qquad \forall j \in J \tag{2.3}$$

where x_{ij} denotes a set of observed variables of individual i choosing alternative j. ε_{ij} is the random utility. The vector of tastes β is assumed to be a variate that varies across individuals according to $g(\beta|\Omega)$, where $g(.)$ is usually the Gaussian distribution with the mean and covariance included in Ω. Accordingly, the probability of individual i choosing alternative j is defined as Eq. (2.4).

$$P_{ij} = \int \frac{e^{x_{ij}\beta}}{\sum_{j' \in J} e^{x_{ij'}\beta}} g(\beta|\Omega) d\beta, \quad \forall i \in P, \qquad \forall j \in J \tag{2.4}$$

Despite a growing number of empirical studies, DCMs are not designed for equity analysis under the Big Data context (Ren and Chow, 2022). With a ubiquitous dataset, attributes from the whole population can be obtained instead of just from a sample (Ahas et al., 2009), and the individual tastes might not be normally distributed due to lacking personal information (Zhao et al., 2018). To this end, modelers should consider individual-specific estimations without complex assumptions of the conditional distribution.

2.2.2. Machine learning methods for behavioral choice

In recent years, there has been an emerging trend of using general-purpose machine learning models (MLs) to analyze individual choices (Wang et al., 2020b). General-purpose MLs for

behavior choice have both pros and cons. The pros are that these models allow flexible relationships between individuals' choices and explanatory variables, resulting in higher prediction accuracy than classical DCMs (Hagenauer and Helbich, 2017; Omrani, 2015; Pulugurta et al., 2013). The cons are that MLs are often criticized as "black-boxes" that are sensitive to hyperparameters and lack interpretability for modelers to explain the behavioral mechanism (Liao and Poggio, 2018; Sun et al., 2019; Wang et al., 2020b).

Besides these pros and cons widely discussed in existing studies, general-purpose machine learning models do not generally address the limitations of DCMs. On the one side, similar to the likelihood functions in DCMs, cross-entropy-based cost functions in MLs are also inefficient to optimize, given a huge dataset. On the other side, though the powerful automatic learning of MLs can capture complex behavior realism, it is at the cost of local irregularity and non-linearity of demand functions (LeCun et al., 2015; Liao and Poggio, 2018). Wang et al. (2020a) have pointed out the impacts of local irregularity on individual tastes. They found that the exploding and vanishing gradients in neural networks can result in extremely high or low sensitivities at the individual level that are opposite to domain knowledge. Moreover, with hundreds of parameters in deep learning models, it is almost infeasible to formulate the utility function, let alone generate demand functions and integrate them into optimization models. An innovative, domain-specific machine learning approach is necessary to deal with the ubiquitous datasets and build the link between demand and supply.

2.2.3. Inverse optimization (IO) for behavioral choice

Inverse optimization (IO) is initially used to impute missing optimization model coefficients from data that represents sub-optimal solutions of that optimization problem (Ahuja and Orlin, 2001; Burton and Toint, 1992). Given an optimization problem, an IO can be formulated to impute its left-hand-side constraint parameters and feasible regions (Ghobadi and Mahmoudzadeh, 2021). A typical IO problem is defined as follows: for a given prior θ_0 of missing coefficients and observed decision variables x^*, determine an updated coefficient set θ such that x^* is optimal while minimizing its L_1 norm from the prior, as shown in Eq. (2.5).

$$\min_{\theta} |\theta_0 - \theta| : x^* = \arg\min \{\theta^T x : Ax \leq b, x \geq 0\} \tag{2.5}$$

where A is the constraint matrix b is the vector of side constraint values. $Ax \leq b$ are constraints ensuring x^* is optimal (or the best choice). L_1 norm from a prior is used to regularize what would

otherwise be an ill-posed problem with infinite solutions. Ahuja and Orlin (2001) proved that Eq. (3.5) can be reformulated as a linear programming (LP) problem.

Though IO is less popular than general-purpose machine learning models, it has already been applied to traffic assignment, route choice, and activity scheduling problems (Chow and Recker, 2012; Hong et al., 2017; Chow, 2018; Xu et al., 2018). For instance, Chow and Recker (2012) proposed a multiagent framework for IO where a sample of individuals' trip scheduling data is obtained and used to infer parameters of individual activity scheduling. Xu et al. (2018) formulated the multiagent inverse transportation problem to estimate heterogeneous route preferences and proved that the IO approach could obtain heterogeneous link cost coefficients even when multinomial or mixed logit models would not be meaningfully estimated. Moreover, the potential of IO in modeling individual choice has been noticed by existing studies. Iraj and Terekhov (2021) emphasized the need for stochastic IO models in scenarios where constraints, objective, and prior parameters can be defined with domain knowledge.

In our project, we proposed a hybrid machine learning/econometric approach designed to estimate coefficients of mode choice from a statewide dataset. The approach is based on the IO method of estimating a random utility model that treats trips along each OD pair conducted by each population segment as an agent. The utility function is linear to ensure its compatibility with optimization models.

2.3. Equity-based optimization in operational research

2.3.1. Equity concerns in operational research

Many well-known operational research (OR) problems such as knapsack, scheduling or assignment problems have been considered from an equity perspective (Karsu and Morton, 2015). The discrete knapsack problem selects a set of items such that the total performance of the set is maximized under capacity constraints. Some implications have included equity is a concern as well as the total output maximization. For instance, Kozanidis (2009) formulated a linear knapsack problem with profit and equity objectives in order to achieve equitable resource allocations. Bertsimas et al. (2014) proposed a modeling framework for general dynamic resource allocation problems where there is a concern of equitably distributing the delay among the resource requests. Facility location problems also consider equity issues to ensure an equitable

service to the population. Geographic equity of access to the service facilities is considered as one of the main requirements for an applicable solution, especially in essential public facility location problems (Smith et al., 2013; Batta et al., 2014). Equity concerns naturally arise in vehicle routing problems considered in disaster relief contexts (Beamon and Balcik, 2008), in which one of the objectives is to ensure equitable service distribution to different affected areas (nodes). For instance, Perugia et al. (2011) developed a vehicle routing problem to model a home-to-work bus service. The objective is to achieve an equitable extra time distribution across customers, in which the extra time is defined as the difference between the bus transport time and the time of driving from home to work.

In transportation system design, equity over service users is also considered. For instance, equitable approaches are used in congestion pricing schemes to ensure fair treatment of the travelers that are categorized by income or geographic locations (Wu et al., 2012). In addition, equitable capacity utilizations among the participating hubs and warehouses is considered in collaborative freight assignments (Chan et al., 2004). Moreover, different concepts are reported such as spatial equity and temporal equity (Zhang and Shen, 2010). The spatial equity concerns the difference in convenience among users accessing a variety of mobility services while the temporal equity measures the difference of travel time, delay, and speed among users traveling with mobility services.

Another pair of concepts is equitability and balance (Karsu and Morton, 2015). Equitability is used for comparing access to services across a set of distinguishable entities. Balance is a special type of equity concern in which individuals or groups are not necessarily treated anonymously since they differ in needs, claims, or preferences. An optimal solution to balance problems may not give each agent the same proportion of transportation resources. To consider user preferences, Delle Site et al. (2022) calculated the expectation of the compensating variation based on a calibrated mode choice model as a key index in transportation equity analysis.

2.3.2. Commonly used inequity metrics

Equity concerns are usually incorporated into optimization models through the use of inequality metrics, which assign a scalar value to any given distribution showing the degree of inequity (Karsu and Morton, 2015). Optimization models that handle equity concerns are either designed as single objective models that minimize an inequity metric with efficiency metrics as constraints or as multicriteria models with a weight between equity and efficiency metrics. As inequity metrics are used to assess the disparity in a distribution, they are related to mathematical

concepts of dispersion and variance. Given w_i as the welfare of individual or group i based on a mobility service, commonly used inequity metrics are listed as follows:

- The range between the minimum and maximum levels of welfare, $(\max_i w_i - \min_i w_i)$ (McLay and Mayorga, 2013). A similar measure is $(\frac{\max_i w_i - \min_i w_i}{\min_i w_i} * 100\%)$ (Ramos and Oliveira, 2011).

- The deviation from the mean, ($\sum_{i \in I} |w_i - \bar{w}|$ or $\sum_{i \in I} (w_i - \bar{w})^2$, where $\bar{w} = \frac{\sum_{i \in I} w_i}{|I|}$) (Ogryczak et al., 2008).

- Variation of welfare, $(\frac{\sum_{i \in I} (w_i - \bar{w})^2}{|I|})$. Equivalently, the standard deviation is also used in some studies (Turkcan et al., 2011).

- Gini coefficient is a widely used income inequality measure. The Gini coefficient has the following formula: $\frac{\sum_{i \in I} \sum_{j \in J} |w_i - w_j|}{2|I| \sum_{i \in I} w_i}$, where I and J denote the same individual or group set (Ohsawa et al., 2008).

- Sum of pairwise (absolute) differences which is the sum of absolute differences between all pairs in consideration (Lejeune and Prasad, 2013). The sum of pairwise differences has the following formular: $\sum_{i \in I} \sum_{j \in J} |w_i - w_j|$, where I and J can be different.

Inspired by these studies, our project first developed a NY statewide mode choice model. We then constructed inequity metrics based on the welfare measures enabled by this model. Finally, we adopted optimization models similar to existing studies to design the service region of a new mobility service.

3. Proposed Methodology

The scope of the methodology is shown in Figure 3.1, which contains the ingestion and validation of Replica's synthetic data, NY statewide mode choice model estimation, and equity-based service region design for a new mobility service. Besides a behavioral equity decision support tool, some intermediate outputs of our project are also deliverable, including a nation-scale data qualification report, an OD-level mode choice dataset, and coefficients reflecting statewide mode choice preferences.

With ubiquitous datasets, our proposed methodology can be used to capture statewide mode choice preferences in which each population segment traveling along each block-group OD pair has a unique set of coefficients. These coefficients can be further used to forecast the ridership, revenue, and equity impacts of new mobility scenarios to support statewide policymaking and system design. To the best of our knowledge, no other methodology outputs all these measures.

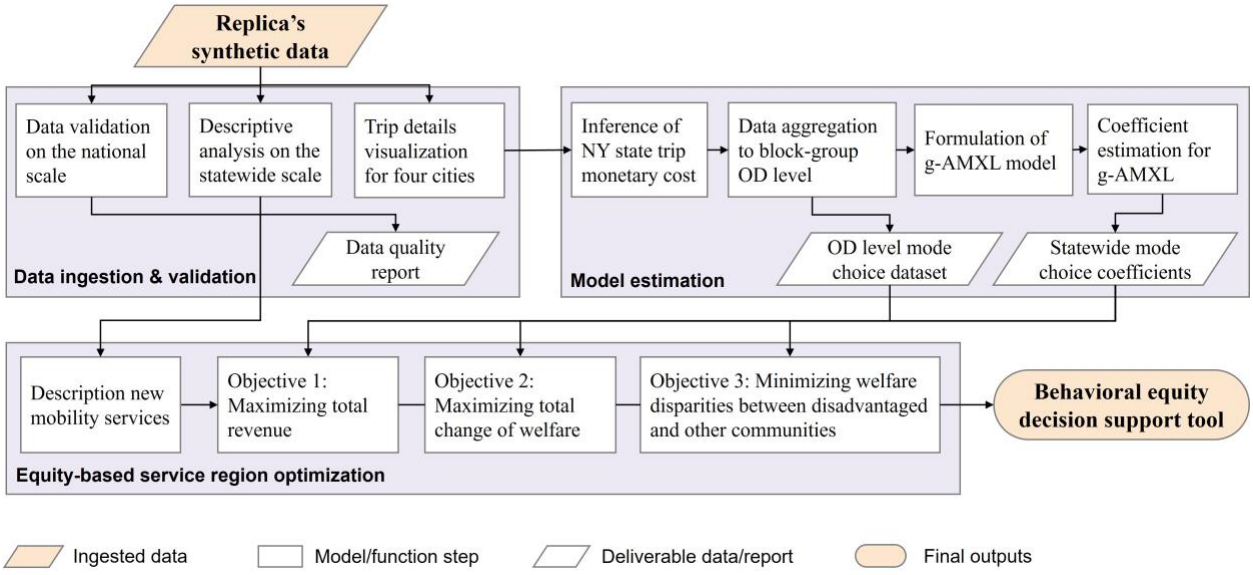


Figure 3.1. Flow chart of building a behavioral equity decision support tool with Replica's synthetic data

3.1. Data preparation

3.1.1. Replica's synthetic population data

The datasets provided by Replica Inc. include a Population table and a Trip table for NY state and the Fall 2019 season (September 2019 – November 2019). The Population table contains records for each synthetic person by Replica. The information in each record includes a set of attributes describing the person's demographics and the block-group assignment for their work, home, and school locations. Table 3.1 lists part of the field definitions of the Population table used in our project. Complete field definitions can be found in Appendix A.

Table 3.1. Field definitions of the Population table (used in our project)

Field name	Data type	Sample value	Description
person_id	String	144080185795050000-01	Unique identifier of a person.
household_id	String	141025892089804000	Unique identifier of a household.
BLOCKGROUP	String	360810201001	The US Census Bureau-assigned FIPS code of the census block-group containing the housing unit.
age	Integer	19	Age, in years old, assigned to the person.
school_grade_attending	String	not_attending_school	Current grade level assigned to a person. Valid values include: • graduate • kindergarten • not_attending_school • school • undergraduate
household_income	Integer	408,500	Total income of the household per year.
household_size	String	3_person	Number of persons that makeup the household. Valid values include: • 1_person • 1_person_group_quarters • 2_person

			• 3_person • 4_person • 5_person • 6_person • 7_plus_person

The Trip Table contains each trip produced by Replica for a given synthetic person. Characteristics included with each trip record include its origin, destination, travel mode, travel time, and distance. Records from the Trips table can be joined to records in the Population table using the unique person_id and household_id fields. Table 3.2 lists part of the field definitions of the Trip table used in our project. Complete field definitions can be found in Appendix A. For public transit trips, we further calculated the access time, egress time, in-vehicle time, and number of transfers based on the field network_link_ids.

Table 3.2. Field definitions of the Trip table (used in our project)

Field name	Data type	Sample value	Description
activity_id	String	15323941267251300000	A randomly assigned unique identifier defined for each trip.
person_id	String	144080185795050000-01	Unique identifier of a person.
mode	String	PUBLIC_TRANSIT	Primary transportation mode used for the trip. In the case of multiple travel modes, only the primary mode of travel across a set of trip segments is included. Valid options are: • PRIVATE_AUTO: Trips made by drivers in private auto vehicles • PUBLIC_TRANSIT: Trips that primarily used public transit, such as buses, light rail, and subways • ON_DEMAND_AUTO: Trips made by passengers in a Taxi or using a Transportation Network Company (TNC) such as Uber or Lyft • BIKING: Trips made by people biking. Replica does not model

			scooter trips and does not separate out e-bike trips • WALKING: Trips made by people walking • CARPOOL: Trips made by passengers in private auto vehicles. Sum Carpool and Private Auto trips to get the total number of people who traveled in private autos • COMMERCIAL: Trips made by medium and heavy trucks
start_time	TIME	2019-01-10 06:08:00 America/New_York_City	Date and 24-hour time of trip start, reported as yyyy-mm-dd hh:mm:ss timezone
end_time	TIME	2019-01-10 07:11:04 America/New_York_City	Date and 24-hour time of trip end, reported as yyyy-mm-dd hh:mm:ss timezone
duration_minutes	Integer	63	Duration of trip in minutes, calculated as the difference between the trip start_time and end_time.
distance_miles	Float	10.79	Distance in miles measured along the trip route.
origin_bgrp	String	360810201001	The US Census Bureau-assigned FIPS code of the block group from which the trip originated.
destination_bgrp	String	360810238330	US Census Bureau-assigned FIPS of the block group in which the trip ended.
network_link_ids	Set	[28492853,28493983,…]	A set of road segment ID that the trip is associated with.

3.1.2. Inference of trip monetary cost

An issue with Replica's Trip Table is that it does not contain trip monetary cost, probably due to the difficulty of inferencing monetary cost for nationwide trips. This is possible if we only focus on NY state. Though we cannot get trip cost directly from Replica's data, we inferred one for each trip based on information such as trip mode, trip length, and trip origin & destination (see Table 3.3).

For private auto trips related to New York City, we assumed the average parking and toll cost for each trip (see (Chow et al., 2020)): $4.94 for each Manhattan related trip and $1.37 for each non-Manhattan-related trip. For private auto trips outside New York City, we assumed the trip cost is $0.07/mile, which is based on 2019 NY gas price and the average mpg of cars and SUVs. The cost of public transit is $1.38/trip for senior population and $2.75/trip for other population. On-demand auto cost is calculated according to the standard metered fare from NYC Taxi and Limousine Commission (TLC) [1]. Since Replica's data cannot separate e-bike trips or scooter trips, we assumed zero cost of biking and walking trips. Carpool trips are assumed to be half the cost of private auto trips.

Table 3.3. Monetary cost inference for private auto, public transit, and on demand auto

1.Private auto		
Item	**Unit price**	**Description**
Manhattan-related cost	$4.94/trip	For trips of which the origin or destination is in Manhattan.
NYC non-Manhattan-related cost	$1.37/trip	For trips of which the origin or destination is in NYC but not in Manhattan.
Cost for other trips	$0.07/mile	For trips of which the origin and destination are outside of NYC. This is calculated based on 2019 NY gas price ($2.542/gallon[2]) divided by the average mpg of cars and SUVs (36.33 mpg[3])
2. Public transit		
Item	**Unit price**	**Description**
Cost for non-senior population	$2.75/trip	For subways and buses trips.
Cost for senior population	$1.38/trip	Riders who are 65 or older have a 50% discount.
3.On demand auto (standard metered fee for taxi in 2019)		
Item	**Unit price**	**Description**
Initial charge	$2.50/trip	An initial charge for the trip.
Metered charge	$2.50/mile	Trip cost charged according to trip length.

[1] https://www.nyc.gov/site/tlc/passengers/taxi-fare.page
[2] https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=emm_epmr_pte_y35ny_dpg&f=m
[3] https://techxplore.com/news/2022-04-vehicles-average-mpg.html, https://www.indyautoman.com/blog/best-mpg-suv, https://nhts.ornl.gov/documentation

MTA state surcharge	$0.50/trip	For all trips that end in New York City or Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange or Putnam Counties.
Improvement surcharge	$0.30/trip	--
Overnight surcharge	$0.50/trip	For trips from 8 p.m. to 6 a.m.
Rush hour surcharge	$0.50/trip	For trips from 4 p.m. to 8 p.m. on weekdays
Tax	8.88% of the trip cost	--
Tips	20% of the trip cost	--

3.1.3. Data aggregation and population segmentation

Since the statewide dataset is quite large, it is impossible to estimate model at the individual level. Hence, we aggregated the mode choice observations into block-group OD level. Moreover, we separated four population segments: NotLowIncome Population, LowIncome Population, Senior Population, and Student Population. First, we used the field school_grade_attending to separated Student Population. We then used the field age to separated Senior population (age>=65). To separate NotLowIncome and LowIncome Population, we referred to 2019 U.S. Federal Poverty Guidelines[4], in which a low-income household is defined as: 1-person family with annual income lower than $12,490, 2-person family with annual income lower than $16,910, 3-person family with annual income lower than $21,330, 4-person family with annual income lower than $25,750, 5-person family with annual income lower than $30,170, 6-person family with annual income lower than $34,590, 7-person family with annual income lower than $39,010, and 8-person family with annual income lower than $43,430.

The dataset contains 120,740 rows in total. Each row contains the mode choice information of an agent, including the population segment ID, block-group ID of the origin and destination, number of trips per day along the OD pair, average travel time of each mode, average monetary cost of each mode, and the mode share.

[4] https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references/2019-poverty-guidelines

Table 3.4. Field definitions of the OD level mode choice dataset

Field name	Data type	Sample value	Description
Segment_ID	String	NotLowIncome	ID of the population segment.
origin_brgp	String	360810201001	ID of the block group from which the trips originated.
destination_brgp	String	360810238330	ID of the block group from which the trips ended.
Trip_num	Integer	502	Number of trips per day along the OD pair.
Dur_private_auto	Float	4.88	Average travel time (min) of private auto.
Dur_access	Float	2.68	Average access time (min) of public transit.
Dur_egress	Float	2.84	Average egress time (min) of public transit.
Dur_in_vehicle	Float	5.12	Average in-vehicle time (min) of public transit.
Num_transfer	Float	0.0	Average number of transfers of public transit.
Dur_on_demand_auto	Float	5.47	Average travel time (min) of on-demand auto.
Dur_biking	Float	7.25	Average travel time (min) of biking.
Dur_walking	Float	21.42	Average travel time (min) of walking.
Dur_carpool	Float	6.77	Average travel time (min) of carpool.
Cost_private_auto	Float	4.94	Average monetary cost ($) of private auto.
Cost_public_transit	Float	2.75	Average monetary cost ($) of public transit.
Cost_on_demand_auto	Float	5.89	Average monetary cost ($) of on-demand auto.
Cost_biking	Float	0	Average monetary cost ($) of biking.
Cost_walking	Float	0	Average monetary cost ($) of walking.
Cost_carpool	Float	2.47	Average monetary cost ($) of carpool.
Pro_private_auto	%	52.07%	Mode share of private auto.
Pro_public_transit	%	10.05%	Mode share of public transit.
Pro_on_demand_auto	%	8.43%	Mode share of on-demand auto.
Pro_biking	%	7.97%	Mode share of biking.
Pro_walking	%	14.45%	Mode share of walking.
Pro_carpool	%	7.03%	Mode share of carpool.

3.2. Estimation of statewide mode choice model

The proposed model is an extension of agent-based mixed logit (AMXL) model in Ren and Chow (2022)'s study, which provides deterministic and individual-specific estimation that can be efficiently integrated into optimization models. We call our model a group-level agent-based mixed logit (g-AMXL) model.

Compared with AMXL, the significance of g-AMXL is two-fold. On the one side, g-AMXL treats a group of choice observations (instead of a single choice observation per individual in AMXL) as an agent, which makes g-AMXL more efficient in dealing with large datasets (with millions of choice observations). On the other side, g-AMXL allows multiple fixed-point priors (instead of only one in AMXL), which enables g-AMXL to identify latent classes of preferences (this can also be used to tease out irregular data points, see Section 3.2.3). The g-AMXL is applied in this study to estimate NY statewide mode choice behavior.

3.2.1. Architecture of group-level agent-based mixed logit (g-AMXL) model

We considered six modes in Replica's datasets, including private auto, public transit, on-demand auto, biking, walking, and carpool. Ten attributes were included into the utility function of mode choice, including travel time for auto, in-vehicle-time for public transit, access time for public transit, egress time for public transit, number of transfers for public transit, travel time for non-vehicle, trip monetary cost, mode constant for auto, mode constant for auto, mode constant for public transit, and mode constant for non-vehicle. Therefore, the utility derived from choosing the six modes can be defined as Eq. (3.1) - (3.6).

$$V_{private\_auto,i} = \theta_{auto\_tt,i} time_i^{private\_auto} + \theta_{cost,i} cost_i^{private\_auto} + asc_{auto,i}\,, \qquad \forall i \in I \tag{3.1}$$

$$V_{public\_transit,i} = \theta_{transit\_at,i} access\_t_i + \theta_{transit\_et,i} egress\_t_i + \theta_{transit\_ivt,i} in\_vehicle\_t_i + \theta_{transit\_trans,i} num\_transfer_i + \theta_{cost,i} cost_i + asc_{transit,i}\,, \qquad \forall i \in I \tag{3.2}$$

$$V_{on\_demand,i} = \theta_{auto\_tt,i} time_i^{on\_demand} + \theta_{cost,i} cost_i^{on\_demand} + asc_{auto,i}\,, \qquad \forall i \in I \tag{3.3}$$

$$V_{biking,i} = \theta_{non\_auto\_tt,i} time_i^{biking} + \theta_{cost,i} cost_i^{biking} + asc_{non\_vehicle,i}\,, \qquad \forall i \in I \tag{3.4}$$

$$V_{walking,i} = \theta_{non\_auto\_tt,i} time_i^{walking} + \theta_{cost,i} cost_i^{walking} + asc_{non\_vehicle,i}\,, \qquad \forall i \in I \tag{3.5}$$

$$V_{carpool,i} = \theta_{auto\_tt,i} time_i^{carpool} + \theta_{cost,i} cost_i^{carpool} + asc_{auto,i}\,, \qquad \forall i \in I \tag{3.6}$$

where i denotes the unique ID of an agent (composed of origin_brgp, destination_brgp, and Segment_ID); I is the set of total agents; $V_{private\_auto,i}$, $V_{public\_transit,i}$, $V_{on\_demand,i}$, $V_{biking,i}$, $V_{walking,i}$, $V_{carpool,i}$ are utilities of trips belonging to agent i with different modes; $time_i^*$ and $cost_i^*$ are the travel time and monetary cost of different modes; $access\_t_i$, $egress\_t_i$, $in\_vehicle\_t_i$, $\theta_{transit\_nt,i}$ are the access time, egress time, in-vehicle time, and number of transfers for public transit. All these observed variables are from Replica's synthetic data, and $\theta_{auto\_tt,i}$, $\theta_{transit\_at,i}$, $\theta_{transit\_et,i}$, $\theta_{transit\_ivt,i}$, $\theta_{transit\_nt,i}$, $\theta_{transit\_trans,i}$, $\theta_{cost,i}$, $asc_{auto,i}$, $asc_{transit,i}$, $asc_{non\_vehicle,i}$ are 10 coefficients per agent to be estimated as a mixed logit model with deterministic tastes.

3.2.2. Estimation framework for g-AMXL

Inspired by the works of Chow and Recker (2012) and Xu et al. (2018), our study formalizes the multiagent inverse utility maximization (MIUM) problem to estimate mode choice coefficients. Consider within each agent $i \in I$, there is a population N of individuals behaviorally seek to select to maximize their overall utilities. In line with the random utility theory, the total utility derived from an individual $n \in N$ choosing mode j is defined in Eq. (3.7).

$$U_{nj} = V_{nj} + \varepsilon_{nj} = \theta_i X_{nj} + \varepsilon_{nj}, \qquad \forall n \in N, j \in J, i \in I \tag{3.7}$$

where U_{nj} is the total utility, which is composed of a deterministic utility V_{nj} and a Gumbel-distributed random utility ε_{nj}. X_{nj} is a set of observed variables related to individual n choosing alternative j. θ_i is a vector of mode choice coefficients of individuals in agent i. By doing this, we capture individual heterogeneities within agent i with ε_{nj} and we assume individuals in the same agent share the same set of coefficients θ_i. According to Berry et.al (1995), the relationship between the utility function and the mode share can be defined as Eq. (3.8).

$$\frac{S_{ij}}{S_{ij^*}} = \exp\left(\frac{V_{ij}}{V_{ij^*}}\right), \qquad \ln(S_{ij}) - \ln(S_{ij^*}) = \theta_i(X_{nj} - X_{nj^*}), \qquad \forall i \in I, j, j^* \in J, j \neq j^* \tag{3.8}$$

where S_{ij} is the market share of mode j in agent i; $\ln(S_{ij}) - \ln(S_{ij^*})$ is called inverted market share, which can be measured as observed variables X_{nj}, X_{nj^*} and coefficients θ_i. The agent-level coefficient set θ_i can be estimated by solving a MIUM problem under L_2-norm as a convex quadratic programming (QP) problem, as illustrated in Eq. (3.9) – (3.12).

$$\min_{\theta_0^k, \theta_i} \sum_{k \in K} \sum_{i \in I^k} (\theta_0^k - \theta_i)^2 \tag{3.9}$$

$s.t.$

$$\theta_i(X_{nj} - X_{nj^*}) \geq \ln(S_{ij}) - \ln(S_{ij^*}) - tol, \qquad \forall i \in I, j, j^* \in J, j \neq j^* \tag{3.10}$$

$$\theta_i(X_{nj} - X_{nj^*}) \leq \ln(S_{ij}) - \ln(S_{ij^*}) + tol, \qquad \forall i \in I, j, j^* \in J, j \neq j^* \tag{3.11}$$

$$\theta_0^k = \frac{1}{|I^k|} \sum_{i \in I^k} \theta_i, \qquad \forall k \in K \tag{3.12}$$

where θ_0^k is the k^{th} fixed-point prior corresponding to a latent class; tol is a manually set tolerance to draw a balance of goodness-of-fit and feasible solutions (a larger value of tol leads to a lower goodness-of fit while a higher proportion of feasible solutions). A recommend range of tol is [0.1,1.5][5]. In our study, we set $tol = 0.5$. Eq. (3.12) makes sure that the estimated agent coefficients have a consistency with one of the fixed-point priors. I^k denotes the agent set of the k^{th} latent class, which can be determined by applying the K-Means algorithm to θ_i. The objective is quadratic while the constraints are linear.

Solving the model in Eq. (3.9) – (3.12) as a single QP would be computationally costly as it would lead to highly diagonal sparse matrix. Instead, we solved Eq. (3.9) – (3.11) $|I|$ times with θ_i as the decision variables in each iteration (which are much smaller QPs). At the end of each iteration, we applied K-Means algorithm to θ_i to identify latent classes, updated the fixed-point priors, and check if the stopping criteria has been reached. If reached, then we output the estimated agent-specific coefficients θ_i. Otherwise, we use the updated fixed-point priors and go to the next iteration. The iterations continue until all priors θ_0^k stabilize (see Xu et al. (2018) for an example of this kind of decomposition for the |K|=1 case). The problem can be solved using any optimizer software or package that can handle QP like Gurobi, CVXPY, etc.

[5] Since we know the range of market share S_{ij} is [0,1], we can calculate the range of $\ln(S_{ij}) - \ln(S_{ij^*})$. 0 ± 0.1 covers 5% of the range and 0 ± 1.5 covers 40% of the range. In other words, $tol = 0.1$ leads to a 95% accuracy and $tol = 1.5$ leads to a 60% accuracy.

The convergent iterative algorithm used in our study is the Method of Successive Average (MSA). The whole estimation approach is summarized in Algorithm 3.1. The stopping rule is set to ε'=0.001 considering time to converge.

Algorithm 3.1. Mode choice g-AMXL estimation

1. Given observed variables X_{ij}, initialize with n=0, tol=0.5, and fixed-point prior $\theta_0^{(n)} = [0, \dots, 0]$
2. For each $i \in I$, solve a QP to get θ_i:

$$\min_{\theta_i} (\theta_0^{(n)} - \theta_i)^2$$
$$s.t.$$
$$\theta_i(X_{nj} - X_{nj^*}) \geq ln(S_{ij}) - ln(S_{ij^*}) - tol, \qquad \forall j, j^* \in J, j \neq j^*$$
$$\theta_i(X_{nj} - X_{nj^*}) \leq ln(S_{ij}) - ln(S_{ij^*}) + tol, \qquad \forall j, j^* \in J, j \neq j^*$$

3. Apply the K-Means algorithm to θ_i to identify k latent classes $I^1, \dots, I^{|K|}$, and update the fixed-point priors to $\theta_0^{k(n)} = \frac{1}{|I^k|} \sum_{i \in I^k} \theta_i$, $\forall k \in K$. Set $n = n + 1$
4. For each $k \in K$ and $i \in I^k$, solve a QP to get θ_i:

$$\min_{\theta_i} (\theta_0^{k(n)} - \theta_i)^2$$
$$s.t.$$
$$\theta_i(X_{nj} - X_{nj^*}) \geq ln(S_{ij}) - ln(S_{ij^*}) - tol, \qquad \forall j, j^* \in J, j \neq j^*$$
$$\theta_i(X_{nj} - X_{nj^*}) \leq ln(S_{ij}) - ln(S_{ij^*}) + tol, \qquad \forall j, j^* \in J, j \neq j^*$$

5. For each $k \in K$ and $i \in I^k$, solve a QP to get θ_i:
6. Set average to $y^{k(n)} = \frac{1}{|I^k|} \sum_{i \in I^k} \theta_i$, $\forall k \in K$, update the fixed priors:

$$\theta_0^{k(n+1)} = \frac{n}{n+1} \theta_0^{k(n)} + \frac{1}{n+1} y^{k(n)}$$
$$\theta_0^{(n+1)} = \frac{1}{|I|} \sum_{k=1}^{|K|} |I^k| * \theta_0^{k(n+1)}$$

7. If MSA stopping criteria for $\theta_0^{(n+1)}$ reached, stop and output $\theta_i^{(n)}$, else, n = n+1 and go to Step 4

3.2.3. An illustrative example

We built a simple example with 8 agents to illustrate how the g-AMXL model and its estimation algorithm work. In this example, each agent refers to an OD pair of a population segment. Only two modes, taxi and transit, are considered for simplicity. Each row of the sample data contains the ID of the agent, travel time and cost of taxi, travel time and cost of transit, and mode share

of the two modes. The sample data containing 8 agents is shown in Table 3.5. The derived utilities of the two modes are defined in Eq. (3.13) – (3.14).

$$V_{taxi,i} = \theta_{time,i} taxi\_time_i + \theta_{cost,i} taxi\_cost_i \,, \qquad \forall i \in I \tag{3.13}$$

$$V_{transit,i} = \theta_{time,i} transit\_time_i + \theta_{cost,i} transit\_cost_i + \theta_{c\_transit,i} \,, \qquad \forall i \in I \tag{3.14}$$

where $V_{taxi,i}$ and $V_{transit,i}$ are utilities derived from choosing taxi and transit. $\theta_{time,i}$ and $\theta_{cost,i}$ are the coefficients of travel time and cost for agent $i \in I$. $\theta_{c\_transit,i}$ is the mode constant for agent i. It is noted that we added two "fake" agents (agent 7 and 8) into the dataset. The mode shares of these two agents are unreasonable since the mode with a longer travel time and a higher cost has a higher market share.

Table 3.5. The sample dataset for illustration

Agent ID	Taxi time	Taxi cost	Transit time	Transit cost	Taxi (%)	Transit (%)
1	10 min	$ 10	30 min	$ 3	80%	20%
2	20 min	$ 15	40 min	$ 3	70%	30%
3	40 min	$ 25	60 min	$ 3	60%	40%
4	10 min	$ 10	30 min	$ 3	20%	80%
5	20 min	$ 15	40 min	$ 3	30%	70%
6	40 min	$ 25	60 min	$ 3	40%	60%
7	10 min	$ 3	30 min	$ 10	10%	90%
8	60 min	$ 25	10 min	$ 3	90%	10%

We ran Algorithm 3.1 with the above settings and k=3. The estimated coefficients are shown in Table 3.6. The estimated market share E_Taxi (%) and E_Transit (%) are quite close to the input data. Moreover, the results reflect diverse tastes at the agent level though the three latent classes: (1) agent 1-3 have negative $\theta_{time,i}$ and $\theta_{cost,i}$ close to zero, indicating a preference for shorter travel time; (2) agent 4-6 have negative $\theta_{cost,i}$ and $\theta_{time,i}$ close to zero, indicating a preference for lower travel cost; (3) agent 7 and 8 have positive $\theta_{cost,i}$ and $\theta_{time,i}$, indicating an "irregular" preference for longer travel time and higher travel cost. In ubiquitous datasets, "irregular" preference is often related to issues in data collection. Therefore, g-AMXL can be used to check the data quality to some extent.

Table 3.6. Results obtained from Algorithm 3.1

Agent ID	$\theta_{time,i}$	$\theta_{cost,i}$	$\theta_{c\_transit,i}$	Cluster	E_Taxi (%)	E_Transit (%)
1	-0.107	$-7.30*10^{-8}$	-0.005	1	89.54%	10.46%
2	-0.067	$-1.71*10^{-8}$	-0.003	1	79.19%	20.81%
3	-0.040	$-1.68*10^{-9}$	-0.002	1	68.94%	31.06%
4	$-7.69*10^{-10}$	-0.301	0.043	2	10.46%	89.54%
5	$-3.90*10^{-10}$	-0.111	0.009	2	20.81%	79.19%
6	$-9.46*10^{-9}$	-0.036	0.001	2	31.06%	89.54%
7	0.095	0.034	0.005	3	10.46%	89.54%
8	0.036	0.016	$-7.19*10^{-4}$	3	89.54%	10.46%

3.3. Optimization models for service region design

We further integrated coefficients estimated by g-AMXL into optimization models for equitable service region optimization. We assume that there will be new mobility services selecting counties in NY state as operating zones, in which a few vehicles should provide bi-direction trip services to meet the demands on the selected OD links. Each vehicle has a maximum service distance and a maximum number of trips per day. To avoid confusion, we use the index uw to denote trips along block-group OD pairs. The total mode utilities in a block-group OD pair are the weighted sum (by trip volume conducted by the segment) of utilities of different population segments. Notations used in this section are shown in Table 3.7.

Table 3.7. Notations used in the optimization models

Parameters observed from Replica's data	
Z	The set of all counties in New York state
N	The set of all block groups in New York state (with 50 trips or higher per day)
N^{dis}	The set of disadvantaged communities (block groups) identified by NYSERDA
N^{other}	The set of other communities (block groups) except for disadvantaged ones
N_i	The set of block groups in county $i \in Z$ (with 50 trips or higher per day)
K^-	The mode choice set without the new mobility service, including six modes
K^+	The mode choice set with the new mobility service, including seven modes
X_{uw}^k	A vector of trip details (including average travel time, average monetary cost, and mode constant) from block group $u \in N$ to $w \in N$ using mode k$\in K^+$
d_{uw}	The travel demand (trips/day) on the link from block group $u \in N$ to $w \in N$, $u \neq w$

l_{uw}	Trip length (km) from block group $u \in N$ and $w \in N$, $u \neq w$, measured by the Manhattan distance between the centroids the block groups
M	A large positive value
Parameters defining the new mobility service	
O	The maximum number of operating zones
$\mathcal{F}_{max}$	The maximum fleet size in total (vehicles/day)
F_{max}, F_{min}	The maximum and minimum fleet size in each operating zone (vehicles/day)
$t_{uw}^{\hat{k}}$	Trip duration (minutes) of the new mobility service on the link from block group $u \in N$ to $w \in N$, $u \neq w$
$c_{uw}^{\hat{k}}$	Trip fee ($/trip) of the new mobility service on the link from block group $u \in N$ to $w \in N$, $u \neq w$
L	The maximum distance (km) a vehicle can serve per day
D	The maximum number of trips a vehicle can serve per day
Parameters calculated from the mode choice model	
β_{uw}	A vector of mode choice coefficients for trips from block group $u \in N$ to $w \in N$, $u \neq w$
V_{uw}^{k}	The utility of traveling from block group $u \in N$ to $w \in N$ using mode k$\in K^+$
$d_{uw}^{\hat{k}}$	The demand (trips/day) of the new mobility service on the link from block group $u \in N$ to $w \in N$, $u \neq w$
s_{uw}^{K}	Social welfare (or consumer surplus) of traveling from block group $u \in N$ to $w \in N$, $u \neq w$, given the mode choice set $K \in \{K^+, K^-\}$
Decision variables (single-service region design)	
y_i	A binary variable that indicates whether county $i \in Z$ is included into the service region
x_{uw}	A binary variable that indicates whether the link from block group $u \in N$ to $w \in N$ is operated
f_{uw}	The fleet size (vehicles/day) on the link from block group $u \in N$ to $w \in N$, $u \neq w$
Decision variables (multi-service region assortment)	
y_i	A binary variable that indicates whether county $i \in Z$ is included into the service region of service A and B
x_{uw}^{A}, x_{uw}^{B}	Binary variables that indicate whether the link from block group $u \in N$ to $w \in N$ is operated by service A and B, respectively
f_{uw}^{A}, f_{uw}^{A}	The fleet size (vehicles/day) of service A and B on the link from block group $u \in N$ to $w \in N$, $u \neq w$

3.3.1. Equity metrics

We constructed six equity metrics mentioned in Karsu and Morton (2015)'s study.

(1) Average welfare: the consumer surplus extracted from the coefficients estimated by g-AMXL model, which is defined in Eq. (3.15) – (3.16)

$$AW = \frac{1}{|N|^2}\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} s_{uw}^{K}, \qquad \forall u,w \in N, u \neq w \tag{3.15}$$

$$where\ s_{uw}^{K} = ln[\frac{1}{|K|}\sum_{k=1}^{|K|} exp(V_{uw}^{k})], \quad \forall u,w \in N, u \neq w, K \in \{K^+, K^-\} \tag{3.16}$$

(2) Welfare range: the range between the minimum and maximum levels of welfare, which is defined in Eq. (3.17)

$$R = max_{uw}\, s_{uw}^{K} - min_{uw}\, s_{uw}^{K}, \qquad \forall K \in \{K^+, K^-\} \tag{3.17}$$

(3) Welfare mean deviation: the range between the minimum and maximum levels of welfare, which is defined in Eq. (3.18)

$$MD = \frac{1}{|N|^2}\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} |s_{uw}^{K} - \overline{s_{uw}^{K}}|, \quad \overline{s_{uw}^{K}} = \frac{\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} s_{uw}^{K}}{|N|^2}, \qquad \forall K \in \{K^+, K^-\} \tag{3.18}$$

(4) Welfare variance: the variance of welfare defined in Eq. (3.19)

$$V = \frac{1}{|N|^2}\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} (s_{uw}^{K} - \overline{s_{uw}^{K}})^2, \quad \overline{s_{uw}^{K}} = \frac{\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} s_{uw}^{K}}{|N|^2}, \qquad \forall K \in \{K^+, K^-\} \tag{3.19}$$

(5) Welfare Gini coefficient: Gini coefficient defined in Eq. (3.20), where L is the set of all block group OD links.

$$Gini = \frac{\sum_{i=1}^{|L|}\sum_{j=1}^{|L|} |s_i^K - s_j^K|}{2|L|\sum_{j=1}^{|L|} s_j^K}, \qquad \forall K \in \{K^+, K^-\} \tag{3.20}$$

(6) Welfare disparity (differences of welfare between disadvantaged and non-disadvantaged communities), which is defined in Eq. (3.21) where L^{dis} and L^{other} are OD links related to disadvantaged and non-disadvantaged communities.

$$DW = \frac{1}{|L^{dis}|}\sum_{i=1}^{|L^{dis}|} s_i^K - \frac{1}{|L^{other}|}\sum_{j=1}^{|L^{other}|} s_j^K, \qquad \forall K \in \{K^+, K^-\} \tag{3.21}$$

3.3.2. Single-service region design problem

For the single-service region design problem, we assume there is only one new mobility service entering the market. We considered three objectives of the new mobility service: (1) maximizing the total revenue of the new service; (2) maximizing the total welfare (change of consumer surplus); (3) minimizing the welfare disparities between disadvantaged and non-disadvantaged communities. Accordingly, three objective functions were formulated (Eq. (3.22) - (3.24)).

$$\max_{y_i, x_{uw}, f_{uw}} \sum_{u=1}^{|N|} \sum_{w=1}^{|N|} c_{uw}^{\hat{k}} d_{uw}^{\hat{k}} x_{uw} \tag{3.22}$$

$$\max_{y_i, x_{uw}, f_{uw}} \sum_{u=1}^{|N|} \sum_{w=1}^{|N|} \left(s_{uw}^{K^+} - s_{uw}^{K^-}\right) d_{uw} x_{uw} \tag{3.23}$$

$$\min_{y_i, x_{uw}, f_{uw}} \sum_{u \in N^{other}} \sum_{w \in N} \left(s_{uw}^{K^+} - s_{uw}^{K^-}\right) d_{uw} x_{uw} - \sum_{u \in N^{dis}} \sum_{w \in N} \left(s_{uw}^{K^+} - s_{uw}^{K^-}\right) d_{uw} x_{uw} \tag{3.24}$$

where y_i, x_{uw}, f_{uw} are decision variables indicating whether county i is included into the service region, whether link uw is operated, and the fleet size on link uw. d_{uw} and $d_{uw}^{\hat{k}}$ denote the total demand and demand for the new mobility service on link uw. $c_{uw}^{\hat{k}}$ denotes the trip fare of the new mobility service on link uw. N^{dis} denotes the set of block groups that are identified as disadvantaged communities by NYSERDA (2021). N^{other} denotes the set of block groups that are identified as disadvantaged communities. $s_{uw}^{K^+}$ and $s_{uw}^{K^-}$ denote the social welfare (or consumer surplus) with and without the new mobility service, as defined in Eq. (3.25), in which V_{uw}^{k} denotes the utility of traveling from block group $u \in N$ to $w \in N$ using mode k.

$$sw_{uw,K} = \ln\left[\sum_{k=1}^{|K|} \exp\left(V_{uw}^{k}\right)\right], \qquad \forall u, w \in N, u \neq w, K \in \{K^+, K^-\} \tag{3.25}$$

Eq. (3.26) - (3.36) are constraints of this linear programming (LP) problem. Eq. (3.26) - (3.27) are to ensure the number of service zones and vehicle fleet size are restricted by the total budget. Eq. (3.28) - (3.29) is to ensure only links within the service zones can be operated and vehicles can only be assigned to operating links. Eq. (3.30) - (3.31) are to ensure a maximum and minimum

fleet size in each service zone. Eq. (3.32) is to ensure a bi-direction trip of the new mobility service. Eq. (3.33) - (3.34) are to ensure that the travel demand on an operating link should be met within the maximum distance and number of trips per vehicle. Eq. (3.35) - (3.36) are to define the type of decision variables.

$$\sum_{i=1}^{|Z|} y_i \leq O \tag{3.26}$$

$$\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} f_{uw} \leq \mathcal{F}_{max}, \qquad u \neq w \tag{3.27}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} x_{uw} \leq My_i, \qquad \forall i \in Z, u \neq w \tag{3.28}$$

$$f_{uw} \leq Mx_{uw}, \qquad \forall u, w \in N, u \neq w \tag{3.29}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} f_{uw} \leq F_{max}y_i, \qquad \forall i \in Z, u \neq w \tag{3.30}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} f_{uw} \geq F_{min}y_i, \qquad \forall i \in Z, u \neq w \tag{3.31}$$

$$f_{uw} = f_{wu}, \qquad \forall u, w \in N, u \neq w \tag{3.32}$$

$$Lf_{uw} \geq d_{uw}^{\hat{k}} l_{uw} x_{uw}, \qquad \forall u, w \in N, u \neq w \tag{3.33}$$

$$Df_{uw} \geq d_{uw}^{\hat{k}} x_{uw}, \qquad \forall u, w \in N, u \neq w \tag{3.34}$$

$$y_i, x_{uw} \in \{0,1\} \tag{3.35}$$

$$f_{uw} \in Z^+ \tag{3.36}$$

3.3.3. Multi-service region assortment problem

For the multi-service region assortment problem, we assume there are two new mobility services (service A and service B) with different pricing policies and performances. The major difference is that now the demand of the new mobility services on link uw depends on whether service A

and B choose to operate on the link. Many studies mentioned that the assortment and pricing optimization under logit models is a challenging problem (Wang, 2012; Agrawal et al., 2019). However, we can avoid the complexity by enumerating all the possible demand configurations if we assume there are only two mobility services entering the market (resulting in four configurations to enumerate).

Eq. (3.37) defines the objective function of maximizing the total revenue of the two services, where the demand of service A on link uw is defined as a combination of possible demands and the decision variables. Eq. (3.38) ensures that if $x_{uw}^{A} = 0$, then $d_{uw}^{A} = 0$, and if $x_{uw}^{A} = 0$, $x_{uw}^{B} = 1$, then $d_{uw}^{A} = d_{uw}^{A\_AB}$. Eq. (3.39) – (3.40) defines the demand of service A when only service A operates on link uw ($d_{uw}^{A\_A}$) and when both service A and B operate on link uw ($d_{uw}^{A\_AB}$), respectively.

$$\max_{y_i,\dots,f_{uw}^{B}} \sum_{u=1}^{|N|}\sum_{w=1}^{|N|} c_{uw}^{A} d_{uw}^{A} + \sum_{u=1}^{|N|}\sum_{w=1}^{|N|} c_{uw}^{B} d_{uw}^{B} \tag{3.37}$$

$$d_{uw}^{A} = d_{uw}^{A\_A} x_{uw}^{A}(1 - x_{uw}^{B}) + d_{uw}^{A\_AB} x_{uw}^{A} x_{uw}^{B} \tag{3.38}$$

$$d_{uw}^{A\_A} = \frac{\exp(V_{uw}^{A})}{\sum_{k=1}^{|K^-|}[\exp(V_{uw}^{k})] + \exp(V_{uw}^{A})} \tag{3.39}$$

$$d_{uw}^{A\_AB} = \frac{\exp(V_{uw}^{A})}{\sum_{k=1}^{|K^-|}[\exp(V_{uw}^{k})] + \exp(V_{uw}^{A}) + \exp(V_{uw}^{B})} \tag{3.40}$$

Eq. (3.41) defines the objective function of maximizing the total change of social welfare (consumer surplus), where the change of social welfare on link uw is also defined as a combination of all the possible change of social welfare and the decision variables. Eq. (3.42) ensures that (1) if both x_{uw}^{A} and x_{uw}^{B} equals to zero, the change of welfare will equal to zero; (2) if $x_{uw}^{A} = 1$ and $x_{uw}^{B} = 0$, $\widehat{cs}_{uw}$ equals to the change of welfare after with only service A (cs_{uw}^{A}); (3) if $x_{uw}^{A} = 0$ and $x_{uw}^{B} = 1$, $\widehat{cs}_{uw}$ equals to the change of welfare after with only service B (cs_{uw}^{B}); (4) if both x_{uw}^{A} and x_{uw}^{B} equals to 1, $\widehat{cs}_{uw}$ equals to the change of welfare after with both service A and B (cs_{uw}^{AB}). Eq. (3.43) – (3.45) defines the three possible changes of welfare.

$$\max_{y_i,\dots,f_{uw}^{B}} \sum_{u=1}^{|N|}\sum_{w=1}^{|N|} \widehat{cs}_{uw} \tag{3.41}$$

$$\widehat{cs}_{uw} = cs_{uw}^{AB} x_{uw}^{A} x_{uw}^{B} + cs_{uw}^{A} x_{uw}^{A}(1 - x_{uw}^{B}) + cs_{uw}^{B}(1 - x_{uw}^{A}) x_{uw}^{B} \tag{3.42}$$

$$cs_{uw}^{AB} = \ln\{\sum_{k=1}^{|K^-|}[\exp(V_{uw}^{k})] + \exp(V_{uw}^{A}) + \exp(V_{uw}^{B})\} - \ln[\sum_{k=1}^{|K^-|}\exp(V_{uw}^{k})] \tag{3.43}$$

$$cs_{uw}^{A} = \ln\{\sum_{k=1}^{|K^-|}[\exp(V_{uw}^{k})] + \exp(V_{uw}^{A})\} - \ln[\sum_{k=1}^{|K^-|}\exp(V_{uw}^{k})] \tag{3.44}$$

$$cs_{uw}^{B} = \ln\{\sum_{k=1}^{|K^-|}[\exp(V_{uw}^{k})] + \exp(V_{uw}^{B})\} - \ln[\sum_{k=1}^{|K^-|}\exp(V_{uw}^{k})] \tag{3.45}$$

Eq. (3.46) defines the objective function of minimizing the welfare disparities between disadvantaged and non-disadvantaged communities, where N^{dis} and N^{other} denote the set of block groups that are identified as disadvantaged and non-disadvantaged communities by NYSERDA (2021). The change of welfare on each link has already been defined in Eq. (3.42)

$$\min_{y_i,\ldots,f_{uw}^{B}} \sum_{u\in N^{other}}\sum_{w\in N}\widehat{cs}_{uw} - \sum_{u\in N^{dis}}\sum_{w\in N}\widehat{cs}_{uw} \tag{3.46}$$

Eq. (3.47) – (3.57) are constraints shared by the above three objective functions, which are similar to the constraints in the single-service region design problem. The only difference is that an index k is added to denote the service A and B. Since we have decision variables multiplied together like $x_{uw}^{A} x_{uw}^{B}$, this is a quadratic programming (QP) problem.

$$\sum_{i=1}^{|Z|} y_i \le O \tag{3.47}$$

$$\sum_{u=1}^{|\mathrm{N}|}\sum_{w=1}^{|\mathrm{N}|} f_{uw}^{k} \le \mathcal{F}_{max}, \qquad \forall k \in [A,B], u \ne w \tag{3.48}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} x_{uw}^{k} \le M y_i, \qquad \forall i \in Z, k \in [A,B], u \ne w \tag{3.49}$$

$$f_{uw}^{k} \le M x_{uw}^{k}, \qquad \forall u,w \in N, k \in [A,B], u \ne w \tag{3.50}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} f_{uw}^{k} \leq F_{max} y_i, \qquad \forall i \in Z, k \in [A,B], u \neq w \tag{3.51}$$

$$\sum_{u=1}^{|N_i|}\sum_{w=1}^{|N_i|} f_{uw}^{k} \geq F_{min} y_i, \qquad \forall i \in Z, k \in [A,B], u \neq w \tag{3.52}$$

$$f_{uw}^{k} = f_{wu}^{k}, \qquad \forall u,w \in N, k \in [A,B], u \neq w \tag{3.53}$$

$$L f_{uw}^{k} \geq l_{uw} d_{uw}^{k}, \qquad \forall u,w \in N, k \in [A,B], u \neq w \tag{3.54}$$

$$D f_{uw}^{k} \geq d_{uw}^{k}, \qquad \forall u,w \in N, k \in [A,B], u \neq w \tag{3.55}$$

$$y_i, x_{uw}^{A}, x_{uw}^{B} \in \{0,1\} \tag{3.56}$$

$$f_{uw}^{A}, f_{uw}^{B} \in Z^{+} \tag{3.57}$$

4. Data Validation and Descriptive Analysis

Datasets provided by Replica Inc. include the synthetic population and their trips on a typical Thursday and Saturday in the Fall 2019 season (September 2019 – November 2019). This section is to check the data quality and present the results of descriptive analysis. On the national scale, we generated a data quality report, to validate the information on population size, sociodemographic information, and trip volume. Descriptive analysis of trip details was conducted on the statewide and city scale. We used the data to identify urban versus rural areas and disadvantaged versus non-disadvantaged communities. Also, we picked four cities (New York City, Buffalo, Rochester, and Syracuse) to visualize their trip distribution, trip purpose, and mode share.

4.1. National data quality report

The national data quality report generated through Replica's pipeline used aggregated census data[6], Longitudinal Employer-Household Dynamics (LEHD) data[7], and Census Transportation Planning Products (CTPP) data[8] to check the accuracy of synthetic demographic characteristics and trip details. The report includes quality check of resident population by county and census tract, household size, household income, age, sex, employment, vehicles, school enrollment, LEHD jobs, origin-destination flows, and additional metrics (e.g., commute mode, race, ethnicity, etc.). The full report can be accessed through the following link:

https://xr2006.github.io/sample/Replica_project/PopulationQuality.html

Table 4.1 lists the sociodemographic and trip characteristics used in our project, in which the Error column is the percentage difference between Replica's data and ground truth data, and the Accuracy level column sets up three levels (99%, 95%, and 90%) to denote the data quality. The population size is 99% accurate by county and by census tract. Most of the other sociodemographic characteristics are at least 95% accurate, except for the number of households with six or more persons. As for trip details, though the commute mode-share by tract is 99%

[6] https://www.census.gov/data.html
[7] https://lehd.ces.census.gov/
[8] https://ctpp.transportation.org/

accurate, the origin-destination flows at the tract level cannot fit perfectly to CTPP data, with a percentage difference of 25.78% (see Figure 4.1). Reasons to explain this difference can be: (1) It is quite hard to calibrate a nationwide model to generate tract-to-tract flows exactly the same as CTPP data; (2) Replica's methodology does not completely rely on CTPP data, which has a sample rate instead of covering the total population. To this end, it is necessary to conduct statewide and city-level analysis to validate Replica's data on a finer scale.

Table 4.1. Field definitions of the Population table (used in our project)

Category	Subcategory	Bucket	Error	Accuracy level
Sociodemographic Characteristics	Population size	By Tract	0.20%	99%
		By County	0.15%	99%
	Household size (by tract)	1_person	0.51%	99%
		2_person	0.24%	99%
		3_person	0.54%	99%
		4_person	1.03%	95%
		5_person	2.75%	95%
		6_person	7.56%	90%
		7_plus_person	15.70%	--
	Household income group (by tract)	< 10,000	1.09%	95%
		10,000-40,000	0.59%	99%
		40,000-75,000	0.63%	99%
		75,000-125,000	0.74%	99%
		> 125,000	0.85%	99%
	Age group (by tract)	< 4	0.40%	99%
		5-14	0.25%	99%
		15-17	0.53%	99%
		18-24	0.32%	99%
		25-34	0.25%	99%
		35-64	0.12%	99%
		> 65	0.24%	99%
Trip Details	Commute mode share (by tract)	Private auto	0.12%	99%
		Public transit	0.46%	99%
		Biking	1.34%	95%

		Walking	0.86%	99%
		Carpool	0.55%	99%
		Not_working	0.01%	99%
	Origin-destination flows (county to county)	Total count	0.24%	99%
		By industry	0.31%	99%
		By mode	0.40%	99%
	Origin-destination flows (tract to tract)	Total count	25.78%	--
		By mode	33.05%	--

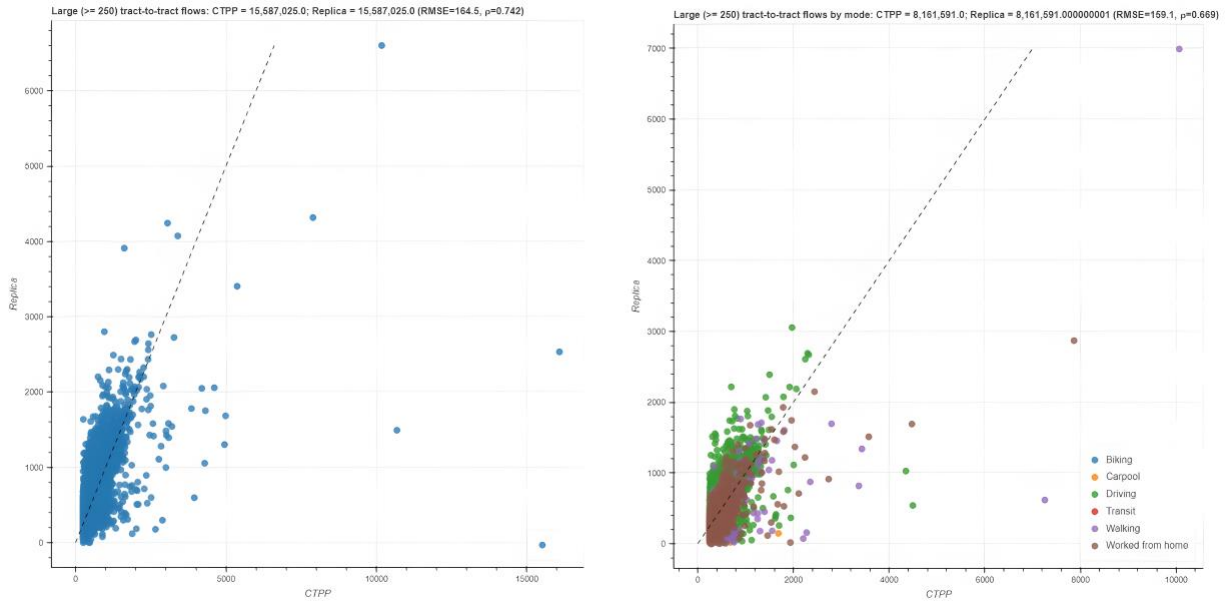


Figure 4.1. Scatterplots of tract-to-tract flows (x-axis denotes CTPP data, y-axis denotes Replica data)

4.2. Statewide descriptive analysis

4.2.1. Trip details on weekdays and weekends

Figure 4.2 presents the size, number of trips, trip rate, and mode share of the four population segments on weekdays and weekends (PrivateAuto_pro denotes the market share of private auto). There are 19,568,859 residents in New York State, of which 6.61% are LowIncome

Population, 52.22% are NotLowIncome Population, 17.09% are Senior Population, and 24.06% are Student Population. The trip rate (trips/day) of Senior Population is the highest (3.773 trips per weekday, 3.646 trips per weekend), followed by NotLowIncome Population (3.605 trips per weekday, 3.613 trips per weekend), LowIncome Population (3.096 trips per weekday, 3.047 trips per weekend), and Student Population (2.374 trips per weekday, 1.338 trips per weekend). Private auto is the dominant trip mode in New York State, followed by walking, carpool, and public transit. The mode share presents slight differences on weekdays and on weekends: private auto market share is generally higher on weekends than on weekdays while public transit market share is generally higher on weekdays than on weekends.

New York State (NY) Weekday Trip Details									
Pop_group	Population_num	Trips_num	Trip_rate	PrivateAuto_pro	Transit_pro	On_demand_pro	Biking_pro	Walking_pro	Carpool_pro
LowIncome	1,293,820	4,005,053	3.096	42.14%	20.23%	0.39%	1.13%	23.91%	12.20%
NotLowIncome	10,220,296	36,839,197	3.605	52.69%	14.75%	2.49%	0.32%	15.96%	13.79%
Senior	3,344,539	12,618,691	3.773	59.65%	10.02%	2.24%	0.28%	12.18%	15.64%
Student	4,710,204	11,183,178	2.374	31.52%	9.84%	1.00%	1.38%	29.03%	27.23%

New York State (NY) Weekend Trip Details									
Pop_group	Population_num	Trips_num	Trip_rate	PrivateAuto_pro	Transit_pro	On_demand_pro	Biking_pro	Walking_pro	Carpool_pro
LowIncome	1,293,820	3,942,350	3.047	46.36%	10.64%	0.51%	2.12%	25.67%	14.70%
NotLowIncome	10,220,296	36,927,584	3.613	57.23%	6.17%	3.07%	0.88%	17.02%	15.64%
Senior	3,344,539	12,193,235	3.646	62.68%	4.27%	2.39%	0.49%	13.13%	17.04%
Student	4,710,204	6,303,389	1.338	55.49%	6.23%	2.00%	1.23%	20.68%	14.36%

Figure 4.2. Trip details in New York State

The block-group level trip origin-destination flows (with a trip volume larger than 50 trips/day) of the four population segments are visualized in Figure 4.3, from which we can find that the trip length of Senior and Student Population is relatively shorter while the trip length of LowIncome Population is relatively longer. Figure 4.4 – 4.9 shows the market share of private auto, public transit, on-demand auto, biking, walking, and carpool on weekdays and weekends. In general, private auto is the major trip mode in most of the areas in New York State. Public transit, biking, and walking are mainly chosen within urban areas (such as New York City, Buffalo, Albany, etc.), in which the population density is obviously higher while the trip length is shorter. It is noted that the market shares of on-demand auto and carpool are also higher within urban areas, though auto trips cover the whole state. This indicates the potential of a new auto service operating outside of urban areas.

Figure 4.3. Trip OD flows by four population segments

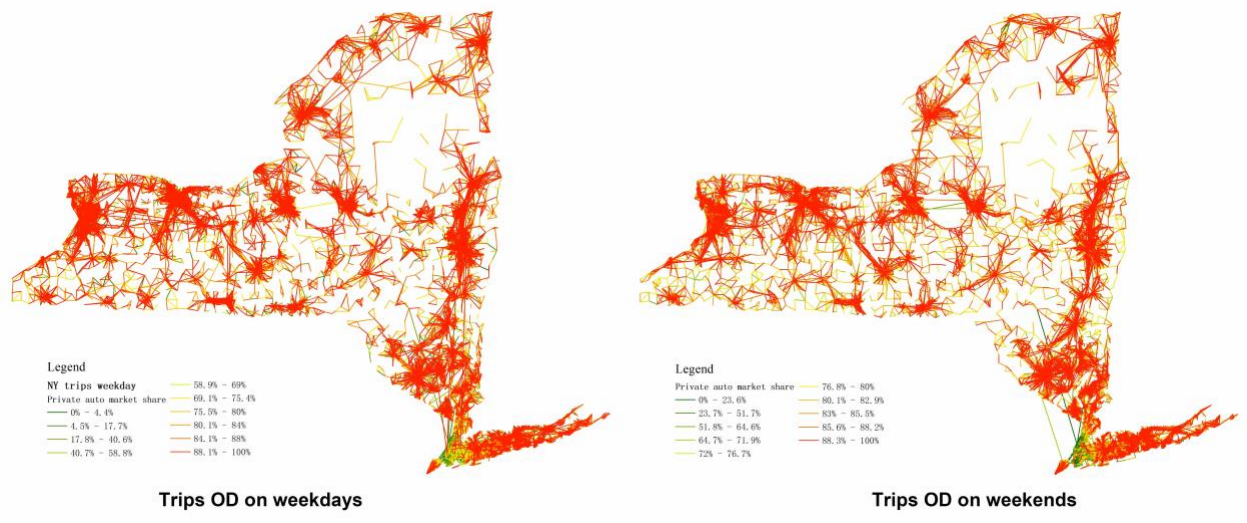


Figure 4.4. Market share of private auto on weekdays and weekends

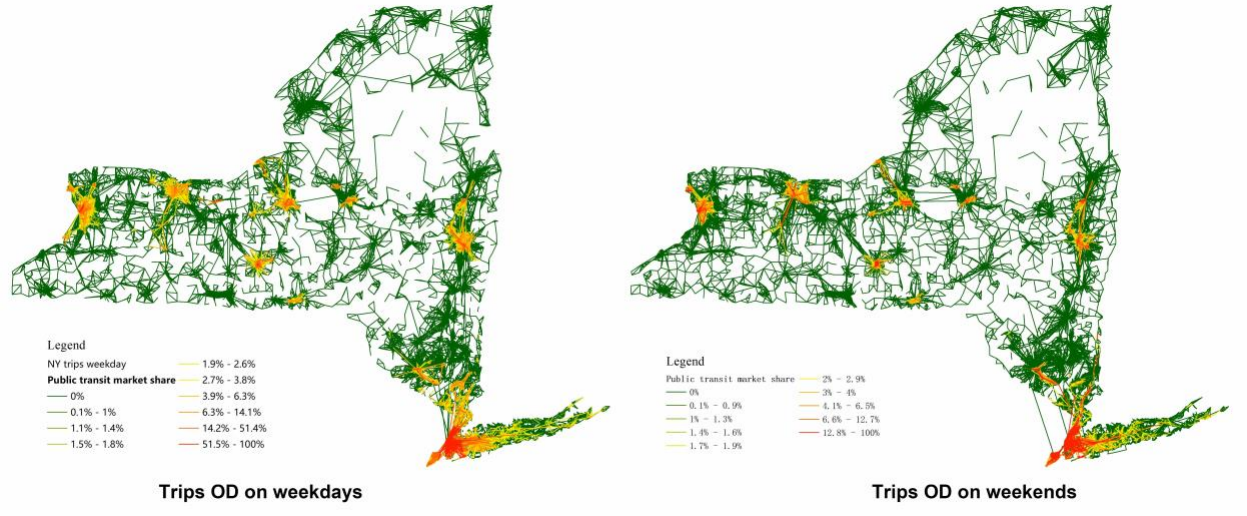


Figure 4.5. Market share of public transit on weekdays and weekends

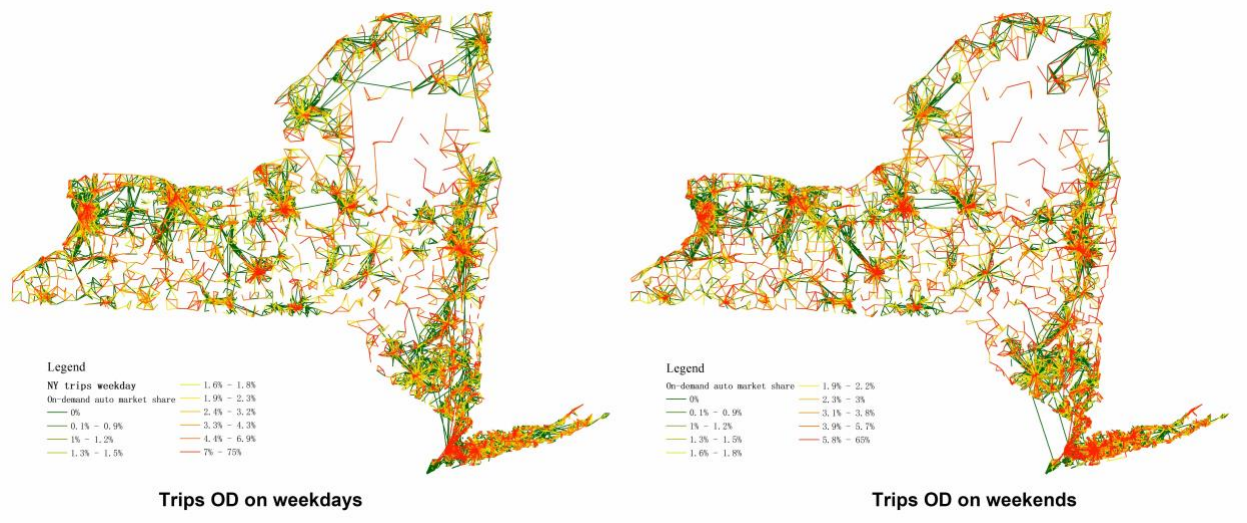


Figure 4.6. Market share of on-demand auto on weekdays and weekends

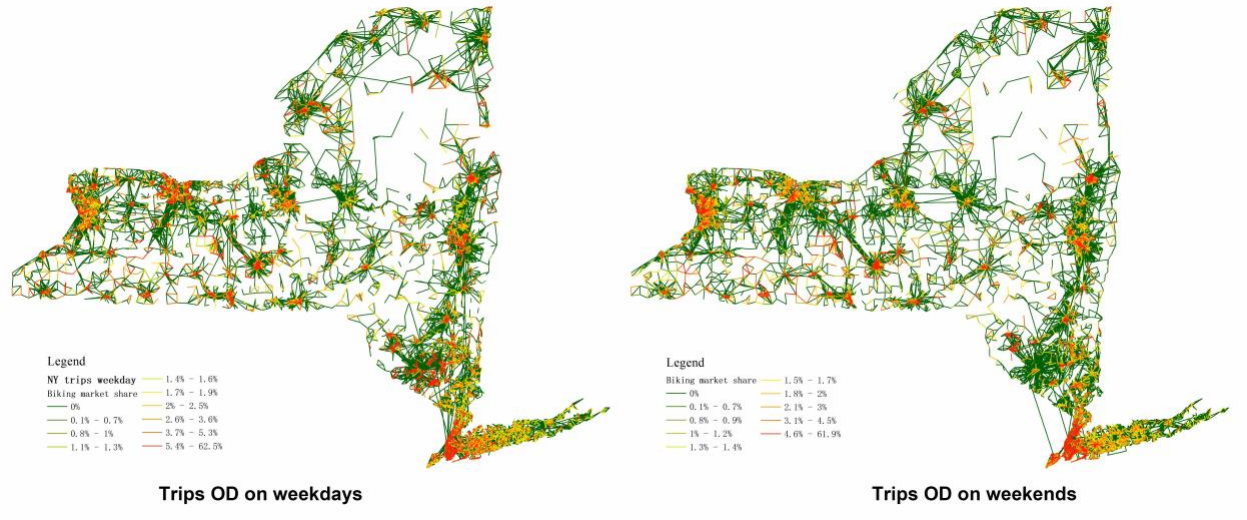


Figure 4.7. Market share of biking on weekdays and weekends

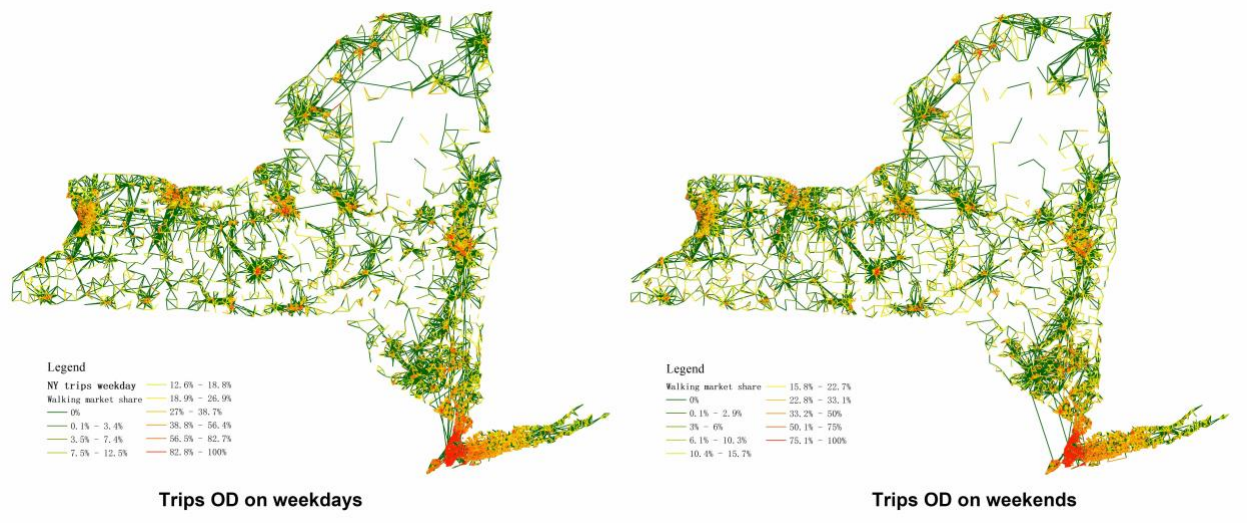


Figure 4.8. Market share of walking on weekdays and weekends

Figure 4.9. Market share of carpool on weekdays and weekends

4.2.2. Urban areas versus rural areas

Urban areas are defined primarily based on housing unit density measured at the census block level (United States Census Bureau, 2020). Three housing unit densities are applied during the delineation process:

(1) Initial urban core: at least 425 housing units per square mile. Based on the national average of 2.5 persons per housing unit, this density threshold is similar to the 1,000 persons per square mile used in 2000 and 2010 when delineating initial urban cores.

(2) High density nucleus: at least one high density nucleus of at least 1,275 housing units per square mile required for qualification. This ensures that each urban area contains a high-density nucleus typical of what one would expect to find within an urban area.

(3) Remainder of urban area: at least 200 housing units per square mile. This is similar to the 500 persons per square mile density used in 2000 and 2010, based on the national average of 2.5 persons per housing unit.

With the same criteria, we identified three kinds of urban areas and labeled the rest areas as rural (see Figure 4.10). The results align with our empirical knowledge.

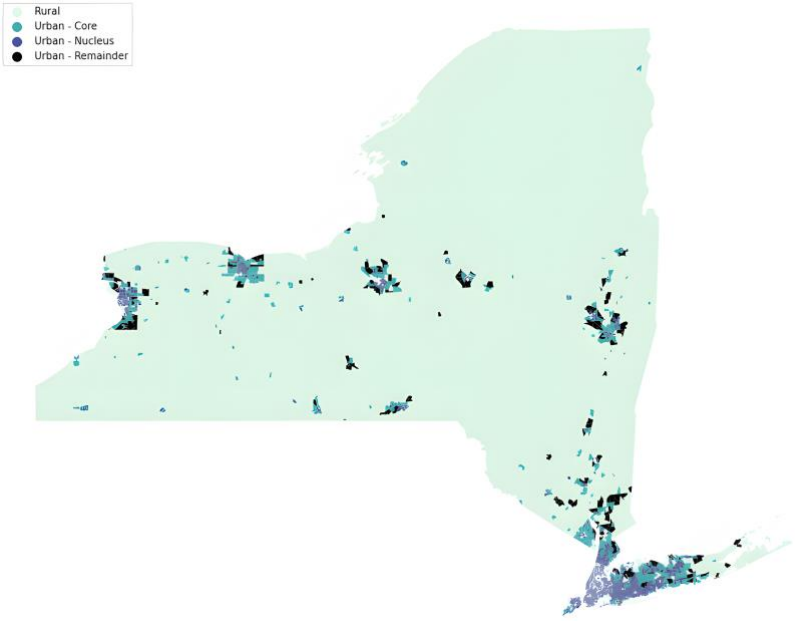


Figure 4.10. Urban and rural areas identified using Replica's data

Figure 4.11 compares the trip counts by mode in urban and rural areas, from which we can find there are few public transit, biking, or on-demand auto trips in rural areas. Though the trip counts in rural areas is much smaller than in urban areas, the unbalanced mode share might cause the inconvenience of traveling and increase transportation inequities. For rural households without vehicles or short of vehicles, it is necessary to provide a new auto service that is on-demand and requires a lower fare.

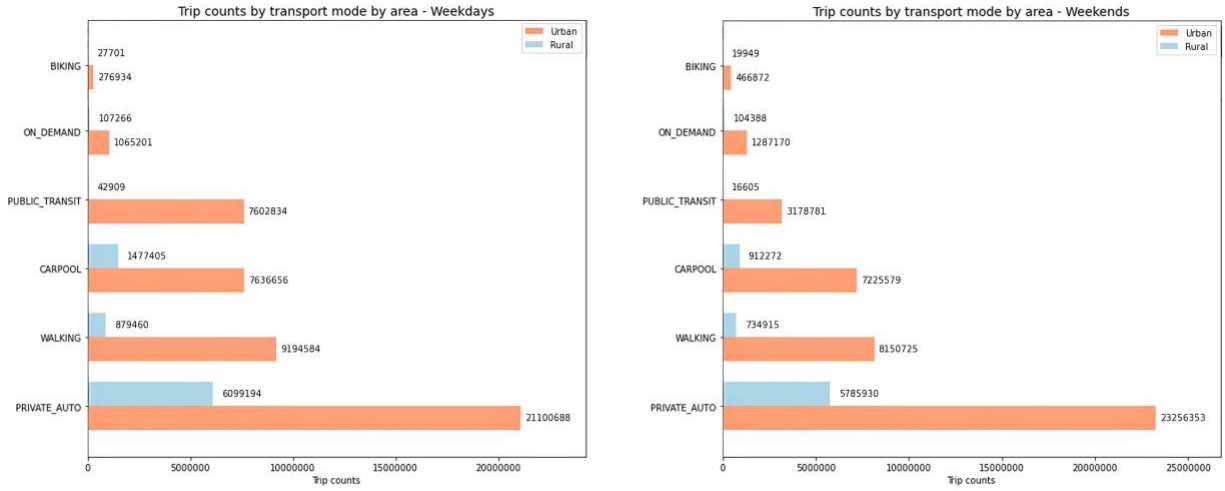


Figure 4.11. Trip counts by mode in urban and rural areas

4.2.3. Disadvantaged communities versus non-disadvantaged communities

New York States (NYSERDA, 2021) has identified interim criteria for disadvantaged communities, which includes communities:

(1) Located within census block groups that meet the HUD 50% AMI threshold, which is the top quartile of census block groups in New York State, ranked by the percentage of LMI Household in each census block. LMI Households are defined as households with annual income at or below 50% of the Area Median Income of the County where the Census Block Group resides.

(2) Located within the DEC Potential Environmental Justice Areas[9].

(3) Located within New York State Opportunity Zones[10].

Figure 4.12 shows the identified disadvantaged communities (census block groups) published by NYSERDA. Since we cannot find the shapefile of the DEC Potential Environmental Justice Areas and New York State Opportunity Zones, we only used annual household income to identify disadvantaged communities (see Figure 4.13). 71.4% of the disadvantaged communities published by NYSERDA can be identified using Replica's data. Considering we haven't included

[9] https://www.dec.ny.gov/public/911.html
[10] https://esd.ny.gov/opportunity-zones

the last two criteria, a consistency of 70% is acceptable and can serve as a validation of Replica's data.

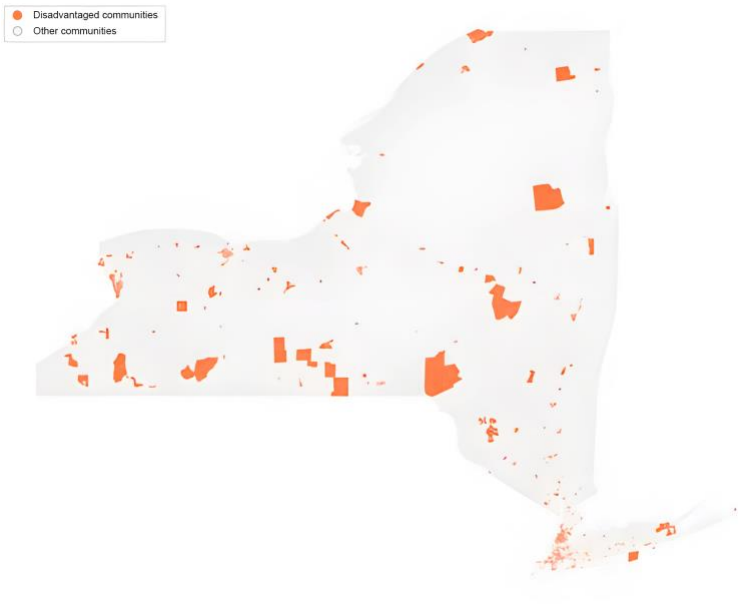


Figure 4.12. Disadvantaged and other communities published by NYSERDA

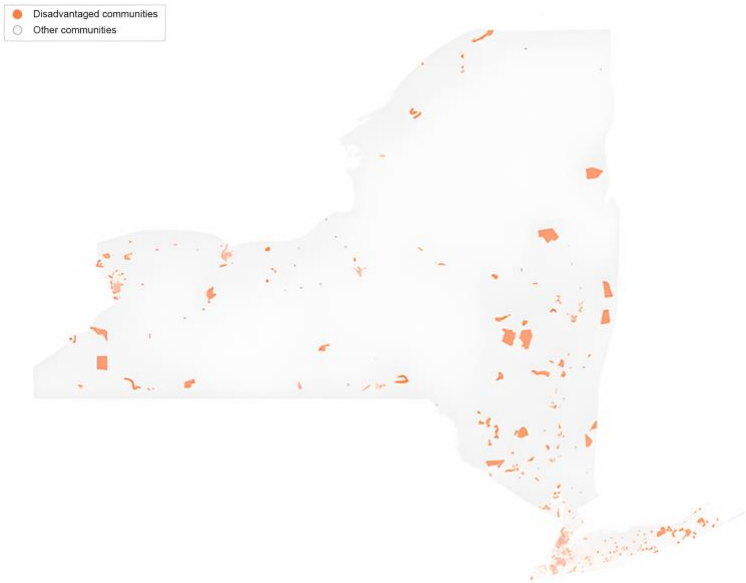


Figure 4.13. Disadvantaged and other communities identified using Replica's data

4.3. City-level descriptive analysis

Trip details of New York City, Buffalo, Rochester, and Syracuse were visualized in this section. These four cities were selected based on their population size[11].

4.3.1. New York City

Figure 4.14 presents the size, number of trips, trip rate, and mode share of the four population segments in New York City. There are 8,417,825 residents in New York State, of which 8.29% are LowIncome Population, 52.55% are NotLowIncome Population, 15.47% are Senior Population, and 23.69% are Student Population. The trip rate (trips/day) of Senior Population is the highest (3.259 trips per weekday, 3.313 trips per weekend), followed by NotLowIncome Population (3.091 trips per weekday, 3.221 trips per weekend), LowIncome Population (2.689 trips per weekday, 2.714 trips per weekend), and Student Population (2.083 trips per weekday, 1.214 trips per weekend). Private auto, public transit, and walking are major trip modes in New York City. The public transit market share is generally higher on weekdays than on weekends. Figure 4.15-4.16 show details of the trip density, mode share, and activity purpose composition in New York City.

New York City (NYC) Weekday Trip Details									
Pop_group	Population_num	Trips_num	Trip_rate	PrivateAuto_pro	Transit_pro	On_demand_pro	Biking_pro	Walking_pro	Carpool_pro
LowIncome	697,481	1,875,270	2.689	17.20%	40.89%	0.45%	1.29%	32.39%	7.79%
NotLowIncome	4,423,822	13,674,758	3.091	22.18%	37.75%	3.89%	0.53%	24.35%	11.30%
Senior	1,301,997	4,242,741	3.259	31.92%	28.19%	3.74%	0.44%	22.92%	12.79%
Student	1,994,525	4,155,254	2.083	13.63%	25.02%	1.41%	2.02%	47.12%	10.80%

New York City (NYC) Weekend Trip Details									
Pop_group	Population_num	Trips_num	Trip_rate	PrivateAuto_pro	Transit_pro	On_demand_pro	Biking_pro	Walking_pro	Carpool_pro
LowIncome	697,481	1,892,953	2.714	25.88%	21.23%	0.65%	3.02%	36.02%	13.19%
NotLowIncome	4,423,822	14,248,858	3.221	32.69%	15.47%	5.33%	1.79%	27.74%	16.98%
Senior	1,301,997	4,313,540	3.313	39.99%	11.53%	4.18%	0.96%	25.17%	18.17%
Student	1,994,525	2,420,551	1.214	32.35%	15.62%	3.21%	2.49%	31.33%	14.99%

Figure 4.14. Trip details in New York City

[11] https://www.moving.com/tips/the-10-largest-cities-in-new-york/

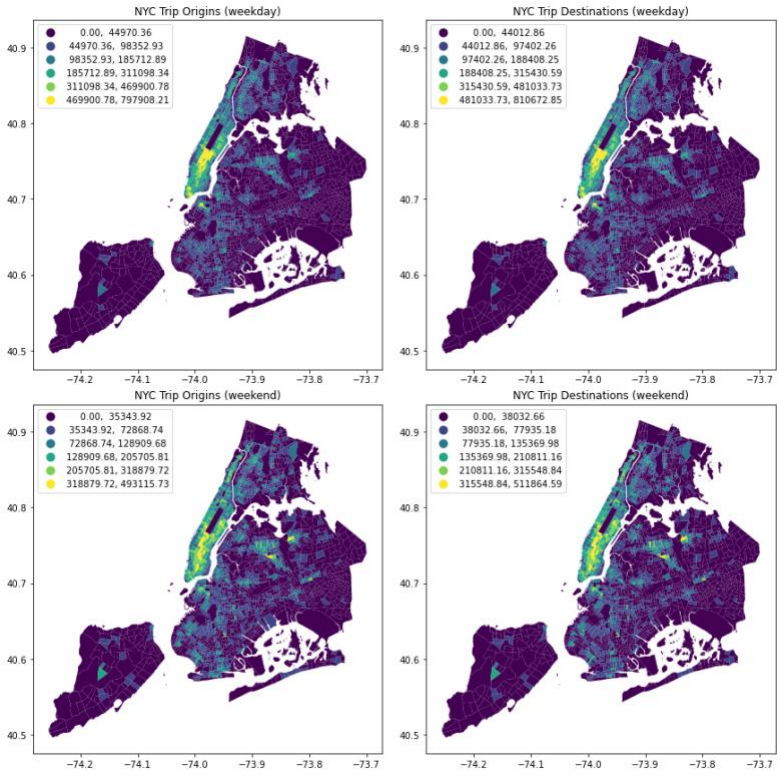


Figure 4.15. Trip density in New York City

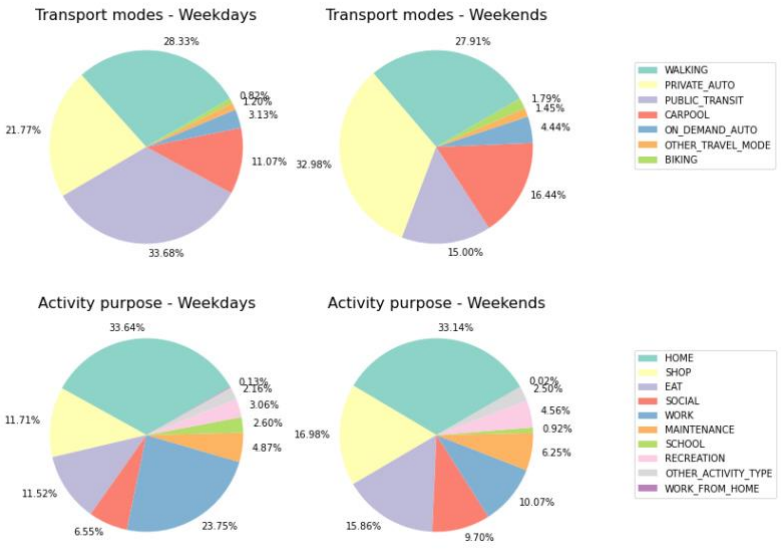


Figure 4.16. Mode share and activity purpose composition in New York City

4.3.2. Syracuse

Figure 4.17-4.18 show details of the trip density, mode share, and activity purpose composition in the county where Syracuse resides.

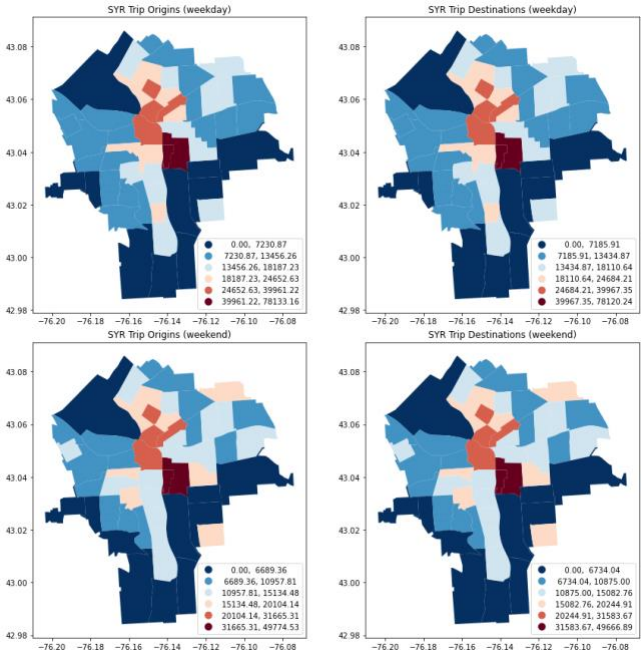


Figure 4.17. Trip density in Syracuse

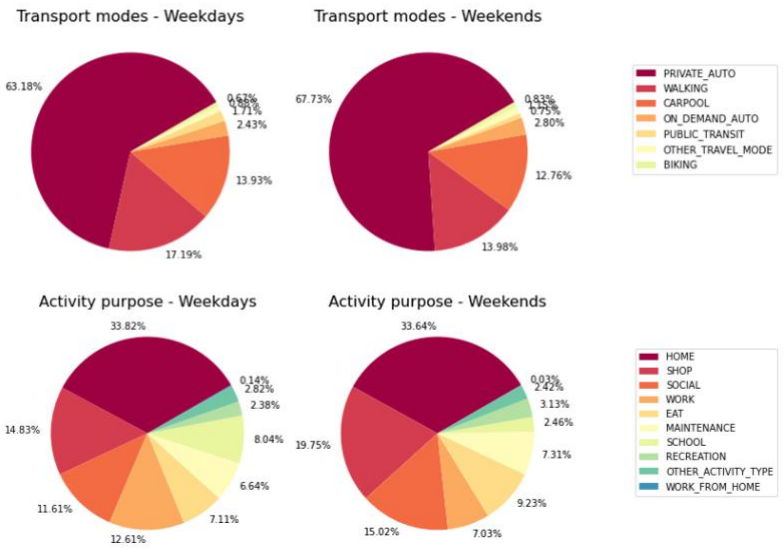


Figure 4.18. Mode share and activity purpose composition in Syracuse

4.3.3. Buffalo

Figure 4.19-4.20 show details of the trip density, mode share, and activity purpose composition in the county where Buffalo resides.

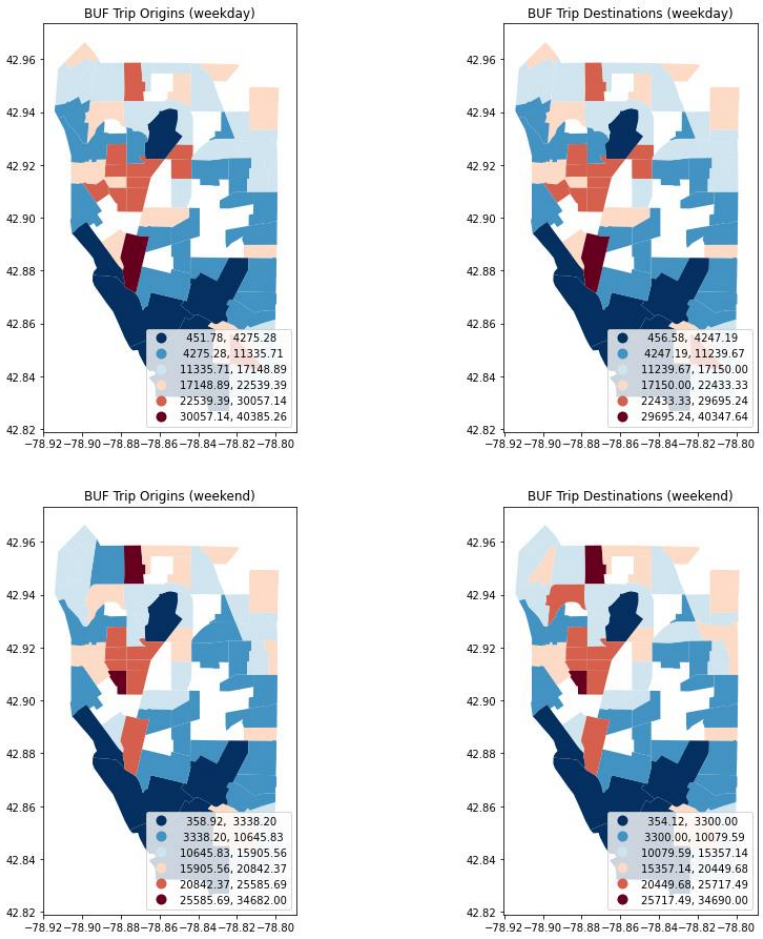


Figure 4.19. Trip density in Buffalo

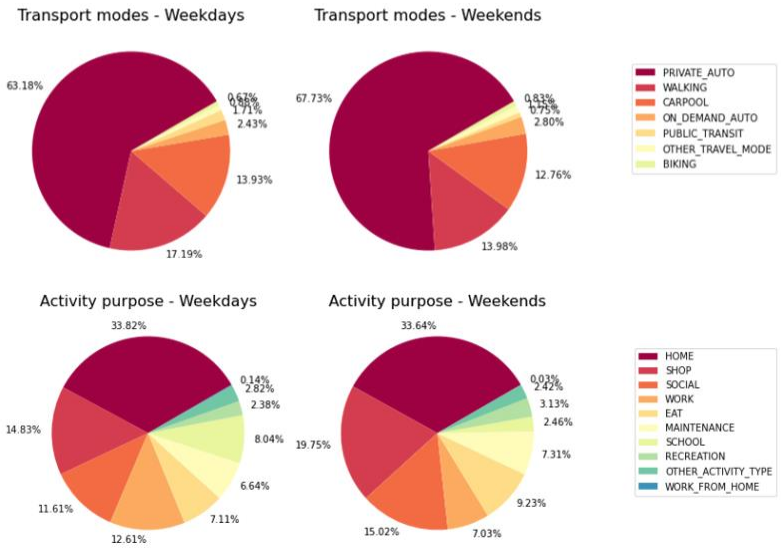


Figure 4.20. Mode share and activity purpose composition in Buffalo

4.3.4. Rochester

Figure 4.21-4.22 show details of the trip density, mode share, and activity purpose composition in the county where Rochester resides.

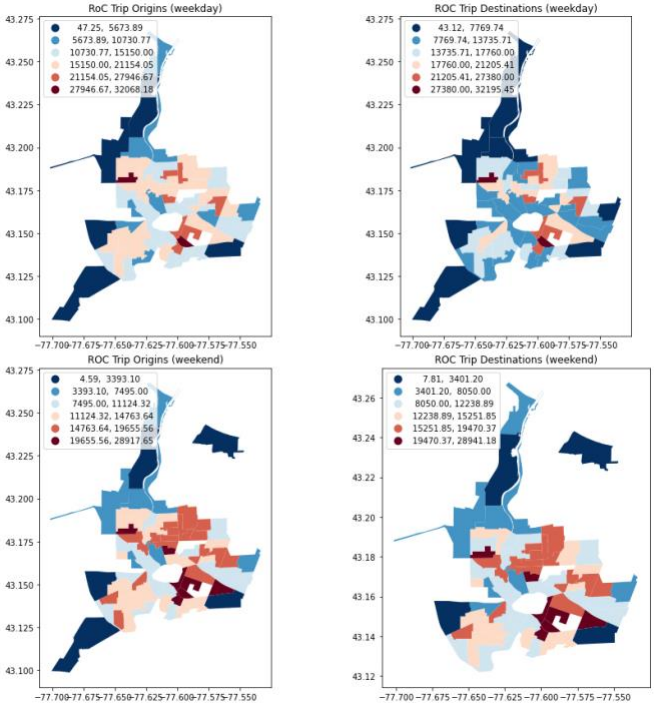


Figure 4.21. Trip density in Rochester

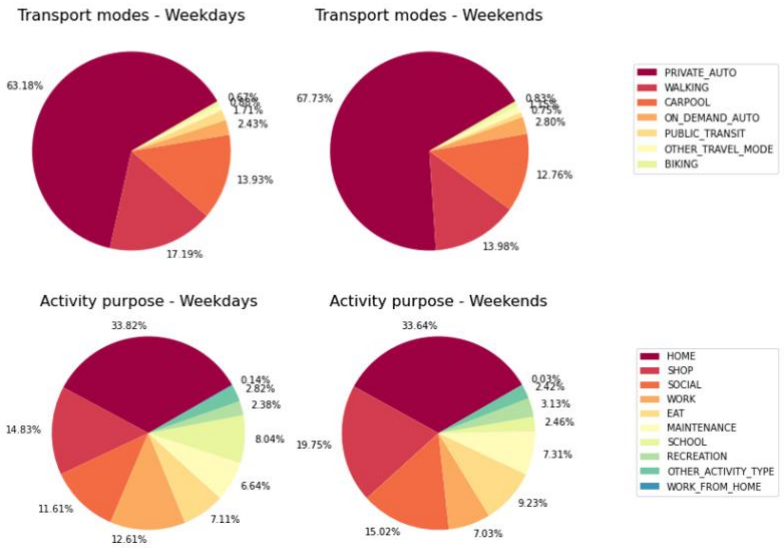


Figure 4.22. Mode share and activity purpose composition in Rochester

5. NY Statewide Mode Choice Modeling

This section presents the model results and the prediction accuracy of our NYS mode choice model. The g-AMXL model with block-group level mode choice dataset converged at the 26th iteration after 2.79 hours of Algorithm 3.1, resulting in two latent classes with different calibrated coefficients per agent that are empirically derived, revealing them to be neither Gumbel nor Gaussian. Instead, the empirical distribution seems to be a combination of a constant (assumed in multinomial logit) and Gaussian distribution (assumed in mixed logit). Moreover, the spatial distribution of agent-level coefficients reveals a regional divergence of value of time and mode preference, indicating potential inequity issues in the transportation system. This is infeasible for conventional discrete choice models (DCMs) to capture.

All of the experiments were conducted on a local machine with Intel (R) Core (TM) i7-10875H CPU and 32GB installed RAM. The Gurobi package was used to solve the LP and QP problems. A Python package for building g-AMXL with aggregated level datasets was created with an academic license to accelerate further studies (link to the tool).

5.1. Basic statistics

5.1.1. Coefficient means and goodness-of-fit

Replica's data on a typical weekday in the Fall 2019 season was used to build the g-AMXL model. The model took 2.79 hours to converge at the 26th iteration on a local machine with Intel (R) Core (TM) i7-10875H CPU and 32GB installed RAM. Since g-AMXL estimates coefficients for each block-group OD pair, we can aggregate the model results into regions that we are interested in. Table 5.1 and Table 5.2 summarize the model results in New York State (NYS) and in New York City (NYC), respectively, in which each entry represents the means of one estimated coefficient, and the number in the parenthesis is the standard error of the mean values[12]. In general, most of the coefficients are significant at the 0.001 level. The 120,740 block-group OD agents can be divided into three categories.

[12] We still calculated the standard error to get the significant level, though the estimated coefficients of g-AMXL are not assumed to be Gaussian distributed.

(1) Latent class_1 agents, of which the coefficient signs show great consistency with existing studies: auto trip time, transit in-vehicle time, and trip cost have negative signs, while auto constant has positive signs. This agent type accounts for 54.37% in NYS and 33.56% in NYC.

(2) Latent class_2 agents, of which some coefficient signs are not easy to be interpreted. For instance, the mean value of auto trip time coefficient is a large positive value, which means auto trip time derives a high utility. This agent type accounts for 37.37% in NYS and 48.15% in NYC.

(3) Infeasible agents, of which coefficients cannot be estimated by g-AMXL since the inverse optimization problems don't have any feasible solution. This agent type accounts for 8.25% in NYS and 19.29% in NYC.

The goodness-of-fit in NYS is greater than in NYC, with the McFadden Rho-Square equals to 0.523 in NYS and 0.372 in NYC. The reason might be that the proportion of class_1 agents in NYS is higher than in NYC. To this end, details of the three agent types are essential to the application of g-AMXL.

Table 5.1. Basic statistics of the model results (New York State)

Mode choice coefficients estimated by g-AMXL			
Coefficient	**Units**	**Latent Class 1**	**Latent Class 2**
$\bar{\theta}_{auto\_tt}$	1/hour	-4.566*** (0.038)	41.519*** (0.041)
$\bar{\theta}_{transit\_ivt}$	1/hour	-1.313*** (0.002)	-0.094*** (0.003)
$\bar{\theta}_{transit\_at}$	1/hour	7.967*** (0.001)	2.892*** (0.001)
$\bar{\theta}_{transit\_et}$	1/hour	8.557*** (0.001)	4.505*** (0.001)
$\bar{\theta}_{transit\_nt}$	1/transfer	8.708*** (0.004)	5.202*** (0.006)
$\bar{\theta}_{non\_auto\_tt}$	1/hour	9.394*** (0.071)	5.708*** (0.106)
$\bar{\theta}_{cost}$	1/U.S. $	-0.706*** (0.003)	-1.038*** (0.004)
$\overline{asc}_{auto}$	N/A	5.635*** (0.008)	1.116*** (0.017)
$\overline{asc}_{transit}$	N/A	-4.031*** (0.010)	-1.091*** (0.014)
$\overline{asc}_{non\_vehicle}$	N/A	-1.604*** (0.011)	-0.023 (0.023)
	*, **, and *** indicate statistical significance at the 0.05, 0.01, 0.001 levels, respectively		
	Summary Statistics Total number of agents = 120,740 Number of class 1 agents = 65,650 (54.37%) Number of class 2 agents = 45,125 (37.37%) Number of infeasible agents = 9,965 (8.25%)		

	Mode choice model performance
	Number of observations: 65,650 + 45,125 = 110,775
	Initial log likelihood = -198,482.16
	Final log likelihood = -94,708.15
	McFadden Rho-Square = 0.523

Table 5.2. Basic statistics of the model results (New York City)

Mode choice coefficients estimated by g-AMXL			
Coefficient	**Units**	**Latent Class 1**	**Latent Class 2**
$\bar{\theta}_{auto\_tt}$	1/hour	-6.744*** (0.148)	41.815*** (0.126)
$\bar{\theta}_{transit\_ivt}$	1/hour	-0.846*** (0.006)	0.356*** (0.012)
$\bar{\theta}_{transit\_at}$	1/hour	8.158*** (0.002)	3.051*** (0.004)
$\bar{\theta}_{transit\_et}$	1/hour	8.774*** (0.003)	4.689*** (0.004)
$\bar{\theta}_{transit\_nt}$	1/transfer	9.522*** (0.011)	5.900*** (0.015)
$\bar{\theta}_{non\_auto\_tt}$	1/hour	17.972*** (0.242)	14.531*** (0.378)
$\bar{\theta}_{cost}$	1/U.S. $	-0.944*** (0.014)	-1.369*** (0.027)
$\overline{asc}_{auto}$	N/A	3.512*** (0.032)	0.419*** (0.073)
$\overline{asc}_{transit}$	N/A	-0.555*** (0.036)	1.853*** (0.053)
$\overline{asc}_{non\_vehicle}$	N/A	-2.957*** (0.043)	-2.268 (0.083)
	*, **, and *** indicate statistical significance at the 0.05, 0.01, 0.001 levels, respectively		
	Summary Statistics		
	Total number of agents = 23,691		
	Number of class 1 agents = 7,950 (33.56%)		
	Number of class 2 agents = 11,408 (48.15%)		
	Number of infeasible agents = 4,333 (19.29%)		
	Mode choice model performance		
	Number of observations: 7,950 + 11,408 = 19,358		
	Initial log likelihood = -34,684.88		
	Final log likelihood = -21,790.99		
	McFadden Rho-Square = 0.372		

5.1.2. Latent class agents and infeasible agents

Figure 5.1 and Figure 5.2 visualize the spatial distribution of three agent types in NYS and in NYC, respectively. We found that the trip length of class_2 agents and infeasible agents is obviously shorter than class_1 agents. This indicates that our g-AMXL model performs better on longer distance trips compared to trips with short distances. Two reasons might account for the difference: (1) individuals' mode choices are more complicated to estimate when the trip length is relatively short, given that small trips are not that sensitive to time and cost compared to long trips; (2) small trips are more difficult to record and generate since self-reported activity schedules often exclude these trips. Though Replica identified and generated these trips with mobile phone data, it is at the risk of decreasing the data quality.

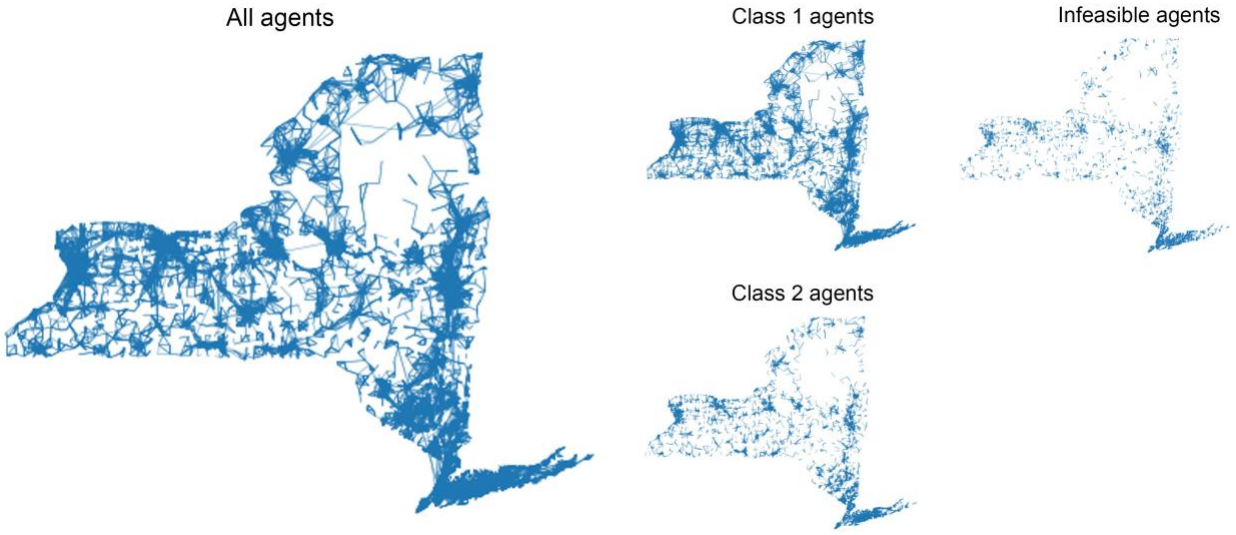


Figure 5.1. Spatial distribution of agents in New York State

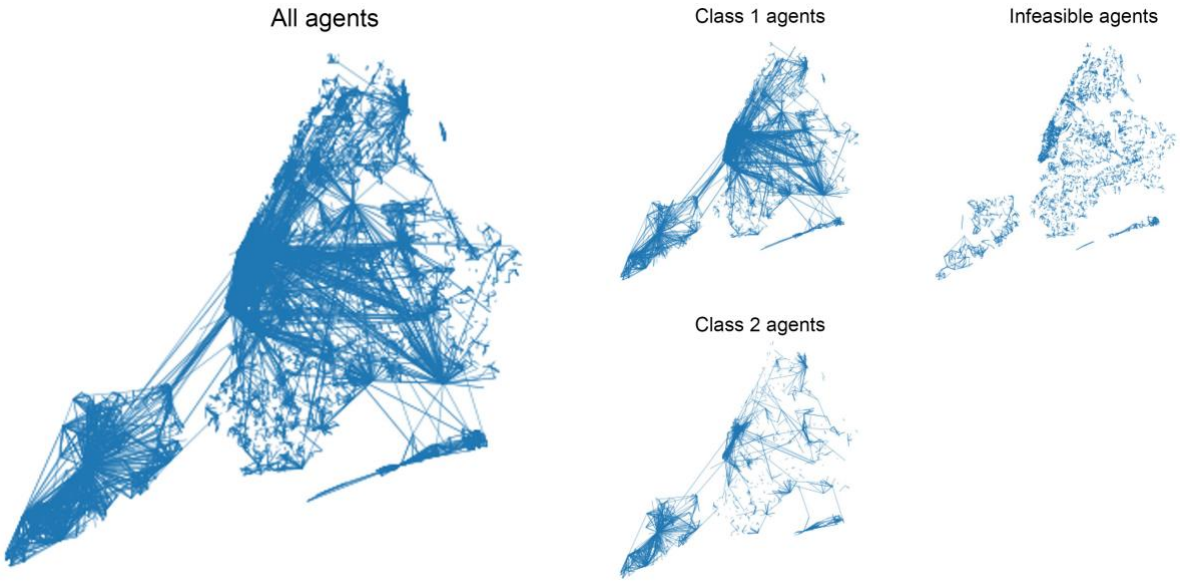


Figure 5.2. Spatial distribution of agents in New York City

Table 5.3 lists the average number of trips, average trip length, and the population composition of the three agent types, which further validated our conjecture. In New York State, the average trip length of class_1 agents is 5.14 km, while the class_2 and infeasible agents only have average trip distances of 2.67 km and 1.79 km. In New York City, the average trip distance is 2.31 km for class_1 agents, 0.74 km for class_2 agents, and 0.64 for infeasible agents. Moreover, the proportion of trips made by Student Population could be another indicator to differentiate agent types (probably because trips made by students are also insensitive to time and cost). In New York State, Student Population accounts for 33.39% trips in class_2 agents and 54.44% trips in infeasible agents, while only account for 6.15% trips in class_1 agents. Therefore, implications of the model results are two-fold. On the one side, agent-level coefficients estimated by g-AMXL are not that accurate for trips with high uncertainties such as small trips and student trips, which necessitates the K-means step to identify latent classes. On the other side, only using trip time, trip cost, and mode constant cannot explain the mode choices of all trips in NYS and NYC, more attributes should be included.

Table 5.3. Trip and population details by agent category

Mode choice agents in New York State			
	Class_1 agent	**Class_2 agent**	**Infeasible agent**
Average number of trips (per agent per day)	93.76	92.74	86.34
Average trip length (km)	5.14	2.67	1.79

% LowIncome Population	9.87%	9.08%	12.96%
% NotLowIncome Population	69.88%	45.69%	28.40%
% Senior Population	14.11%	11.84%	4.20%
% Student Population	6.15%	33.39%	54.44%
Mode choice agents in New York City			
	Class_1 agent	**Class_2 agent**	**Infeasible agent**
Average number of trips (per agent per day)	83.30	85.31	79.91
Average trip length (km)	2.31	0.74	0.64
% LowIncome Population	12.52%	9.56%	18.23%
% NotLowIncome Population	67.32%	52.79%	40.09%
% Senior Population	6.58%	5.16%	5.75%%
% Student Population	13.58%	32.49%	35.93%

5.2. Distribution of agent-specific coefficients

5.2.1. Statistical distribution

Figure 5.3 (a)-(c) shows the mean value of each coefficient in each iteration. The estimated coefficients per agent are empirically derived, revealing to be neither Gumbel nor Gaussian. Instead, the empirical distribution seems to be a combination of a constant (assumed in MNL) and Gaussian distribution (assumed in MXL).

According to Figure 5.3 (d)-(i), coefficients can be divided into three categories: (1) highly-concentrated coefficients, such as transit in-vehicle time (transit_ivt), transit access time (transit_at), transit egress time (transit_et), and trip cost (cost). These coefficients are concentrated around their mean values with small variations, reflecting homogenous tastes among agents; (2) even-distributed coefficients, such as auto mode constant (constant_auto), transit mode constant (constant_transit), and non-vehicle mode constant (constant_non_vehicle). These coefficients have larger variations and follow normal distributions, reflecting heterogeneous tastes among agents, and; (3) asymmetrical-distributed coefficients, such as auto trip time (auto_tt), non-vehicle trip time (non_vehicle_tt), and number of transfer for transit (transit_nt). These coefficients vary obviously among different latent classes and reflect heterogeneous tastes that are not Gaussian distributed (probably because there are some unobserved attributes). To this end, g-AMXL provides a flexible approach for modelers to capture inter-agent homogeneities and heterogeneities, which are infeasible in MNL and MXL.

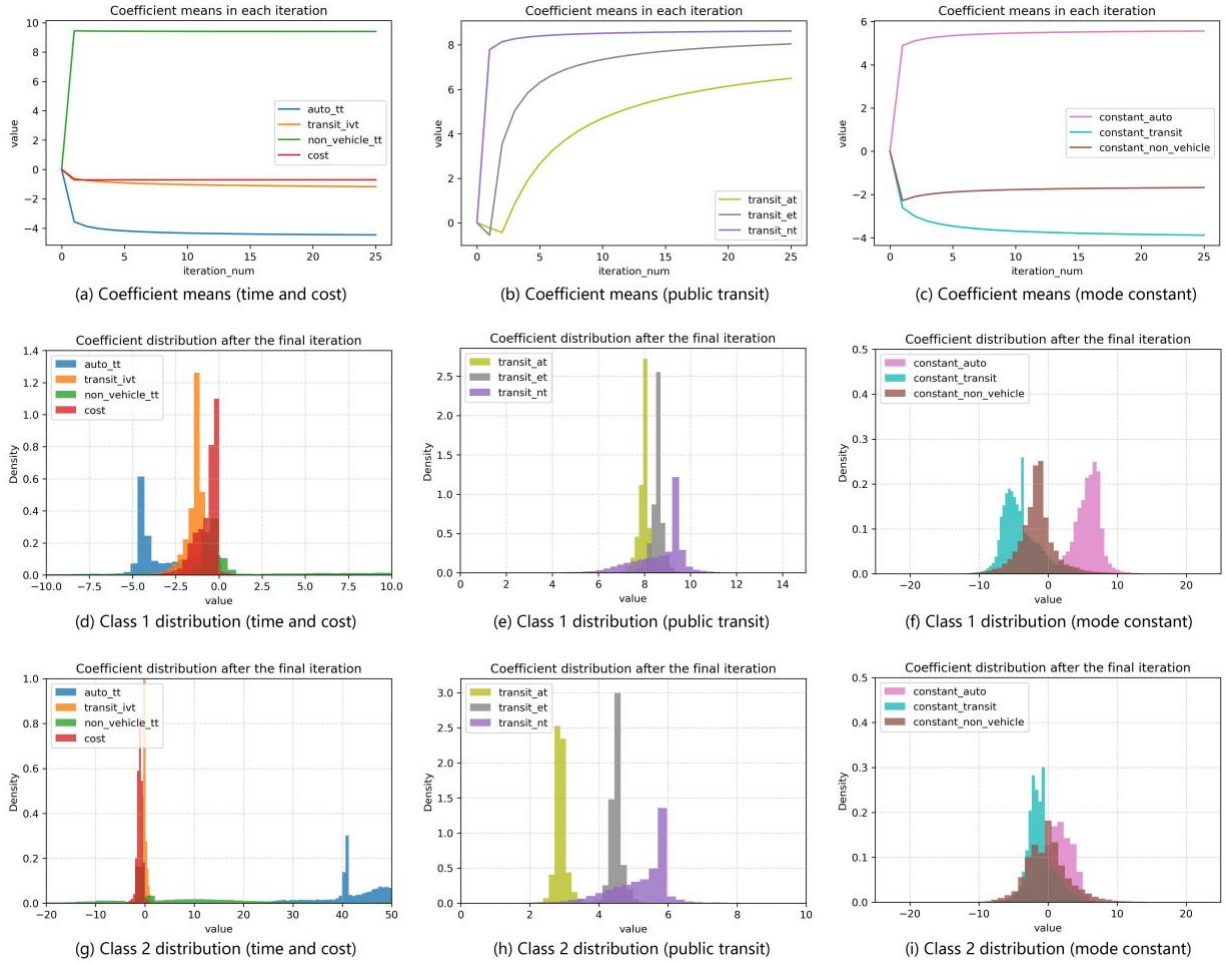


Figure 5.3. Mean values and the distribution of estimated coefficients. In (a)-(c), x-axis is the number of iterations, y-axis is the value of fixed-point prior θ_0. In (d)-(i), x-axis is the value of estimated coefficients, y-axis is the probability density.

5.2.2. Spatial distribution

Since each set of coefficients is linked to a block-group OD pair, we can further check the spatial distribution. Table 5.4 lists the average value of time (VOT, measured as $\bar{\theta}_{auto\_tt}/\bar{\theta}_{cost}$) of four population segments in New York State and New York City. The results are consistent with existing studies and our empirical knowledge by: (1) the average VOT in New York City is generally higher than in New York State; (2) NotLowIncome Population have the highest VOT (13.78$/hour in NYS, 18.74$/hour in NYC), while LowIncome Population have relatively lower VOT (3.05 $/hour in NYS, 14.59$/hour in NYC).

Table 5.4. Value of time (VOT) of different population segments

	Average VOT in New York State	Average VOT in New York City
NotLowIncome Population	13.78 $/hour	18.74 $/hour
LowIncome Population	3.05 $/hour	14.59 $/hour
Senior Population	9.98 $/hour	5.31 $/hour
Student Population	10.78 $/hour	17.07 $/hour

Figure 5.4-5.7 present the spatial distribution of value of time (VOT), coefficient of auto mode constant ($asc_{auto,i}$), coefficient of transit mode constant ($asc_{transit,i}$), and coefficient of non-vehicle mode constant ($asc_{non\_vehicle,i}$), which is infeasible for DCMs to capture. The spatial distribution of coefficients reveals the regional differences of mode choice preferences:

(1) The value of time in New York City is generally higher than other areas. Within NYC, trips related to Manhattan and trips pointing to JFK airport have relatively higher value of time.

(2) The coefficient of auto mode constant is positive (prefer auto modes) in rural areas, while negative (do not prefer auto modes) in urban areas, especially in Manhattan in NYC.

(3) The coefficient of transit mode constant is only positive (prefer transit mode) in NYC and downtown areas of other cities, while negative (do not prefer transit mode) in other areas.

(4) The coefficient of non-vehicle mode constant is relatively higher (prefer biking and walking) in NYC, while no obvious pattern is found throughout NYS and NYC.

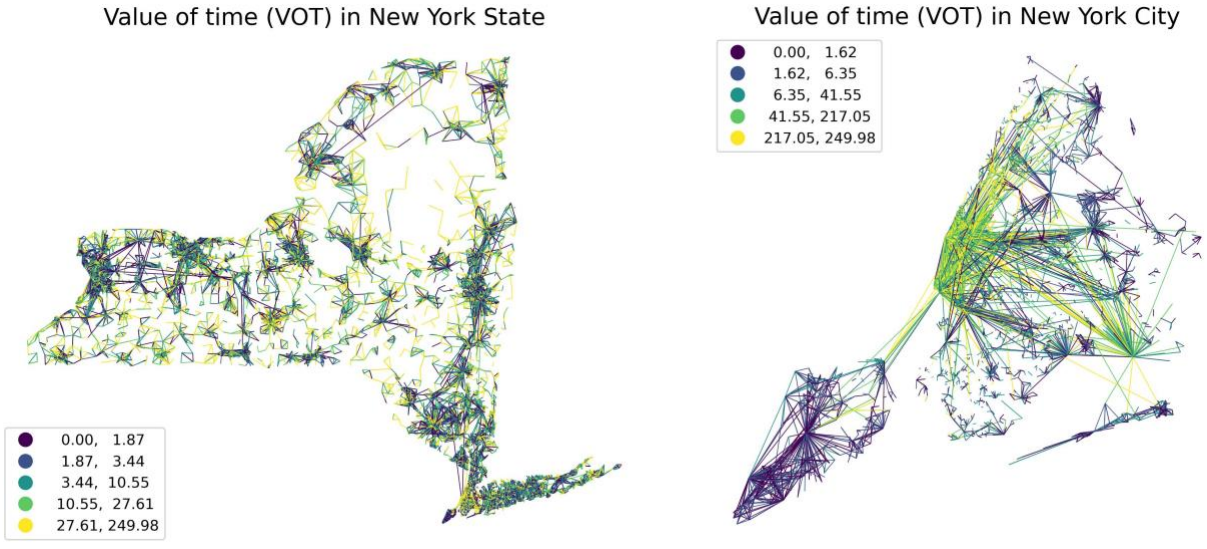


Figure 5.4. Spatial distribution of value of time (VOT)

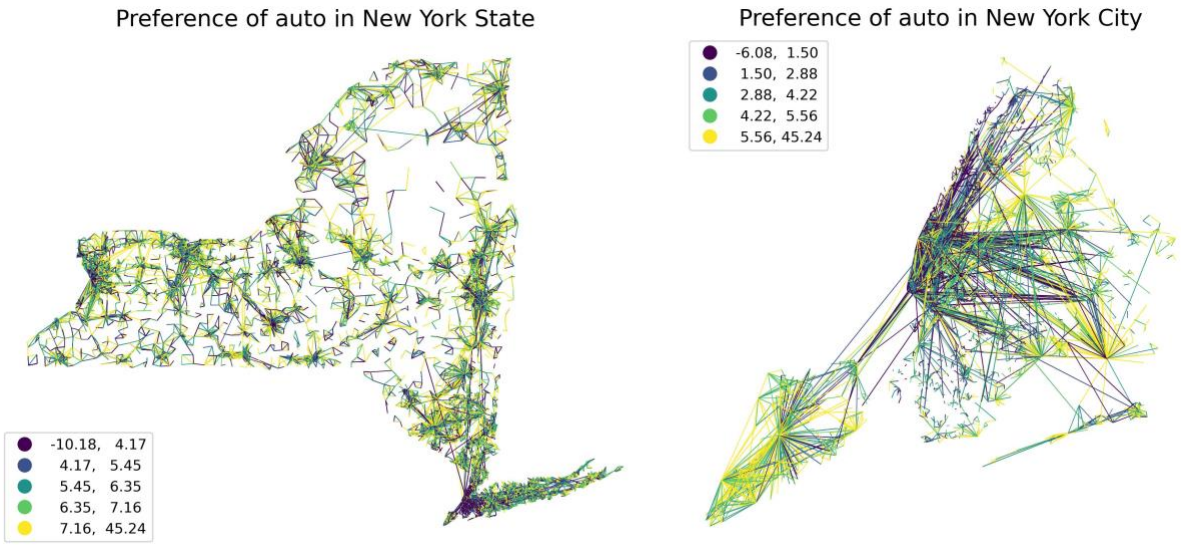


Figure 5.5. Spatial distribution of auto mode coefficient ($asc_{auto,i}$)

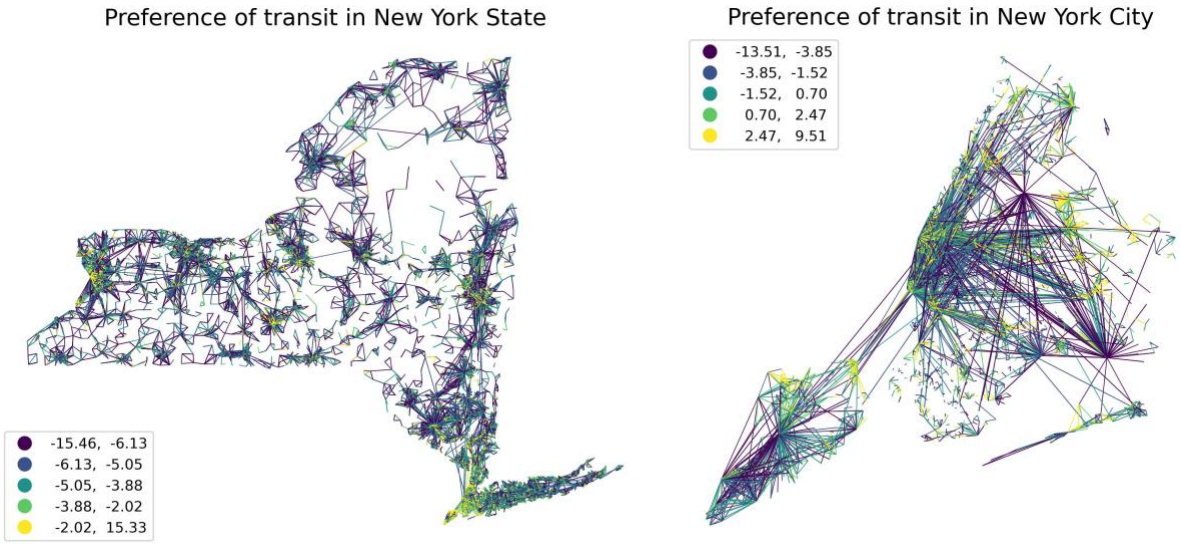


Figure 5.6. Spatial distribution of transit mode coefficient ($asc_{transit,i}$)

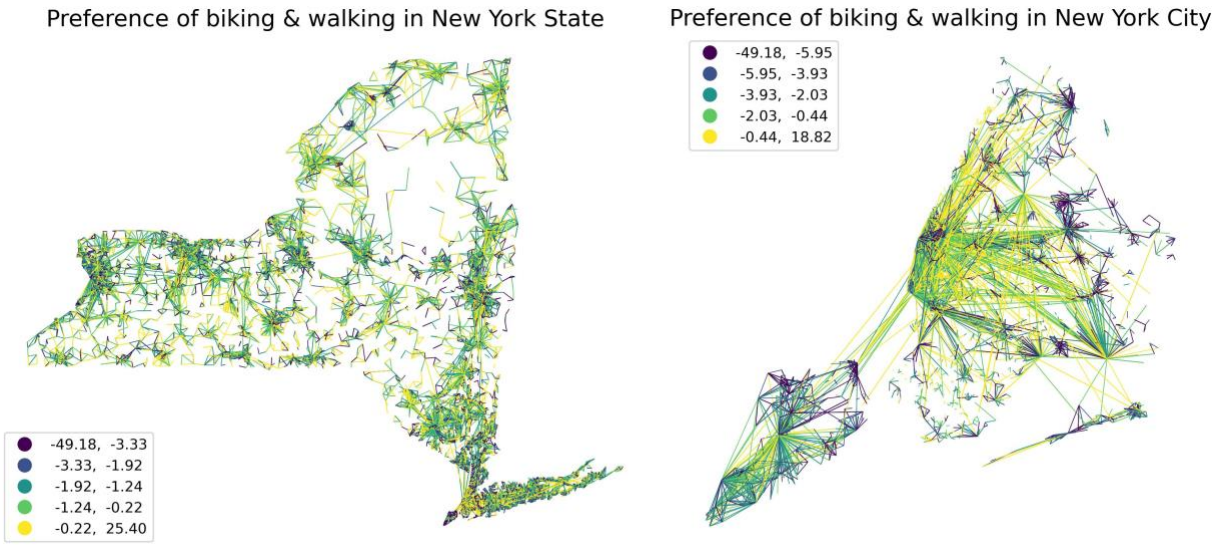


Figure 5.7. Spatial distribution of non-vehicle mode coefficient ($asc_{non\_vehicle,i}$)

5.3. Prediction accuracy

5.3.1. Performance measures

The prediction accuracy of block-group OD pair mode share is defined as Eq. (5.1).

$$acc_{uw} = \sum_{h=1}^{|H|} \sum_{k=1}^{|K|} min\,(P_{uw,h}^{k}, S_{uw,h}^{k}), \qquad \forall u, w \in N, u \neq w \tag{5.1}$$

where acc_{uw} is the prediction accuracy of mode share from block group u to w; K is the mode choice set; N is the set of all census block groups; H is the set of population segment ID; $P_{uw,h}^{k}$ and $S_{uw,h}^{k}$ are predicted and observed proportion of mode k from block group u to w given population segment h.

Based on the acc_{uw}, two performance metrics can be calculated: (1) the overall prediction accuracy, which is measured as the weighted sum of block-group OD pair accuracies according to the trip volume. Eq. (5.2) defines the overall accuracy, where d_{uw} is the trip volume per day from block group u to w; (2) the mean absolute error of the prediction, which is measured as the total error per OD pair divided by the number of mode alternatives as defined in Eq. (5.3).

$$ACC = \frac{\sum_{u=1}^{|N|}\sum_{w=1}^{|N|}(d_{uw} * acc_{uw})}{\sum_{u=1}^{|N|}\sum_{w=1}^{|N|} d_{uw}} \tag{5.2}$$

$$mas_{uw} = \frac{1 - acc_{uw}}{|K|}, \qquad \forall u, w \in N, u \neq w \tag{5.3}$$

5.3.2. In-sample and out-of-sample accuracy

We measured both in-sample and out-of-sample accuracy considering that the AMXL model might have the risk of overfitting (Ren and Chow, 2022). Since we built the g-AMXL with Replica's trip data on weekday, we used weekday's dataset to calculate the in-sample accuracy and used weekend's dataset to calculate the out-of-sample accuracy. Moreover, we eliminated infeasible agents when measuring prediction accuracy since we cannot estimate coefficients for these agents.

Table 5.5 shows the performance metrics of our model. The in-sample accuracy of g-AMXL model is quite competitive, with an overall accuracy of 90.28% in NYS and 88.63% in NYC. The mean absolute error is 0.016 in NYS (with a standard deviation of 0.013), which means the estimation error per mode share per agent is around 1.6%. When it comes to out-of-sample accuracy, however, the performance of g-AMXL drops significantly, with an overall accuracy of 76.98% in NYS and 65.52% in NYC.

Table 5.5. Model Performance in NYS and NYC

	New York State	New York City
Overall accuracy (in-sample)	90.28%	88.63%
Overall accuracy (out-of-sample)	76.98%	65.52%
Mean absolute error (in-sample)	0.016 (0.013)	0.018 (0.015)
Mean absolute error (out-of-sample)	0.040 (0.037)	0.062 (0.046)

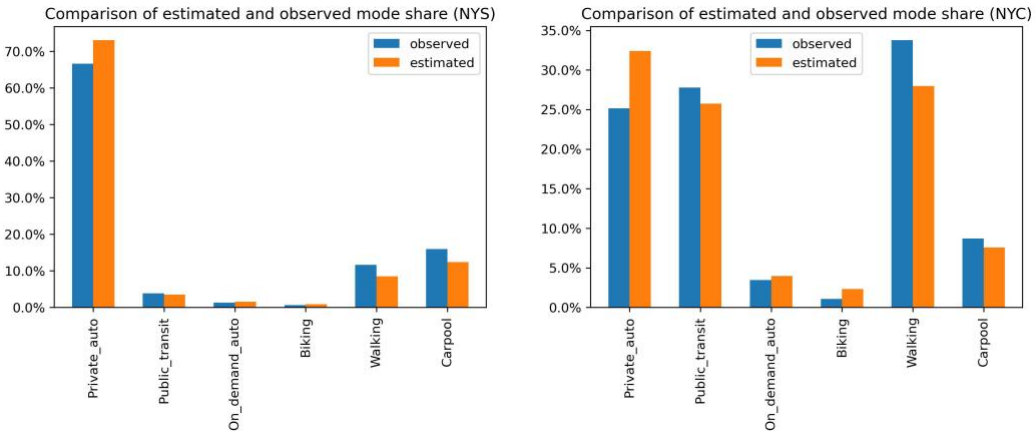


Figure 5.8. Estimated and observed mode share (in-sample, weekday)

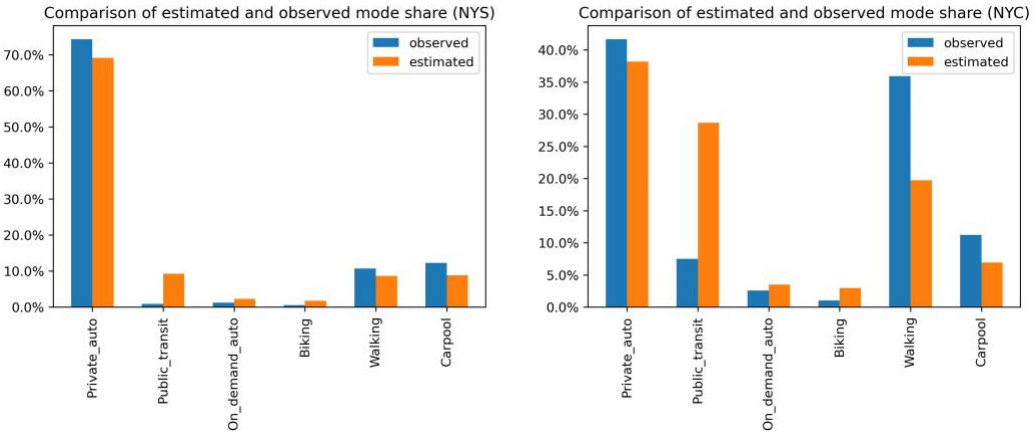


Figure 5.9. Estimated and observed mode share (out-of-sample, weekend)

It is noted that the difference between in-sample and out-of-sample accuracy in our project is not due to the overfitting issue mentioned in Ren and Chow (2022)'s study. Figure 5.8 and Figure 5.9 show the estimated and observed mode share on the weekday (in sample) and the weekend (out of sample), from which we can find obvious differences between observed mode share on the weekday and the weekend: public transit proportion on weekday is higher than on weekend, and walking proportion on weekday is higher than on weekend. Therefore, the drop of out-of-sample accuracy might be due to the difference of training and test data instead of overfitting issues of the g-AMXL. We propose two ways to overcome this issue: either to build separate models for weekday and weekend, or to include attributes such as activity purpose and day-of-week dummy variables.

6. Equity-based Service Region Optimization Tool

This section showcases another advantage of the g-AMXL model, that is, the estimated coefficients can efficiently be integrated into system optimization models. Extracting demand from a mixed logit model is possible but would require simulation to predict the response of travelers to determine the optimal solution (Pacheco et al., 2021). This lacks efficiency and consistency since simulation is time consuming and the randomness would produce different results. In the g-AMXL model, we can directly get the demand response through deterministic calculation using agent-level coefficients, which enables system optimization on the statewide scale.

The proposed tool is written in Python codes and wrapped into a package available with academic license (link to the tool). The use case of the tool includes single-service region design and multi-service region assortment. We assume that there will be new mobility services entering into the market. Each service selects counties in NY state as the service zones, in which a few vehicles should provide bi-direction trip services to meet the demands on the selected OD links. Each vehicle has a maximum service distance and a maximum number of trips per day. Our tool takes about 8 min for each service region optimization on a local machine with Intel (R) Core (TM) i7-10875H CPU and 32GB installed RAM. The outputs include the optimal service region and equity impact metrics given a budget level and one of the three objectives including: (1) maximizing the total revenue; (2) maximizing the total welfare (change of consumer surplus); (3) minimizing the welfare disparity between disadvantaged and non-disadvantaged communities. These enable transportation planners and policymakers to design new mobility services that consider statewide transportation equity.

6.1. Pre-settings of the tool

There are four setup steps before running optimization using our tool: (1) setting the price level and performance of a new mobility service (for single-service use case) or two mobility services (for multi-service use case); (2) setting the maximum number of service zones (O) and maximum fleet size ($\mathcal{F}_{max}$) to define the budget level; (3) selecting one of the three aforementioned objectives; (4) adjusting default parameters if necessary. Table 6.1 lists the inputs required by our equity-based decision support tool. Users should input these parameters manually to set up a service region optimization problem (see instructions of the tool).

Table 6.1. Inputs of the equity-based decision support tool

(1) Price level and performance	
$c_{uw}^{k^*}$	Trip fare ($) of the new mobility service k^* from block group u to w
$t_{uw}^{k^*}$	Trip duration (hour) of the new mobility service k^* from block group u to w
(2) Budget level	
O	The maximum number of operating zones
$\mathcal{F}_{max}$	The maximum fleet size in total (vehicles/day)
(3) Objective function	
Obj	One of the three strings: “Objective 1”, “Objective 2”, “Objective 3”
(4) Default parameters	
F_{max}	The maximum fleet size in each service zone, default to $2\mathcal{F}_{max}/O$
F_{min}	The minimum fleet size in each service zone, default to $\mathcal{F}_{max}/2O$
L	The maximum distance (km) a vehicle can serve per day, default to 200 km
D	The maximum number of trips a vehicle can serve per day, default to 10 trips

To showcase how the proposed tool works, the rest of this section presented two examples, one for single-region design and another for multi-service region assortment. For each example, we defined the new mobility services and ran the optimization under the three objectives and three budget levels (nine scenarios per use case). The outputs include equity metrics for users to track the revenue and equity impacts of the mobility services, as well as online map visualizing the optimal service region in space.

6.2. Example of single-service region design

For the single-service region design, we assume that there will be a new mobility service entering into the market. The trip fare of the mobility service is half of the on-demand mode, and the trip duration of the mobility service equals to the trip duration of on-demand mode plus a five-minute waiting time[13]. Three aforementioned objectives are considered, and three budget levels are set,

[13] We use relative trip fare and duration for simplicity, while users can define the price level and performance using their own cost functions.

including: (1) $O = 5,\ \mathcal{F}_{max} = 2{,}000$; (2) $O = 10,\ \mathcal{F}_{max} = 5{,}000$; (1) $O = 10,\ \mathcal{F}_{max} = 10{,}000$. All of the other parameters are set to default.

Table 6.2 summarizes the number of operating links, VMT per vehicle per day, total revenue, and equity metrics of the single-service region optimization under different objectives and budget levels. Each entry represents the value of a metric, and the number in the parenthesis is the percentage change compared to baseline (without any new mobility services). Total revenue of objective 1, average welfare of objective 2, and welfare disparity of objective 3 are in bold font, from which we can see the optimization tool indeed maximize or minimize the objective function. It is noted that the percentage change of the metrics is relatively small (most of them are smaller than 1%), which is because the new mobility service can only impact a small part of the total trips. For instance, a maximum fleet size of 2,000 vehicles can at most serve 20,000 trips with a maximum vehicle trip of 10 per vehicle per day, only accounting for 0.5% of the total trips in NYS. Therefore, some statewide equity metrics like average welfare, welfare mean deviation, and welfare deviation changed slightly after the introduction of the new mobility service, though their signs align with our expectation. In general, the equity metrics indicate that maximizing total revenue and maximizing total welfare will increase transportation inequities by increasing the welfare disparity by up to 0.40% and 0.59%, respectively. On the other hand, minimizing welfare disparity between disadvantaged communities and other communities helps to decrease transportation inequities, by effectively decreasing the welfare disparity by up to 7.37%, though this is at the cost of losing total revenue.

Table 6.2. Metrics of single-service region design optimization

	Baseline	Objective 1	Objective 2	Objective 3
A. 5 zones, 2,000 vehicles				
Number of operating links	-	694	1,387	1,492
VMT/vehicle (km/day)	-	63	57	33
Total revenue ($)	-	**148,635**	136,442	73,223
Average welfare	5.443	5.452 (+0.16%)	**5.454 (+0.19%)**	5.446 (+0.4%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)
Welfare mean deviation	1.429	1.431 (+0.04%)	1.430 (+0.05%)	1.430 (0.00%)
Welfare variation	3.918	3.921 (+0.07%)	3.896 (-0.56%)	3.919 (+0.01%)
Welfare Gini coefficient	0.188	0.188 (0.06%)	0.188 (-0.19%)	0.188 (-0.04%)

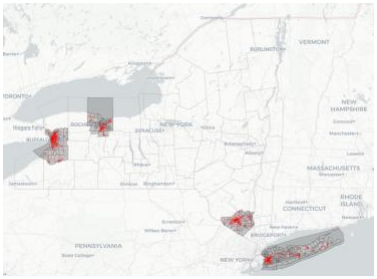
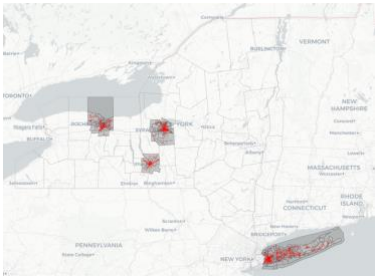
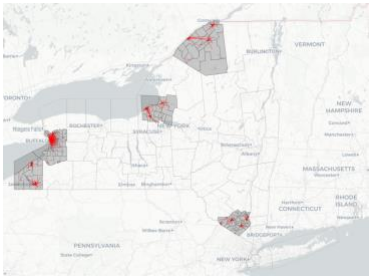
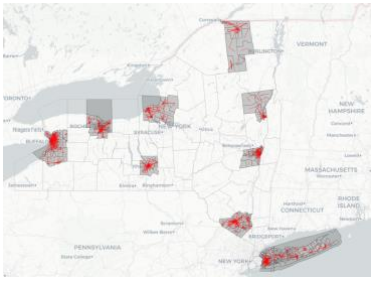
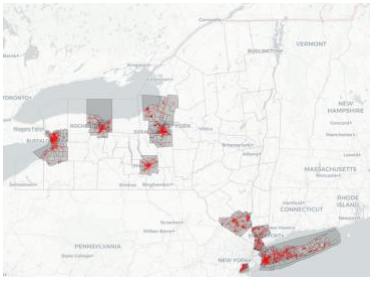
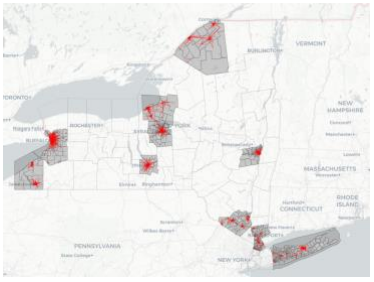
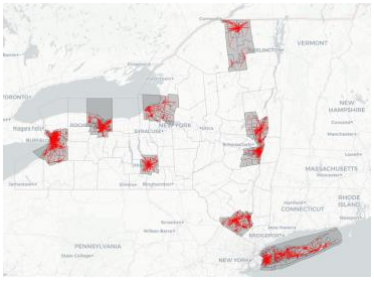
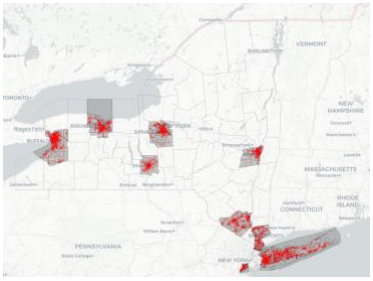
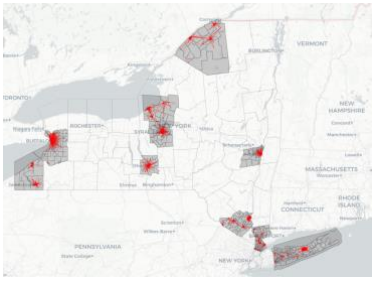
	Baseline	Objective 1	Objective 2	Objective 3
Welfare disparity	0.482	0.483 (+0.02%)	0.484 (+0.37%)	**0.471 (-2.48%)**
B. 10 zones, 5,000 vehicles				
Number of operating links	-	2,503	3,445	3,439
VMT/vehicle (km/day)	-	62	53	33
Total revenue ($)	-	**350,623**	308,357	171,805
Average welfare	5.443	5.460 (+0.29%)	**5.462 (+0.33%)**	5.449 (+0.10%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)
Welfare mean deviation	1.429	1.432 (+0.14%)	1.431 (+0.04%)	1.429 (-0.03%)
Welfare variation	3.918	3.927 (+0.25%)	3.901 (-0.41%)	3.916 (-0.04%)
Welfare Gini coefficient	0.188	0.188 (-0.09%)	0.187 (-0.27%)	0.188 (-0.13%)
Welfare disparity	0.482	0.483 (+0.06%)	0.484 (+0.38%)	**0.459 (-4.84%)**
C. 10 zones, 10,000 vehicles				
Number of operating links	-	6,151	7,050	5,202
VMT/vehicle (km/day)	-	55	46	28
Total revenue ($)	-	**593,845**	538,666	237,878
Average welfare	5.443	5.470 (+0.49%)	**5.476 (+0.59%)**	5.452 (+0.14%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)
Welfare mean deviation	1.429	1.432 (+0.14%)	1.432 (+0.14%)	1.429 (-0.04%)
Welfare variation	3.918	3.914 (-0.10%)	3.904 (-0.36%)	3.915 (-0.06%)
Welfare Gini coefficient	0.188	0.187 (-0.34%)	0.187 (-0.43%)	0.188 (-0.18%)
Welfare disparity	0.482	0.485 (+0.40%)	0.486 (+0.59%)	**0.447 (-7.37%)**

Note: each entry represents the value of a metric, and the number in the parenthesis is the percentage change compared to baseline

Table 6.3 lists the optimal service region given different objectives and budget levels, in which each figure is a snapshot of the online maps visualizing operated links and serviced counties (see the online map by clicking the link behind each figure). We can find that maximizing total revenue and maximizing total welfare result in similar service regions. Both of them cover Buffalo, Rochester, Syracuse, Ithaca, Albany, and Long Island. It is noted that New York City is not included into the service region, probability because the auto mode share and trip distance in NYC is relatively low, resulting in lower revenue and lower increase of total welfare. Also, the

similarity between service regions under objective 1 and 2 can be explained by the concept of compensating variation (CV). On the contrary, optimal service regions under objective 3 show considerable difference by covering more links and counties in rural areas.

Table 6.3. Visualization of optimal strategies (single-service region design)

A. 5 zones, 2,000 vehicles		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity
		
(Click here for online map)	(Click here for online map)	(Click here for online map)
B. 10 zones, 5,000 vehicles		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity
		
(Click here for online map)	(Click here for online map)	(Click here for online map)
C. 10 zones, 10,000 vehicles		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity
		
(Click here for online map)	(Click here for online map)	(Click here for online map)

6.3. Example of multi-service region assortment

For the multi-service region assortment, we assume that there will be two new mobility services entering into the market. Service A is the same as what we've defined in Section 6.2. Service B is relatively more expensive but has higher performance. The trip fare of the service B is 1.2 times of the on-demand mode, and the trip duration of the mobility service equals to the trip duration of driving. We also consider three objectives and three budget levels, and all the other parameters are set to default.

Table 6.4 summarizes the number of operating links, VMT per vehicle per day, total revenue, and equity metrics of the multi-service region optimization under different objectives and budget levels. Each entry represents the value of a metric, and the number in the parenthesis is the percentage change compared to baseline (without any new mobility services). Total revenue of objective 1, average welfare of objective 2, and welfare disparity of objective 3 are in bold font, from which we can see the optimization tool indeed maximize or minimize the objective function. Compared with the single-service example, the total revenue (under objective 1) is higher while the decrease of welfare disparity (under objective 3) is smaller, which aligns with our expectation after adding a service with higher trip fare while shorter trip duration. In general, the equity metrics indicate that minimizing welfare disparity between disadvantaged communities and other communities helps to decrease transportation inequities, by effectively decreasing the welfare disparity by up to 6.19%. However, maximizing total revenue and maximizing total welfare won't always increase or decrease transportation inequities.

Table 6.4. Metrics of multi-service region assortment optimization

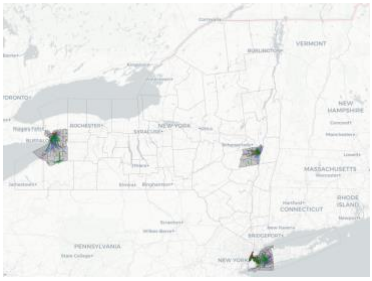
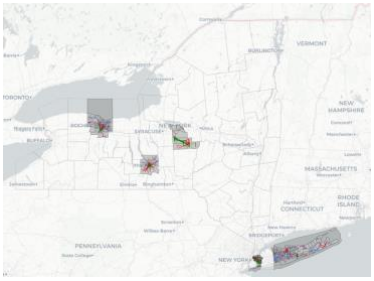
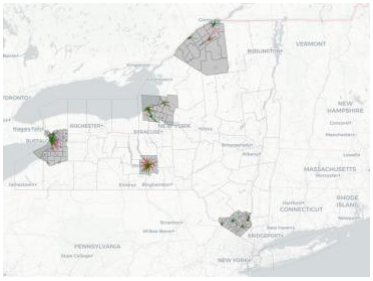
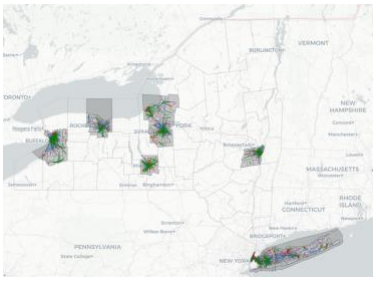
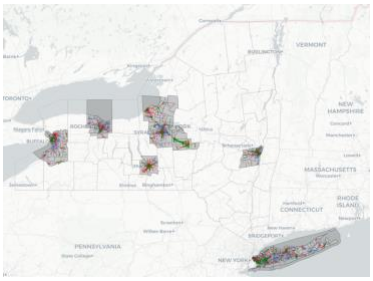
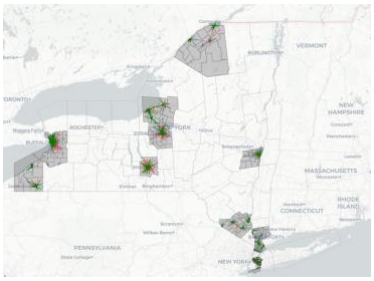
	Baseline	Objective 1	Objective 2	Objective 3
A. 5 zones, 2,000 vehicles (1,000 for service A, 1,000 for service B)				
Number of operating links	-	694 (service A) 571 (service B)	652 (service A) 619 (service B)	658 (service A) 827 (service B)
VMT/vehicle (km/day)	-	92 (service A) 49 (service B)	57 (service A) 37 (service B)	37 (service A) 27 (service B)
Total revenue ($)	-	**110,069 (service A)** **176,977 (service B)**	64,025 (service A) 132,993 (service B)	42,305 (service A) 63,175 (service B)
Average welfare	5.443	5.452 (+0.15%)	**5.453 (+0.17%)**	5.446 (+0.04%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)

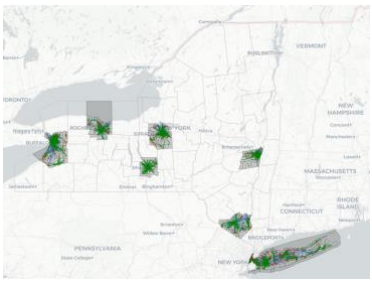
	Baseline	Objective 1	Objective 2	Objective 3
Welfare mean deviation	1.429	1.430 (+0.02%)	1.428 (-0.14%)	1.429 (-0.03%)
Welfare variation	3.918	3.917 (-0.03%)	3.870 (-1.21%)	3.916 (-0.05%)
Welfare Gini coefficient	0.188	0.187 (-0.20%)	0.187 (-0.32%)	0.188 (-0.07%)
Welfare disparity	0.482	0.481 (-0.30%)	0.481 (-0.15%)	**0.472 (-2.22%)**
B. 10 zones, 5,000 vehicles (2,500 for service A, 2,500 for service B)				
Number of operating links	-	1,808 (service A) 1,647 (service B)	1,713 (service A) 1,793 (service B)	1,828 (service A) 2,188 (service B)
VMT/vehicle (km/day)	-	89 (service A) 40 (service B)	60 (service A) 31 (service B)	32 (service A) 24 (service B)
Total revenue ($)	-	**255,170 (service A)** **346,743 (service B)**	172,953 (service A) 283,608 (service B)	92,383 (service A) 121,634 (service B)
Average welfare	5.443	5.458 (+0.27%)	**5.461 (+0.31%)**	5.449 (+0.09%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)
Welfare mean deviation	1.429	1.430 (+0.01%)	1.428 (-0.06%)	1.429 (-0.08%)
Welfare variation	3.918	3.920 (+0.06%)	3.896 (-0.56%)	3.913 (-0.12%)
Welfare Gini coefficient	0.188	0.187 (-0.27%)	0.186 (-0.33%)	0.187 (-0.16%)
Welfare disparity	0.482	0.483 (+0.04%)	0.480 (-0.36%)	**0.471 (-4.95%)**
C. 10 zones, 10,000 vehicles (5,000 for service A, 5,000 for service B)				
Number of operating links	-	3,601 (service A) 3,804 (service B)	3,574 (service A) 4,291 (service B)	3,294 (service A) 3,358 (service B)
VMT/vehicle (km/day)	-	73 (service A) 26 (service B)	53 (service A) 20 (service B)	34 (service A) 19 (service B)
Total revenue ($)	-	**411,354 (service A)** **473,777 (service B)**	319,323 (service A) 374,240 (service B)	122,634 (service A) 165,073 (service B)
Average welfare	5.443	5.469 (+0.46%)	**5.473 (+0.53%)**	5.451 (+0.12%)
Welfare range	72.597	72.597 (0.00%)	72.597 (0.00%)	72.597 (0.00%)
Welfare mean deviation	1.429	1.431 (+0.10%)	1.429 (-0.02%)	1.428 (-0.10%)
Welfare variation	3.918	3.925 (+0.23%)	3.894 (-0.61%)	3.913 (-0.14%)
Welfare Gini coefficient	0.188	0.186 (-0.37%)	0.186 (-0.48%)	0.187 (-0.21%)
Welfare disparity	0.482	0.484 (+0.08%)	0.479 (-0.82%)	**0.453 (-6.19%)**

Note: each entry represents the value of a metric, and the number in the parenthesis is the percentage change compared to baseline

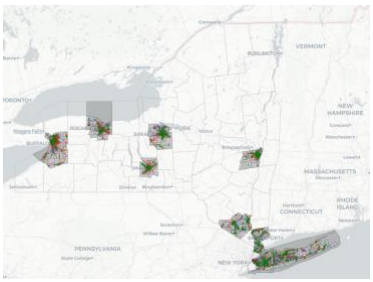
Table 6.5 lists the optimal service region given different objectives and budget levels, in which each figure is a snapshot of the online maps visualizing operated links and serviced counties (see the online map by clicking the link behind each figure). Blue color denotes the links on which only service A operates, red color denotes the links on which only service B operates, and color green denotes the links on which both service A and B operate. We can find that maximizing total revenue and maximizing total welfare result in similar service regions, while with the enter of service B, Manhattan was included into the service region (probability because the high value of time in Manhattan). Also, optimal service regions under objective 3 show considerable difference by covering more links and counties in rural areas.

Table 6.5. Visualization of optimal strategies (multi-service region assortment)

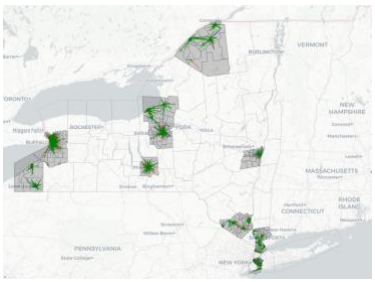
A. 5 zones, 2,000 vehicles (1,000 for service A, 1,000 for service B)		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity
		
(Click here for online map)	(Click here for online map)	(Click here for online map)
B. 10 zones, 5,000 vehicles (2,500 for service A, 2,500 for service B)		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity
		
(Click here for online map)	(Click here for online map)	(Click here for online map)
C. 10 zones, 10,000 vehicles (5,000 for service A, 5,000 for service B)		
1.Maximizing total revenue	2.Maximizing total welfare	3.Minimizing welfare disparity



(Click here for online map)



(Click here for online map)



(Click here for online map)

7. Conclusion

Transportation equity refers to the principle of ensuring that all population groups have similar access to safe, affordable, and reliable transportation options. A statewide decision support tool can play a critical role in achieving transportation equity by providing transportation planners and policymakers with a comprehensive, data-driven, and transparent approach. Though the availability of large-scale ICT data makes it possible to develop behavioral models for a range of different population segments, innovative methods are required to deal with big datasets, capture regional heterogeneities, and efficiently optimize the system design.

In this collaborative project with Replica Inc., we first ingested Replica's synthetic datasets in New York State, which include the synthetic population and their trips on a typical Thursday and Saturday in the Fall 2019 season (September 2019 – November 2019). We then proposed g-AMXL model to deterministically fit heterogeneous coefficients for trips along each census block-group OD pair conducted by each population segment. Finally, the estimated coefficients were directly integrated into our service region optimization tool. Outputs of the proposed tool include the optimal service region and equity impact metrics given a budget level and one of the three objectives including: (1) maximizing total revenue; (2) maximizing total welfare; (3) minimizing welfare disparity. These enable transportation planners and policymakers to design new mobility services that consider statewide transportation equity.

The significance of g-AMXL model is three-fold. First, the g-AMXL takes OD level (instead of individual level) trip data as inputs, which is efficient in dealing with ubiquitous datasets containing millions of observations. This enables us to model mode choice with 48,898,993 trips made by 19,568,859 residents in New York State. Despite the large data size, our model took only 2.79 hours to converge at the 26th iteration.

Second, preference heterogeneities are based on non-parametric aggregation of coefficients per agent instead of having to assume a distributional fit. The spatial distribution of agent-level coefficients reveals a regional divergence of value of time and mode preference, which is infeasible for conventional discrete choice models (DCMs) to capture. Moreover, statistics of these coefficients helps us to figure out irregular agents that might indicate unobserved attributes or poor data quality.

Third, g-AMXL can be directly integrated into system design optimization models as constraints instead of dealing with simulation-based approaches required by mixed logit (MXL)

models. With agent-level coefficients, single-service region design can be formulated as a linear programming (LP) problem, and multi-service region assortment can be formulated as a quadratic programming (QP) problem. Our decision support tool took about 8 min to solve a statewide service region optimization problem, which enables us to quickly compare optimal strategies and their equity impacts under different objectives and budget levels.

The practical contribution of our project is that we checked the quality of Replica's synthetic data before building models with these emerging datasets. We found that the population size and sociodemographic information are quite reliable, while the origin-destination flows at the tract level cannot fit perfectly to CTPP data. To this end, models dealing with ubiquitous datasets should be capable to identify bad data points. The proposed g-AMXL has certain capability due to its agent-level estimation and K-Means algorithm. We found that coefficients cannot be estimated for infeasible agents (account for 8.25% in NYS), and most of these agents are small trips or trips made by students. A practical use of our g-AMXL model is to figure out trips with irregular coefficients and then conduct case studies to see what happens there.

Moreover, our equity-based decision support tool provides a data-driven approach for transportation planners and policymakers to consider statewide transportation equity. According to our experiments, maximizing total revenue and maximizing total welfare result in similar service regions, which can be explained by the concept of compensating variation. However, they might increase transportation inequities by increasing the welfare disparity. Minimizing welfare disparity between disadvantaged communities and other communities can effectively decrease the welfare disparity by up to 7.37%, though this is at the cost of losing revenue. These provide quantitative supports for the future system design.

There are many new research opportunities and use cases to be addressed. While the project looks only at mode choice on weekdays, the g-AMXL model can be customized to examine more diverse choice scenarios specific to population segments. Also, a comparison of model results on weekdays and weekends or in different years would help to explore the change of preferences as well as check the stability of g-AMXL model. Moreover, information for economic interpretation obtained by the agent-level coefficients, such as elasticity, marginal rate of substitution, and change of social welfare, should be comprehensively analyzed. Future research should look at further collaboration with synthetic data providers and government departments to develop more use cases for the equity-based decision support tool.

8. Summary of research outputs and tech transfer

As an outcome of this research project, several research outputs were produced along with dissemination. This section summarizes those results.

Table 8.1. Summary of research outputs

Output type	Description	Link/source
Paper	AMXL for NYS	Not available yet
Paper	Service region design based on NYS AMXL	Not available yet
Paper	Ren, X., & Chow, J. Y. (2022). A random-utility-consistent machine learning method to estimate agents' joint activity scheduling choice from a ubiquitous data set. Transportation Research Part B: Methodological, 166, 396-418.	https://doi.org/10.1016/j.trb.2022.11.005
Data	Block-group level predicted mode share for NYS	https://zenodo.org/record/7718935#.ZAuf8BXMJPZ
Data	Calibrated block-group level coefficients for NYS mode choice	https://zenodo.org/record/7718923#.ZAueZBXMJPZ
Tool	AMXL and g-AMXL model for discrete choice modeling with ubiquitous dataset	Not available yet
Tool	Equity-based decision support tool for statewide mobility service design	Not available yet
Presentation	102nd TRB Annual Meeting: Agent-based mixed logit model (AMXL) for discrete choice analysis with ubiquitous data set	Paper submitted
Presentation	2022 TRISTAN Conference	Abstract submitted
Presentation	2023 Tongji Event: Agent-based mixed logit model (AMXL) for discrete choice analysis with ubiquitous data set	https://c2smart.engineering.nyu.edu/event/2nd-nyu-tju-urban-transportation-forum/

References

Agrawal, S., Avadhanula, V., Goyal, V. and Zeevi, A., 2019. MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5), pp.1453–1485.

Ahas, R., Silm, S., Saluveer, E. and Järv, O., 2009. Modelling Home and Work Locations of Populations Using Passive Mobile Positioning Data. *Lecture Notes in Geoinformation and Cartography*, 0(199079), pp.301–315.

Ahrens, A. and Lyons, S., 2021. Do rising rents lead to longer commutes? A gravity model of commuting flows in Ireland. *Urban Studies*, 58(2), pp.264–279.

Ahuja, R.K. and Orlin, J.B., 2001. Inverse Optimization. *Operations Research*, 49(5), pp.771–783.

Anda, C., Medina, S.A.O. and Axhausen, K.W., 2021. Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 128, p.103118.

Barajas, J.M., Natekal, A. and Abrams, C., 2022. *An Assessment of how State and Regional Transportation Agencies Advance Equity in Transportation Plans, Processes, and Implementation*. University of California, Davis. Institute of Transportation Studies.

Batta, R., Lejeune, M. and Prasad, S., 2014. Public facility location using dispersion, population, and equity criteria. *European Journal of Operational Research*, 234(3), pp.819–829.

Beamon, B.M. and Balcik, B., 2008. Performance measurement in humanitarian relief chains. *International Journal of Public Sector Management*, 21(1), pp.4–25.

Berry, S., Levinsohn, J. and Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pp.841–890.

Bertsimas, D., Gupta, S. and Lulli, G., 2014. Dynamic resource allocation: A flexible and tractable modeling framework. *European Journal of Operational Research*, 236(1), pp.14–26.

Bowman, J.L. and Ben-Akiva, M.E., 2001. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1), pp.1–28.

Burton, D. and Toint, P.L., 1992. On an instance of the inverse shortest paths problem. *Mathematical Programming 1992 53:1*, 53(1), pp.45–61.

Chan, F.T.S., Chung, S.-H. and Wadhwa, S., 2004. A heuristic methodology for order distribution in a demand driven collaborative supply chain. *International Journal of Production Research*, 42(1), pp.1–19.

Chow, J.Y.J. and Djavadian, S., 2015. Activity-based Market Equilibrium for Capacitated Multimodal Transport Systems. *Transportation Research Procedia*, 7, pp.2–23.

Chow, J.Y.J., Ozbay, K., He, Y., Zhou, J., Lee, M., Wang, D. and Sha, D., 2020. Multi-agent simulation-based virtual test bed ecosystem: MATSim-NYC.

Chow, J.Y.J. and Recker, W.W., 2012. Inverse optimization with endogenous arrival time constraints to calibrate the household activity pattern problem. *Transportation Research Part B: Methodological*, 46(3), pp.463–479.

Delle Site, P., de Palma, A. and Kilani, K., 2022. Consumers' welfare and compensating variation: survey and mode choice application.

Ding, C., Wang, D., Liu, C., Zhang, Y. and Yang, J., 2017. Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. *Transportation Research Part A: Policy and Practice*, 100, pp.65–80.

Durán-Heras, A., García-Gutiérrez, I. and Castilla-Alcalá, G., 2018. Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method. *Computers, Environment and Urban Systems*, 68, pp.78–88.

Esztergár-Kiss, D., Rózsa, Z. and Tettamanti, T., 2020. An activity chain optimization method with comparison of test cases for different transportation modes. *Transportmetrica A: transport science*, 16(2), pp.293–315.

Ettema, D., Bastin, F., Polak, J. and Ashiru, O., 2007. Modelling the joint choice of activity timing and duration. *Transportation Research Part A: Policy and Practice*, 41(9), pp.827–841.

Garrido, S., Borysov, S.S., Pereira, F.C. and Rich, J., 2020. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120, p.102787.

Ghobadi, K. and Mahmoudzadeh, H., 2021. Inferring linear feasible regions using inverse optimization. *European Journal of Operational Research*, 290(3), pp.829–843.

Hagenauer, J. and Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, pp.273–282.

He, B.Y., Zhou, J., Ma, Z., Chow, J.Y.J. and Ozbay, K., 2020. Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population. *Transportation Research Part A: Policy and Practice*, 141, pp.444–467.

Hörl, S. and Balac, M., 2021. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, 130, p.103291.

Iraj, E.H. and Terekhov, D., 2021. Comparing Inverse Optimization and Machine Learning Methods for Imputing a Convex Objective Function.

Joubert, J.W. and De Waal, A., 2020. Activity-based travel demand generation using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, 120, p.102804.

Karner, A., London, J., Rowangould, D. and Manaugh, K., 2020. From transportation equity to transportation justice: within, through, and beyond the state. *Journal of planning literature*, 35(4), pp.440–459.

Karsu, Ö. and Morton, A., 2015. Inequity averse optimization in operational research. *European journal of operational research*, 245(2), pp.343–359.

Kozanidis, G., 2009. Solving the linear multiple choice knapsack problem with two objectives: profit and equity. *Computational Optimization & Applications*, 43(2).

Lecun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature 2015 521:7553*, 521(7553), pp.436–444.

Lee, J.H., Davis, A., McBride, E. and Goulias, K.G., 2017. *Exploring social media data for travel demand analysis: A comparison of Twitter, household travel survey, and synthetic population data in California*.

Lejeune, M.A. and Prasad, S.Y., 2013. Effectiveness–equity models for facility location problems on tree networks. *Networks*, 62(4), pp.243–254.

Liao, Q. and Poggio, T., 2018. When Is Handcrafting Not a Curse?

McLay, L.A. and Mayorga, M.E., 2013. A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing & Service Operations Management*, 15(2), pp.205–220.

Namazi-Rad, M.-R., Tanton, R., Steel, D., Mokhtarian, P. and Das, S., 2017. An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data. *Computers, Environment and Urban Systems*, 63, pp.3–14.

New York State Department of Transportation, 2023. *Transportation Equity Act for the 21st Century*. [online] Available at: <https://www.dot.ny.gov/programs/tea21/what-tea> [Accessed 5 February 2023].

Nurul Habib, K., 2018. A comprehensive utility-based system of activity-travel scheduling options modelling (CUSTOM) for worker's daily activity scheduling processes. *Transportmetrica A: Transport Science*, 14(4), pp.292–315.

NYSERDA, 2021. *Disadvantaged Communities*. [online] Available at: <https://www.nyserda.ny.gov/ny/disadvantaged-communities> [Accessed 5 February 2023].

Ogryczak, W., Wierzbicki, A. and Milewski, M., 2008. A multi-criteria approach to fair and efficient bandwidth allocation. *Omega*, 36(3), pp.451–463.

Ohsawa, Y., Ozaki, N. and Plastria, F., 2008. Equity-efficiency bicriteria location with squared Euclidean distances. *Operations research*, 56(1), pp.79–87.

Omrani, H., 2015. Predicting Travel Mode of Individuals by Machine Learning. *Transportation Research Procedia*, 10, pp.840–849.

Pacheco, M.P., Bierlaire, M., Gendron, B. and Sharif Azadeh, S., 2021. Integrating advanced discrete choice models in mixed integer linear optimization. *Transportation Research Part B: Methodological*, 146, pp.26–49.

Perugia, A., Moccia, L., Cordeau, J.-F. and Laporte, G., 2011. Designing a home-to-work bus service in a metropolitan area. *Transportation Research Part B: Methodological*, 45(10), pp.1710–1726.

Pulugurta, S., Arun, A. and Errampalli, M., 2013. Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model. *Procedia - Social and Behavioral Sciences*, 104, pp.583–592.

Ramos, T.R.P. and Oliveira, R.C., 2011. Delimitation of service areas in reverse logistics networks with multiple depots. *Journal of the Operational Research Society*, 62(7), pp.1198–1210.

Ren, X. and Chow, J.Y.J., 2022. A random-utility-consistent machine learning method to estimate agents' joint activity scheduling choice from a ubiquitous data set. *Transportation Research Part B: Methodological*, 166, pp.396–418.

Replica Inc., 2020. *Replica Methodology*. [online] Available at: <https://www.sacog.org/sites/main/files/file-attachments/replica_methodology_2020.pdf?1602683559> [Accessed 5 February 2023].

Replica Inc., 2023. *Synthetic Population Demo*. [online] Available at: <https://replicahq.com/> [Accessed 5 February 2023].

Saadi, I., Farooq, B., Mustafa, A., Teller, J. and Cools, M., 2018. An efficient hierarchical model for multi-source information fusion. *Expert Systems with Applications*, 110, pp.352–362.

Smith, H.K., Harper, P.R. and Potts, C.N., 2013. Bicriteria efficiency/equity hierarchical location models for public service application. *Journal of the Operational Research Society*, 64(4), pp.500–512.

Sun, J., Guo, J., Wu, X., Zhu, Q., Wu, D., Xian, K. and Zhou, X., 2019. Analyzing the Impact of Traffic Congestion Mitigation: From an Explainable Neural Network Learning Framework to Marginal Effect Analyses. *Sensors 2019, Vol. 19, Page 2254*, 19(10), p.2254.

Train, K.E., 2009. *Discrete choice methods with simulation*. Cambridge university press.

Turkcan, A., Zeng, B., Muthuraman, K. and Lawley, M., 2011. Sequential clinical scheduling with service criteria. *European Journal of Operational Research*, 214(3), pp.780–795.

U.S. DOT Volpe Center, 2022. *Transportation for Social Equity*. [online] Available at: <https://explore.dot.gov/views/TransportSECensusMetricsDashboard/TransportSE?%3AisGuestRedirectFromVizportal=y&%3Aembed=y> [Accessed 5 February 2023].

United States Census Bureau, 2020. *2020 Urban Area FAQs*. [online] Available at: <https://www2.census.gov/geo/pdfs/reference/ua/2020_Urban_Areas_FAQs.pdf> [Accessed 10 February 2023].

Västberg, O.B., Karlström, A., Jonsson, D. and Sundberg, M., 2020. A dynamic discrete choice activity-based travel demand model. *Transportation science*, 54(1), pp.21–41.

Wang, R., 2012. Capacitated assortment and price optimization under the multinomial logit model. *Operations Research Letters*, 40(6), pp.492–497.

Wang, S., Mo, B. and Zhao, J., 2020. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112, pp.234–251.

Wang, S., Wang, Q. and Zhao, J., 2020. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118, p.102701.

Wu, D., Yin, Y., Lawphongpanich, S. and Yang, H., 2012. Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transportation Research Part B: Methodological*, [online] 46(9), pp.1273–1287. https://doi.org/https://doi.org/10.1016/j.trb.2012.05.004.

Xu, S.J., Nourinejad, M., Lai, X. and Chow, J.Y.J., 2018. Network Learning via Multiagent Inverse Transportation Problems. *Transportation Science*, 52(6), pp.1347–1364.

Yoon, S.Y., Deutsch, K., Chen, Y. and Goulias, K.G., 2012. Feasibility of using time–space prism to represent available opportunities and choice sets for destination choice models in the context of dynamic urban environments. *Transportation*, 39, pp.807–823.

Zhang, H.M. and Shen, W., 2010. Access control policies without inside queues: Their properties and public policy implications. *Transportation Research Part B: Methodological*, 44(8–9), pp.1132–1147.

Zhao, Y., Pawlak, J. and Polak, J.W., 2018. Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices. *Transportation Planning and Technology*, 41(1), pp.58–79.

Appendix A. Details of Replica's data

Complete field definitions of the Replica's Population and Trip table are listed in Table A.1 and Table A.2.

Table A.1. Complete field definitions of the Population table

Field name	Data type	Sample value	Description
person_id	String	144080185795050000-01	Unique identifier of a person.
household_id	String	141025892089804000	Unique identifier of a household.
BLOCKGROUP	String	360810201001	The US Census Bureau-assigned FIPS code of the census block-group containing the housing unit.
BLOCKGROUP_work	String	360810202321	The US Census Bureau-assigned FIPS code of the census block-group containing the person's workplace. (For employed persons only.)
BLOCKGROUP_school	String	360650063001	The US Census Bureau-assigned FIPS code of the census block-group containing the person's school. (For students only.)
age_group	String	18_24	Age range a person falls within. Value ranges include: • lte_4 • 5_14 • 15_17 • 18_24 • 25_34 • 35_64 • 65_plus
age	Integer	19	Age, in years old, assigned to the person.
sex	String	M	Sex assigned to the person, male (M) or female (F).
race	String	white	Race assigned to the person. Valid values include: • american_indian_alaska_native

			• asian • black_african_american • hawaiin_pacific • other_race_alone • two_or_more_races • white
ethnicity	String	not_hispanic_or_latino	Ethnicity of a person. Valid values include: • not_hispanic_or_latino • hispanic_or_latino
individual_income	Integer	7,500	Total annual income of a person.
employment	String	employed	Employment status of a person. Valid values include: • employed • notinlf • under_16 • unemployed
education	String	some_college	Education level of a person. Valid values include: • advanced_degree • bachelors_degree • high_school • k_12 • no_school • some_college • under_3
school_grade_attending	String	not_attending_school	Current grade level assigned to a person. Valid values include: • graduate • kindergarten • not_attending_school • school • undergraduate
industry	String	naics51	Industry sector a person works within, reported in NAICS code. For a list of NAICS codes see: https://www.naics.com/search-naics-codes-by-industry.
household_role	String	child	The role of the person in the household. Valid values include:

			• child • head_of_household • immobile • non_relative • relative • spouse
commute_mode	String	driving	Typical commute mode used by a person. Valid values include: • biking • carpool • driving • not_working • transit • walking • worked_from_home
tenure	String	owner	Tenure of household, owner or renter.
migration	String	same_house	Indicates mobility status (if the household has lived in the same unit 1 year ago). Values include: • N/A (Lived in house <1 year; Also applies to group quarters) • same_house (non-movers) • outside_us • within_us
household_income	Integer	408,500	Total income of the household per year.
household_size	String	3_person	Number of persons that makeup the household. Valid values include: • 1_person • 1_person_group_quarters • 2_person • 3_person • 4_person • 5_person • 6_person • 7_plus_person
family_structure	String	married_couple	Household family structure. Valid values include: • family_single

			• living_alone • married_couple • nonfamily_single
vehicles	String	3_plus	Number of vehicles owned by a household. Valid values include: • zero • 1 • 2 • 3_plus
language	String	english	Indicates the household language. Values include: • english • spanish • indo-european • asian-pacific • other

Table A.2. Complete field definitions of the Trip table

Field name	Data type	Sample value	Description
activity_id	String	15323941267251300000	A randomly assigned unique identifier defined for each trip.
person_id	String	144080185795050000-01	Unique identifier of a person.
household_id	String	141025892089804000	Unique identifier of a household.
mode	String	PUBLIC_TRANSIT	Primary transportation mode used for the trip. In the case of multiple travel modes, only the primary mode of travel across a set of trip segments is included. Valid options are: • PRIVATE_AUTO: Trips made by drivers in private auto vehicles • PUBLIC_TRANSIT: Trips that primarily used public transit, such as buses, light rail, and subways

			• ON_DEMAND_AUTO: Trips made by passengers in a Taxi or using a Transportation Network Company (TNC) such as Uber or Lyft • BIKING: Trips made by people biking. Replica does not model scooter trips and does not separate out e-bike trips • WALKING: Trips made by people walking • CARPOOL: Trips made by passengers in private auto vehicles. Sum Carpool and Private Auto trips to get the total number of people who traveled in private autos • COMMERCIAL: Trips made by medium and heavy trucks
travel_purpose	String	WORK	The destination activity assigned to a synthetic person. Valid options are: • WORK • SOCIAL • SHOP • SCHOOL • HOME • COMMERCIAL • EAT • LODGING • OTHER
tour_type	String	COMMUTE	Valid options are: • WORK_BASED • COMMUTE • OTHER_HOME_BASED
start_time	TIME	2019-01-10 06:08:00 America/New_York_City	Date and 24-hour time of trip start, reported as yyyy-mm-dd hh:mm:ss timezone
end_time	TIME	2019-01-10 07:11:04 America/New_York_City	Date and 24-hour time of trip end, reported as yyyy-mm-dd hh:mm:ss timezone

duration_minutes	Integer	63	Duration of trip in minutes, calculated as the difference between the trip start_time and end_time.
distance_miles	Float	10.79	Distance in miles measured along the trip route.
origin_bgrp	String	360810201001	The US Census Bureau-assigned FIPS code of the block group from which the trip originated.
destination_bgrp	String	360810238330	US Census Bureau-assigned FIPS of the block group in which the trip ended.
network_link_ids	Set	[28492853,28493983,…]	A set of road segment ID that the trip is associated with.
origin_land_use_l1	String	residential	The primary land use category of the trip origin. Valid options are: • residential • commercial • mixed_use • industrial • civic_institutional • transportation_utilities • open_space • agriculture • other
destination_land_use_l1	String	commercial	The primary land use category of the trip destination. Valid options are: • residential • commercial • mixed_use • industrial • civic_institutional • transportation_utilities • open_space • agriculture • other

Appendix B. Selecting the number of latent class (K)

Figure B.1 shows the results of the g-AMXL model when we set the value of K from 2 to 5 in Algorithm 3.1.

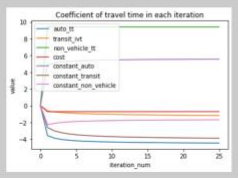
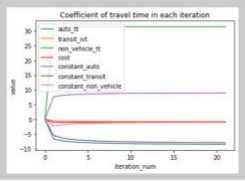
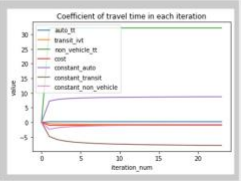
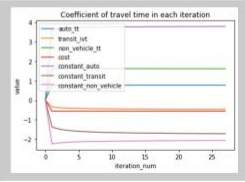
	K=2	K=3	K=4	K=5
Computational time	2.79 hours	2.47 hours	2.65 hours	2.97 hours
McFadden Rho-Square	0.599	0.598	0.598	0.601
Mean absolute error (%)	3.29%	3.34%	3.36%	3.18%
More model details	percent of good agents: 0.721 percent of fixed-point agents: 0.197 percent of zero agents: 0.083 percent of cluster 0 in good agents: 0.751 percent of cluster 1 in good agents: 0.249 cluster 0 VOT: 6.31$/hour, number of good agents 65398 cluster 1 VOT: -39.24$/hour, number of good agents 21637	percent of good agents: 0.721 percent of fixed-point agents: 0.195 percent of zero agents: 0.085 percent of cluster 0 in good agents: 0.245 percent of cluster 1 in good agents: 0.527 percent of cluster 2 in good agents: 0.228 cluster 0 VOT: 8.31$/hour, number of good agents 21342 cluster 1 VOT: 1.94$/hour, number of good agents 45860 cluster 2 VOT: -39.43$/hour, number of good agents 19833	percent of good agents: 0.721 percent of fixed-point agents: 0.198 percent of zero agents: 0.081 percent of cluster 0 in good agents: 0.515 percent of cluster 1 in good agents: 0.221 percent of cluster 2 in good agents: 0.219 percent of cluster 3 in good agents: 0.045 cluster 0 VOT: -1.56$/hour, number of good agents 44811 cluster 1 VOT: -39.04$/hour, number of good agents 19251 cluster 2 VOT: -0.21$/hour, number of good agents 19070 cluster 3 VOT: 35.77$/hour, number of good agents 3983	percent of cluster 0 in good agents: 0.501 percent of cluster 1 in good agents: 0.217 percent of cluster 2 in good agents: 0.198 percent of cluster 3 in good agents: 0.039 percent of cluster 4 in good agents: 0.044 cluster 0 VOT: -1.38$/hour, number of good agents 43610 cluster 1 VOT: -40.02$/hour, number of good agents 18901 cluster 2 VOT: -0.27$/hour, number of good agents 17276 cluster 3 VOT: -7.22$/hour, number of good agents 3405 cluster 4 VOT: 36.35$/hour, number of good agents 3843
Plot of θ_0^n				

Figure B.1. Performance of g-AMXL given K from 2 to 5