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

Urban Microtransit Cross-Sectional Study for Service Portfolio Design

USDOT Award No. 69A3551747124

August 2021

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
ORC-ID: 0000-0002-6471-3419

Co-PI: Rae Zimmerman
New York University
ORC-ID: 0000-0001-5825-3383

Srushti Rath
New York University
ORC-ID: 0000-0002-7603-339X

Bingqing Liu
New York University
ORC-ID: 0000-0002-7808-4967

Gyugeun Yoon
New York University
ORC-ID: 0000-0003-1622-9021

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 Via Transportation is gratefully acknowledged.

Executive Summary

Due to transportation technologies having such heterogeneous impacts on different communities, there needs to be better tools to evaluate the deployment of emerging technologies with limited data. Microtransit is one such technology. We propose a novel methodology to "upscale" the limited data available so that further decision-support analysis and modeling can be achieved by microtransit companies working with cities around the U.S. where none existed previously. The methodology involves simulating data using a calibrated day-to-day adjustment process for a set of cities in which data are available.

The day-to-day adjustment process simulates both first/last mile access trips and direct trips with the adjustments made to match occupancy data. A within-day microtransit simulator developed for the Federal Transit Administration is enhanced to be more parametric in design to be calibrated to different cities. A scenario generation process is developed to come up with the scenarios from which the data are generated.

The method is tested in a case study in collaboration with Via Transportation based on data they shared for Salt Lake City, Austin, Cupertino, Sacramento, and Columbus, as well as publicly available data from Jersey City. For those cities, public data is collected to estimate mode choice models that include Auto, Bike, Transit, Microtransit, and Walk for each city. Public data include U.S. Census Transportation Planning Products, American Community Survey, Transitfeeds, Smart Location Database from EPA, OpenStreetMap, and OpenTripPlanner. The models are estimated initially using maximum likelihood without the Microtransit mode since the data do not include it; afterwards, the Microtransit alternative specific constant is updated to minimize least squares from ridership data shared by Via. The average microtransit ridership error over the 6 cities with the estimated constant is 0.004 while the error with a constant of 0 is 603.59, showing that the method fits quite well.

The within-day simulation parameters are also adjusted in terms of walking limit (0.5 – 0.1 miles), dwell time (15 – 5 sec), and the weight placed on operator cost over user cost (0.8 – 0.2). This is done for the 6 cities resulting in an average ridership error of the output simulations to be 18.4%, which is acceptable. Different cities end up with different calibrated parameters. For example, Salt Lake City commuters have a walking limit of 0.5 mile, dwell time of 15 sec, and operator weight of 0.5 while Sacramento has walk limit of 0.1 mile, dwell time 5 sec, and operator weight 0.2.

The simulator is then used to synthesize 326 scenarios. The scenarios are randomly generated from four of the cities: Salt Lake City has 71 scenarios, Austin has 79, Sacramento has 100, and Cupertino has 76. For these scenarios, a forecast model is estimated. We use a linear regression with employment density, household density, mean household income, street density, transit station density, ratio of households with one or more automobiles, a trip equilibrium index, the underlying pricing policy from

Via (which could be PP1 (fixed fare and paid first/last mile rides) or PP2 (fixed fare with free first/last mile rides)). First order features are also used and selected via Lasso regularization. The resulting ridership model has a R^2 of 0.72 from the data based on 47 features. The CV of the model applied to the 4 original cities' data is 45%. Meanwhile, the VMT model has R^2 of 0.90 based on 55 features, and a Coefficient of Variation (CV) of 37%. Considering the comparison is to only four cities' output data as a proof of concept, these results look like an adequate fit. Further refinements in the future can be made by identifying city clusters from which different models can be estimated.

The forecast models are then used to identify two other alternative portfolios that have the same predicted total VMT as the 4 cities currently being operated, but their ridership can be expanded 1.4 to 1.9 times higher. For example, the first example portfolio consists of service regions in Seattle, Chicago, Boston, and Birmingham. The second portfolio includes Seattle, Boston, Detroit, St. Louis, D.C., and Arlington. The case study proves that the day-to-day adjustment model can be made to fit to limited data, and further that it can help reveal important relationships between public data and measures like ridership and vehicle-miles-traveled. Such results can be used by microtransit companies to identify cities to reach out to and to provide quantitative support to convince them of the potential value of the service; it can also be used by federal agencies like the Federal Transit Administration to target priority areas for supporting microtransit deployment.

Improvements can be made to the scenario generation process by identifying city typologies and customizing models for those types. We find that certain cities like Columbus just behave very differently from the other cities. Understanding a city's transportation typology is immensely valuable for planners and policy makers whose decisions can potentially impact millions of city residents. Despite the value of understanding a city's typology, labeled data (city and its typology) is scarce, and spans at most a few hundred cities in the current transportation literature. We propose a supervised machine learning approach to predict a city's typology given the information in its Wikipedia page. Our method leverages recent breakthroughs in natural language processing, namely sentence-BERT, and shows how text-based information from public sites like Wikipedia can be effectively used as a data source for city typology prediction tasks that can be applied to over 2000 cities worldwide. The method makes supervised learning of city typology labels (such as congestion, auto-heavy, transit-heavy, and bike-friendly cities) tractable even with a few hundred labeled samples. Based on data from 197 cities for training and 85 cities as a test set, we show that model for predicting whether a city is congestion-based, auto-heavy, transit-heavy, or bike-friendly just from Wikipedia data alone has accuracies of 080, 0.85, 0.62, and 0.76, respectively. The model is then used to classify 2100 cities around the world, which significantly expands the visibility to what can be evaluated.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Project Background

Transportation technologies are not “one-size-fits-all” solutions in general because their effectiveness depend on the deployment region. On-demand transit, i.e., “microtransit”, exhibits this characteristic. Microtransit can be defined as shared public or private sector transportation services that offer fixed or dynamically allocated routes and schedules in a demand-responsive manner i.e., in response to individual or aggregate consumer demand, using smaller vehicles (multi-passenger /pooled shuttles or vans) and capitalizing on widespread mobile GPS and internet connectivity (Volinski, 2019; Chow, et al., 2020; Yoon, et al., 2021). The broader market of demand-responsive transportation (e.g., shared taxis, ridesourcing, carshare, micromobility, microtransit) has gained significant interest in the global urban mobility sector because of these mobile technologies.

Since these technologies are not one-size-fits-all, the reception for such technologies have been mixed. Some ventures have been successful. For example, Via Transportation, Inc. (founded in 2012) (Via, 2021) continues to operate at full capacity in over 35 countries in partnership with over 90 transit agencies (see Figure 1.1(a)). Their services include door-to-door, first-last mile trips to transit stations, and virtual stops. Transdev (2021), founded in 2011, operates multiple microtransit services (including first-last mile services) in the U.S, the Netherlands, France, and Australia. On the other hand, there have been failures as well: Kutsuplus in Helsinki (Haglund, et al., 2019), Car2Go in North America (Krok, 2016), Bridj (Bliss, 2017), and Chariot (Marshall, 2019). Usability and adoption of such services vary from city to city in terms of cost and benefit. Currie and Fournier (2020) provide a lifespan analysis on 120 demand-responsive transportation systems (including microtransit) from 19 countries over the period 1970-2019; their analysis highlights the failure rates in the UK is 67% while that in Europe and the USA/Canada is 23% and 50%, respectively.

(a) (b)

The list of cities shown in Figure 1.1(a) represent an example of a "microtransit deployment portfolio". A portfolio consists of a list of active product projects sharing common resources that is continuously updated; new projects need to be evaluated and prioritized and existing projects may be accelerated, abandoned, or de-prioritized (Cooper, et al., 1998; Chow, et al., 2011). With microtransit deployment as the portfolio product, how can mobility providers decide which city agencies to work with for deploying new service?

To address this microtransit deployment portfolio problem, a solution is needed that can make the most of the limited data that may be available. We propose a new methodology to "upscale" the available data using simulation, in a similar manner to how deep learning algorithms can be used to upscale low-quality images into high-quality ones.

This topic of scenario generation has also been applied to generating test cases for machine learning models, particularly in testing autonomous vehicle algorithms (Rocklage, et al., 2017; Tuncali, et al., 2018; Nalic, et al., 2020). Our methodology is novel in that the market equilibrium model is extended from earlier work (Chow, et al., 2020; Djavadian & Chow, 2017a; Djavadian & Chow, 2017b; Caros & Chow, 2021) to allow parameterizing degree of virtual stop access distance and outputting the degree of usage of microtransit as a first/last mile access mode. Analysis of the generated scenario data reveals interesting insights relating a deployment's ridership and vehicle-miles-traveled (VMT) to service region design, pricing policy, and proximity of fixed route transit stations. Furthermore, the method can be readily adapted to any emerging transportation technology deployment planning process.

1.2 Role of Multimodal Connectivity in Improving the Performance of Ridesharing

Three factors are addressed in this research: the role of multi-modal connectivity in improving the performance of ride sharing, human behaviors affecting modal connections, and market opportunities to expand ride sharing services. Microtransit services such as Via can provide can be reinforced by connecting to other modes and reducing competition among modes of travel. To reinforce positive performance, a multimodal framework is portrayed here that interfaces with Via usage applicable to some of the target cities (see Figure 1.2). This work covers major modes of travel, i.e., automobiles and other road-based vehicles and rail, as well as microtransit which is becoming a significant component of multimode travel, and within those broad categories are numerous subcategories with varying characteristics (Litman 2021a). Microtransit is defined as "lightweight, single-person vehicles" (U.S. DOT, BTS 2020: 3-12). The benefits of microtransit are lower environmental impacts, smaller space requirements, greater ease of operability in dense areas (INRIX 2019), simplicity of operation, and portability.

Human behaviors are identified including choice of modes of travel and routes. Behavior is a key factor in understanding multi-modal connections. Behavior covers a wide range of factors such as cost, environmental sensitivity, safety and security, and convenience.

Market opportunities are briefly identified that extend beyond passenger transport to encompass food delivery services as a future market (Zimmerman 2021b). This has been a need during the COVID-19 pandemic.

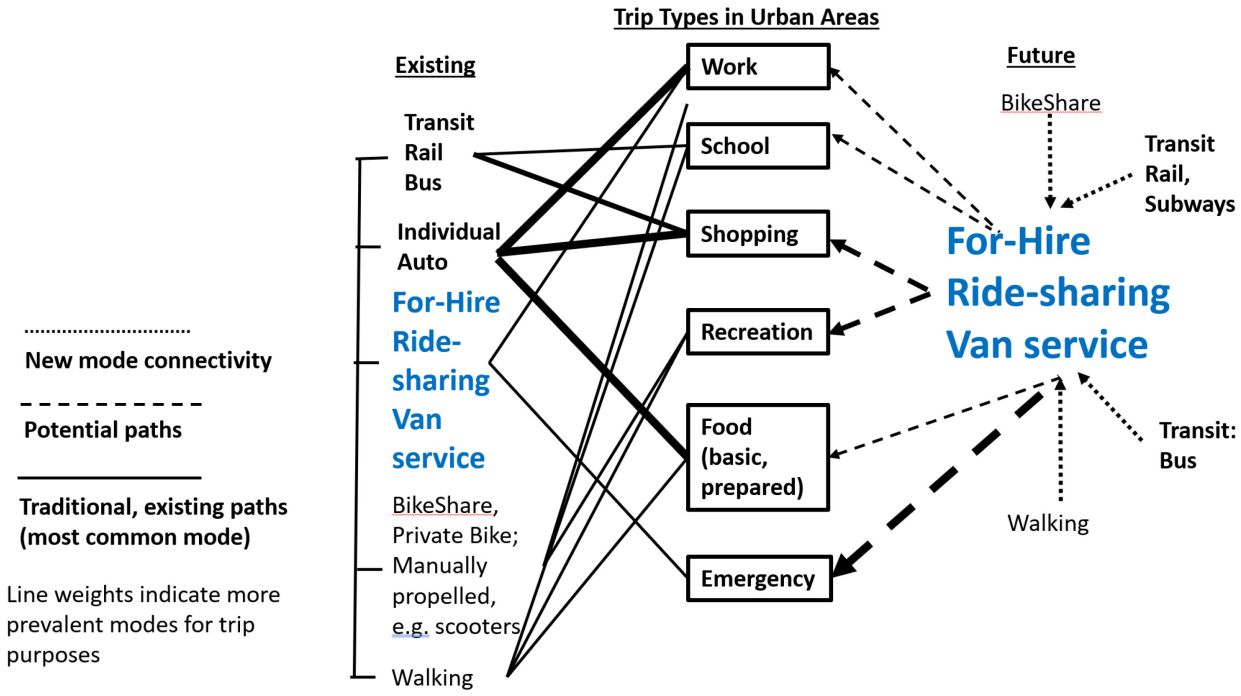


Figure 1.2. Multi-modal transportation framework: existing and potential van service examples, trip type and interconnectivity with other modes, urban areas (adapted from Chow, et al., 2021, Zimmerman, 2019b)

The project report is organized to provide an overview of the forecast models for microtransit, and simulation-based market equilibrium forecasting which includes an in-depth discussion of multimodal connectivity requirements. It is followed by a discussion on the proposed methodology including day-to-day market equilibrium model to handle both first/last mile access trips and direct trips, an updated within-day microtransit simulator with more parametric in design, and a scenario generator to get surrogate data for developing the service portfolio design model. This is followed by an in-depth case study using data shared by Via (for Salt Lake City, Austin, Cupertino, Sacramento, Columbus, and Jersey City) to illustrate how the portfolio design model can be used analyze deployment portfolios in multiple cities. Finally, a section is devoted for understanding and predicting city typologies where we propose a novel method using Wikipedia data for large-scale city classification.

- Section 2: overview of microtransit forecasting and multimodal connectivity
- Section 3: proposed methodology
- Section 4: use case study of Via microtransit deployment portfolio
- Section 5: city typology prediction using Wikipedia
- Section 6: conclusion
- Section 7: summary of research outputs and technology transfer

2. Overview of Multimodal Connectivity and Microtransit Forecasting

2.1. Multimodal Connectivity Incorporating Ride Sharing

Commuting patterns in the absence of ride sharing often emphasize or incorporate multi-modal connectivity. Such connectivity patterns will often vary by type of trip and mode (U.S. DOT, FHWA, 2018: 18; U.S. EPA, 2013: 29). The existence of multi-modal facilities that include such connectivity are key to promoting multi-modality and integrating ride sharing modes. Figure 1.2 highlights some of these relationships as both existing and potential with a focus on ride sharing modes. Figure 1.2 lists a half dozen types of trips. Existing ride sharing services are strongest for work, food, and shopping trips. In the future, ride sharing could connect with non-motorized forms of microtransit as well as conventional transit and expand into servicing more types of trips.

The potential for interconnectivity of transportation modes is reflected in the increasing number of modes, the increasing connections among them, and the existence of modal connectivity infrastructure to support connectivity. The variability of connectivity of multi-modal transportation reflects the many dimensions of multiple modes including different services, technologies often broadly categorized as motorized and non-motorized, facilities, pricing, networks, and supporting communication structures (Litman, 2021b). New modes are continuing to emerge, and the trend has been quantified nationally by NACTO (2019b: 5) for just a few of these - shared bikes, e-bikes and e-scooters: “84 million trips on shared bikes and e-scooters in the United States, more than double the number of trips taken in 2017 . . . with 38.5 million trips on shared e-scooters.”

In light of these trends NACTO has underscored the need to regulate their use to avoid conflicts (NACTO, 2019b).

The U.S. DOT BTS analysis of its Intermodal Passenger Connectivity Database (IPCD) nationally reported that as of 2019 bikeshare dominated interconnected transportation systems, followed by rail – heavy, commuter and light rail in that order (U.S. DOT BTS 2020, p. I-23). The prevalence of rail and bus connectivity has been strong especially in major urban areas such as the New York area (Zimmerman et al., 2014). Zimmerman et al. (2014) found that in NYC and its region, subway and bus transfers are among the most common. However, they depend on such connections existing. The research found that in NYC bus stops near subway stations are often sparser in poorer areas of the city. That bikeshare and scooters are engaged in multimodal connectivity to transit is indicated in a NACTO survey that found that about a quarter of scooters and two thirds of station-based bike share users connect to transit.

People are often flexible in their use of multiple modes and such usage is often situation dependent, changing over time. Immediately following the New York City World Trade Center attacks on September 11, 2001, for example, Zimmerman and Sherman (2011) found that survivors leaving the area began by using single travel modes, primarily walking, until other modes opened up and they quickly switched to multiple

modes to leave the area. The configuration or location of transfers and type of mode used often are related to whether they are for the first or last mile of travel. Micromobility modes are more commonly used for small distances, whereas road and rail-based modes are typically used for longer distances (INRIX, 2019: Figure 2).

The viability of incorporating new technologies into multi-modal connectivity including ride sharing for electric vehicles involve technological factors such as battery technology and the existence of charging technology (Yergin, 2021). Autonomous vehicles have been explored for multimodal mobility (Kortum, 2018) and will require public acceptability, street reconfigurations (Reid, 2021), and other adaptations. Multimodal connectivity also faces equity issues as some research has shown in that not all modes and their connections to other modes are distributed equally across demographic groups. Zimmerman et al. (2014) identified this for bus-subway connections across over 400 subway stations in New York City. The study by Ferenchak and Marshall (2021) of 29 cities over a ten-year period from 2010-2019 in the U.S. found inequities in the siting of bike share facilities. Multi-modal connectivity will also be affected by regulations that are put in place for some of the newer technologies (NACTO, 2019b: 14).

2.2. Human Behavior

Human behavior can influence connections among multiple modes (Zimmerman, 2019a; Litman, 2021). For example, these behaviors are often a function of time, convenience, cost and cost stability of travel, which are characteristics of car sharing explored in other sections of the report for VIA. The importance of or sensitivity to these factors to users can vary, producing variations in demand. Factors that shape human behavior applicable to infrastructure in general, including transportation, have been identified by Zimmerman (2019a) as: Safety and security, Environmental compatibility, Cost directly and indirectly as economic benefits, Affordability, Availability, Accessibility, Comfort, Convenience, Aesthetics, and Equity. Examples of behavioral factors that relate directly to transportation, including ride sharing, are summarized as follows from Zimmerman (2019a) and applicable to ride sharing. Safety and security reflect perceptions about the risk and severity of accidents for different travel modes. Environmental compatibility refers to the extent to which different modes exert pressures on the environment through emissions and disruptive use, and the values that transportation users place on those environmental attributes. Non-motorized modes are generally considered more environmentally compatible except where they physically exert pressures on the environment. Other technologies are considered to support environmental values as well, such as electric vehicles. While the technological aspects were discussed in the previous section, there are behavioral elements as well related to the extent to which people are likely to adopt electric vehicles (EVs) over fuel-based cars (Yergin, 2021). Economic benefits appear in the form of property values (Chatman and Noland 2013). Costs also encompass savings resulting from when a vehicle can be charged, e.g., time of use (Boylan, 2019). Convenience includes choice of routes, for example the research for transit by Guo (2011). Accessibility is associated with many metrics and refers to how easily one can obtain services. Proximity is one such measure that Via incorporates in terms of how far users are from the vehicle they need. Transit-on-

demand services that Via has offered in low density areas include services in Arlington, TX (Smith, 2021) and many other cities that offer proximity of pick up and drop off.

2.3. Expanded Markets for Ridesharing

An important potential area for the expansion of intermodal frameworks is food delivery services as a future market for such multimodal services. Ride sharing services combined with other modes can potentially provide passenger and product delivery alternatives to reduce cost and travel time across the food delivery supply chain (Zimmerman, 2021a, b). Many parts of the food chain depend on a combination of transportation technology and behaviors. The factors that influence the distribution portion of the food chain have been addressed and analyzed extensively and are presented and summarized by Zimmerman et al. (2016, 2018). During the pandemic, dramatic changes in transport within the food distribution system occurred that often affected food processing and packaging. For example, manufacturers often circumvented wholesalers, sending food produce directly to consumers which had considerable implications for transportation modes and their connectivity (Zimmerman, 2021b).

2.4. Forecast Models for Microtransit

In microtransit deployment portfolio management, the perspective shifts to a market of multiple cities. Forecasts need to be made for multiple different cities and for different operating modes. Conventional forecasting practices (Volinski, 2019; Chow, et al., 2020; Yoon, et al., 2021) only consider the public agency perspective, which are not applicable to the deployment portfolio planning problem. Cross-sectional models for forecasting microtransit measures across multiple cities simply do not exist.

Forecast models for individual cities are also limited, and for good reason. Analytical models tend to resort to simplified operations and homogeneous conditions (Daganzo & Ouyang, 2019) or are used for explaining ex post conditions (Haglund, et al., 2019; Pinto, et al., 2020; Pantelidis, et al., 2020; Ma, et al., 2021). Microtransit can have many dimensions of complexity: routing, dispatch, pricing, rebalancing, fleet sizing, service region coverage, etc. Four step models are not equipped to make predictions for users based on these complex factors mainly because that equilibrium cannot be easily captured in a static model that exhibits not only route and mode choice, but also transfers, wait time, and departure time choice. To overcome this drawback, city simulations draw on complex multi-agent simulations of activity behavior (Chow & Djavadian, 2015; Cich, et al., 2017). However, these tools are computationally expensive and data hungry.

One area that has limited exploration is in forecasting the role of microtransit as a first/last mile access mode (Shaheen & Chan, 2016). To date there are no forecast models that distinguish between microtransit as a direct service or as a first/last mile access mode. Yan, et al. (2019) makes forecasts of multimodal trips using ride-sourcing strictly to access public transit.

2.5. Simulation-based Market Equilibrium Forecasting

Simulation-based methods are proven to be effective for evaluating complex mobility systems (Horn, 2002; Jung & Chow, 2019; Ma, et al., 2019; Markov, et al., 2021). However, many such studies only consider fixed demand to simulate the supply side "within-day" without any equilibration.

To capture the equilibrium between demand and mobility services, day-to-day adjustment mechanisms have been used to describe a transportation system through its dynamic evolution (Smith, 1984; Watling & Hazelton, 2003). Under such mechanisms, users and operators in the system adjust their behavior according to past experiences. Such mechanisms can lead the system to evolve and converge at different states depending on the initial conditions and the behavior characteristics of the users and operators (Smith, et al., 2014). Day-to-day adjustment models have been used to model complex transportation systems because they explicitly capture the relationship between system state and the behavior of users and operators (Horowitz, 1984; Mahmassani & Chang, 1986; Mahmassani, 1990; Cantarella & Cascetta, 1995). However, these earlier studies focus only on the road traffic network.

Djavadian and Chow (2017a,b) proposed an agent-based day-to-day adjustment process of flexible transport service and showed that the sampling distribution of different agent populations reaches a stochastic user equilibrium (SUE). Users' choices of mode and departure time are adjusted from day to day to maximize utility and minimize delay. Caros and Chow (2021) extended that model to capture operator learning of optimal cost weights to anticipate elastic user demand in evaluating modular autonomous vehicles in Dubai.

Similar mechanisms are adopted in this study to model the market equilibrium of a transportation system with a microtransit subsystem.

3. Proposed Methodology

The scope of the methodology is shown in Figure 3.1. The ideal setting is that there is enough data that insights (e.g., forecast model as shown in the top dashed box) can be drawn between public data available for any U.S. city and measures important to the portfolio, e.g., aggregate daily ridership or the fleet's VMT. The problem is that data needed for such an analysis or portfolio forecast model is "low-quality"; e.g., in our study we have only data from 6 U.S. cities (of which only 4 are usable).

Our proposed methodology can be used to obtain information including degree of first/last mile access, fleet size, fleet vehicle miles traveled, average traveler journey times (wait, access, in-vehicle), operation cost, revenue, and other derivative measures. To the best of our knowledge, no other forecast methodology outputs all these measures.

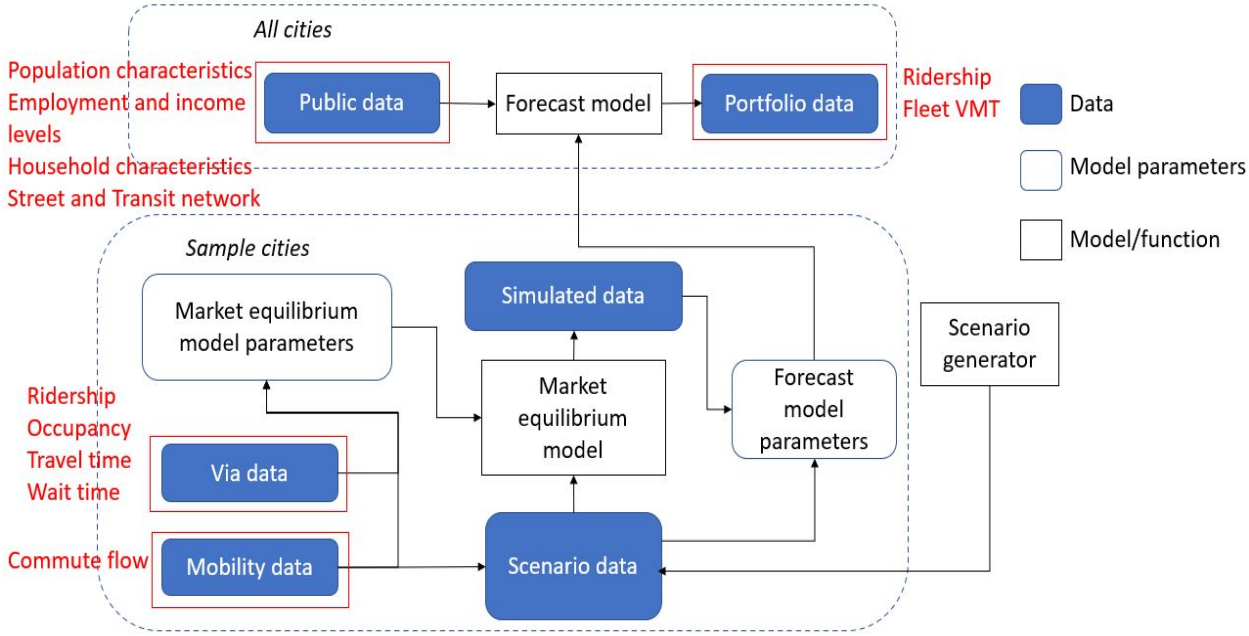


Figure 3.1. Process diagram showing modeling needed to generate scenario data for a portfolio-level forecast model

3.1. General Model Design

A day-to-day adjustment process characterizes the dynamics in adjustments made by both travelers (users) and the operators each day as a dynamic system. Djavadian and Chow (2017a,b) showed that such adjustment processes can reach a stochastic user equilibrium with an asymptotic number of sampled agent populations. The framework is shown in Figure 3.2.

At the end of the period, experienced micro-transit in-vehicle time, wait time, and walk time for each user, and average occupancy for the vehicles, are computed. These values are used to update the mode choices, departure times, and fleet sizes for the next day (note that only microtransit is simulated, so the attributes for all the other modes—Auto, Transit, Bike, Walk, Others—are fixed). The utility functions are generally specified as shown in Eq. (3.1) – (3.6) (with most statistically insignificant attributes for each city removed) as a mode choice model for a given agent n.

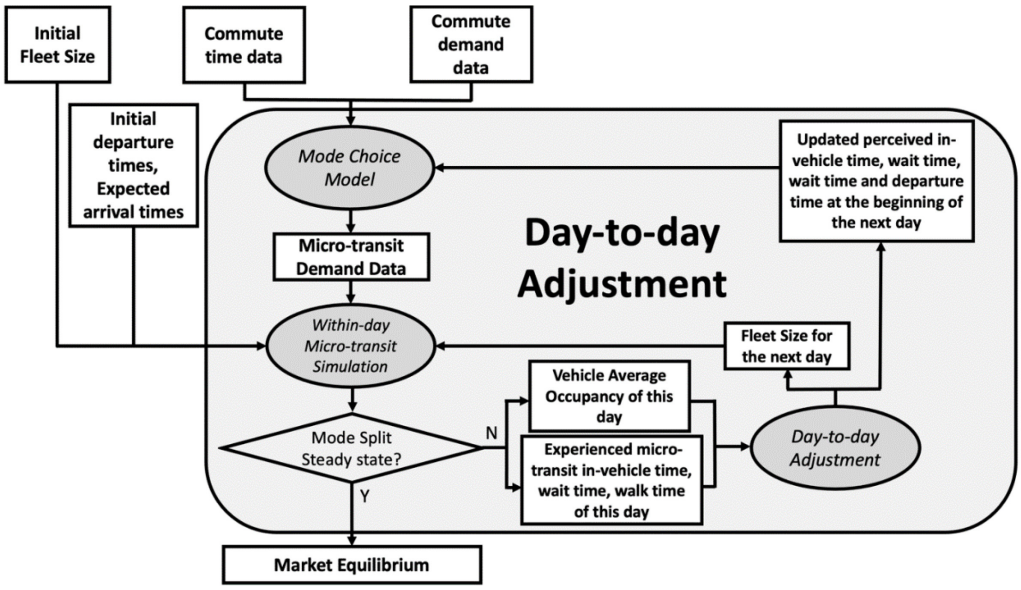


Figure 3.2. Framework of the day-to-day adjustment with oval functions, rectangles for data, and a diamond for decision

$$U_{auto,n} = asc_{auto} + \beta_{tt_{auto}} \times TT_{auto,n} + \beta_{interzone} \times Interzone_n + \varepsilon_{auto,n} \quad (3.1)$$

$$U_{transit,n} = asc_{transit} + \beta_{tt_{transit}} \times TT_{transit,n} + \beta_{AE} \times AET_{transit,n} + \beta_{wait} \times WT_{transit,n} + \beta_{cost} \times CO_{transit,n} + \beta_{interzone} \times Interzone_n + \varepsilon_{transit,n} \quad (3.2)$$

$$U_{bike,n} = asc_{bike} + \beta_{tt_{bike}} \times TT_{bike,n} + \beta_{interzone} \times Interzone_n + \varepsilon_{bike,n} \quad (3.3)$$

$$U_{walk,n} = asc_{walk} + \beta_{tt_{walk}} \times TT_{walk,n} + \beta_{interzone} \times Interzone_n + \varepsilon_{walk,n} \quad (3.4)$$

$$U_{MT,n} = asc_{MT} + \beta_{tt_{auto}} \times TT_{MT,n} + \beta_{AE} \times AET_{MT,n} + \beta_{MTwait} \times WT_{MT,n} + \beta_{cost} \times CO_{MT,n} + \varepsilon_{MT,n} \quad (3.5)$$

$$U_{others,n} = \varepsilon_{others,n} \quad (3.6)$$

where $asc_{\langle mode \rangle}$ denote the mode specific constant for $\langle mode \rangle = \{auto, transit, bike, walk, microtransit\ (MT), other\}$; $TT_{\langle mode \rangle}$ is modal travel times from origin to destination; $CO_{\langle mode \rangle}$ denote the corresponding modal travel costs; $WT_{\langle mode \rangle}$ refer to wait times for transit and microtransit modes; $AET_{\langle mode \rangle}$ are the access/ egress time for transit and microtransit; and $Interzone_n$ is a categorical variable for interzonal trips i.e., 1 when a trip's origin and destination are in different zones (census tracts) and 0 otherwise. The attributes are tracked with an index d to represent the perceived value at the start of day d. The parameters $\beta_{tt_{\langle mode \rangle}}, \beta_{AE}, \beta_{wait}, \beta_{MTwait}, \beta_{cost}, \beta_{interzone}$ need to be estimated for each city or cluster of cities.

One novel treatment of the demand model is that it includes agents from both (1) direct door-to-door trips within a designated service region S as well as (2) first/last mile trips connecting with transit stations located within the service region to other locations in the greater region $Z \supset S$. This is illustrated with Figure 3.3.

Census tracts are used as the geographic units. At the start of the simulation, the origin and destination coordinates of each user are generated randomly within their origin and destination census tracts. The adjustment of fleet size is based on average occupancy provided by the data. At the end of each day, fleet size is adjusted towards the ideal average occupancy based on the occupancy of the past day as shown in Eq. 3.7.

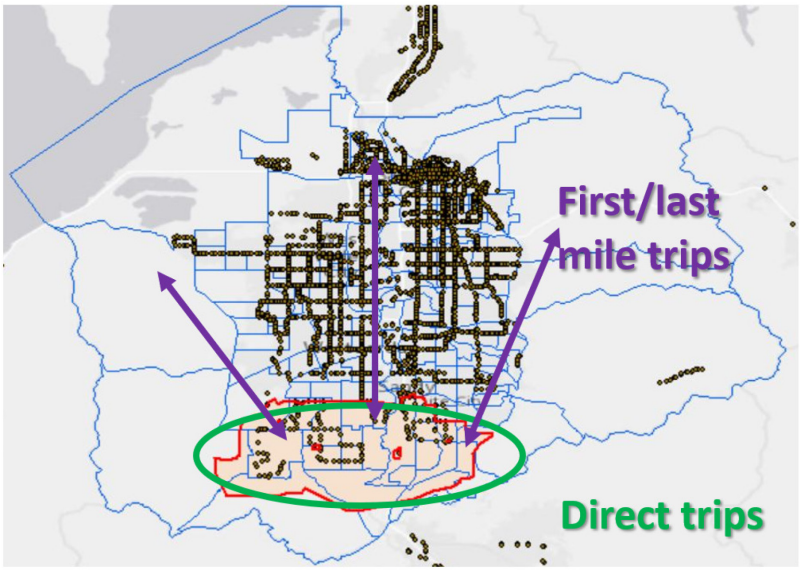


Figure 3.3. Illustration of a designated service region in red (via direct trips) along with blue-highlighted zones in the greater region accessed by public transit (via first/last mile trips)

where FS^d stands for the microtransit fleet size on simulation day d.

Microtransit in vehicle time, wait time and walk time for each user is also updated from day to day. The method adopted is similar to Djavadian and Chow (2017a,b). For a user who used microtransit on day d for commute, he/she learns from the experience on day d and update his/her perceived in-vehicle time, wait time, and walk time with a learning rate θ (shown in Eqs. (3.8) – (3.10)). Learning rate is set as 0.1 in this study.

$$TT_{MT,n}^{d+1} = (1-\theta)TT_{MT,n}^{d} + \theta\ ETT_{MT,n}^{d} \quad (3.8)$$

$$WT_{MT,n}^{d+1} = (1-\theta)WT_{MT,n}^{d} + \theta\ EWT_{MT,n}^{d} \quad (3.9)$$

$$AET_{MT,n}^{d+1} = (1-\theta)AET_{MT,n}^{d} + \theta\ EAET_{MT,n}^{d} \quad (3.10)$$

where $TT_{MT,n}^{d}$, $WT_{MT,n}^{d}$, and $AET_{MT,n}^{d}$ stand for perceived microtransit (MT) in-vehicle time, wait time, and walk time for user n at the beginning of day d. $ETT_{MT,n}^{d}$, $EWT_{MT,n}^{d}$, and $EAET_{MT,n}^{d}$ stand for experienced microtransit (MT) in-vehicle time, wait time, and walk time for user n on day d.

Having introduced the key attributes, let us adopt a generic symbol X to represent each attribute for convenience. For a user n' who did not use microtransit but used other modes on day d for commute,

his/her perceived times are updated with the population's average perceived times $\bar{X}_{MT}^{d}$ at the end of day d (shown in Eq. (3.11)).

$$X_{MT,n'}^{d+1} = (1-\theta)X_{MT,n'}^{d} + \theta\ \bar{X}_{MT}^{d} \tag{3.11}$$

Population perceived in-vehicle time, wait time, and walk time represent the overall perception of the population in the service region, which is the successive average of average in-vehicle, wait, and walk time of the past n days (shown in Eq. (3.12)).

$$\bar{X}_{MT}^{d} = \left(1-\frac{1}{d}\right)\bar{X}_{MT}^{x-1} + \frac{1}{d}E\bar{X}_{MT}^{d} \tag{3.12}$$

Departure time of each user is a continuous variable that is updated based on his/her expected arrival time. The departure time of a passenger on day $(d+1)$ is computed based on the experienced commute time of day d as shown in Eq. (3.13).

$$DT_{n}^{d+1} = AT_{n} - PT_{n}^{d} \tag{3.13}$$

where DT_{n}^{d} stands for the departure time of user n on simulation day d, AT_{n} stands for the desired arrival time of user n. PT_{n}^{d} is the perceived commute time at the end of day d for user n, which depends on the mode taken in Eq. (3.14).

$$PT_{n}^{d} = TT_{MT,n}^{d} + WT_{MT,n}^{d} + AET_{MT,n}^{d} \tag{3.14}$$

At the end of each day, we check if the system has reached a steady state. The adjustment stops when the daily microtransit ridership change keeps below 1% (shown in Eq. (3.15)) for 5 consecutive days.

$$\frac{Ridership_{MT}^{d+1} - Ridership_{MT}^{d}}{Ridership_{MT}^{d}} \leq 1\% \tag{3.15}$$

3.2. Estimation of Mode Choice Model

The choice model is estimated using a combination of microtransit operator data and publicly available data. The following estimation algorithm is used.

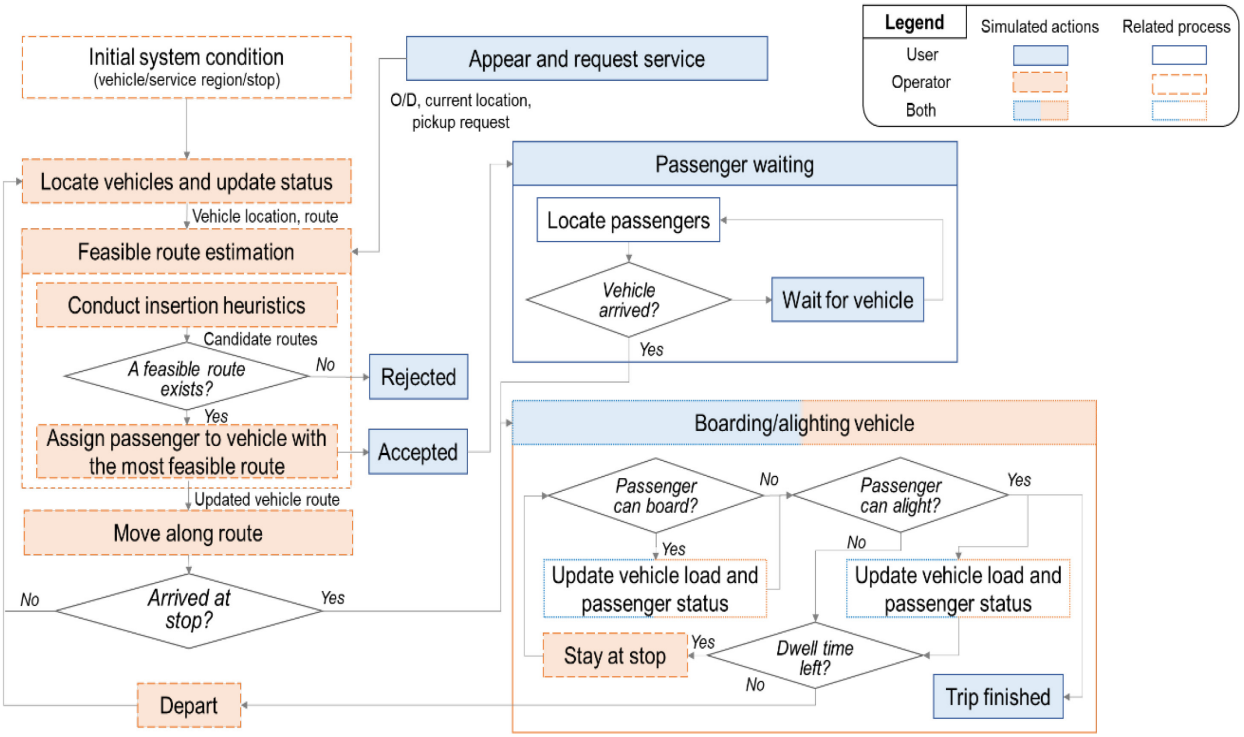
Algorithm 3.1. Mode choice estimation

1. Obtain the commute flow data for auto, transit, bike, walk, and others between census tracts within the region (CTPP, 2016) including transit flows from within service region to outside the service region (within the boundaries of the county/counties covered by the region) and vice-versa.
2. Assume some of the parameters' relationships: $\beta_{\text{MTwait}} = \beta_{\text{wait}}$ (in transit if significant, else $\beta_{\text{MTwait}} = 1.53\beta_{\text{tt}_{\text{auto}}}$), $\beta_{wait} = 1.59\beta_{tt_{transit}}$, $\beta_{AE} = 1.78\beta_{tt_{transit}}$, $\beta_{cost} = 3\beta_{tt_{transit}}$ (Wardman, 2004).
3. Two population segments are constructed: commute flows within the service region S as direct trips with access to all 6 modes in Eqs. (3.1) – (3.6); transit flows from S to $Z \supset S$ and vice versa are assigned OD flows to/from the nearest transit station, assumed to have access to walk, bike, or microtransit only.
4. Estimate Eqs. (3.1) – (3.4), (3.6) (without microtransit) using conventional maximum likelihood estimation with the added constraints in step 2.
5. Add in the microtransit utility function Eq. (3.5) and use root finding via bisection method to ensure the ridership difference (between predicted and actual) is minimized by adjusting asc_{MT}, keeping in mind that the ridership is obtained from the sum of both population segments.

3.3. Within-day Simulator

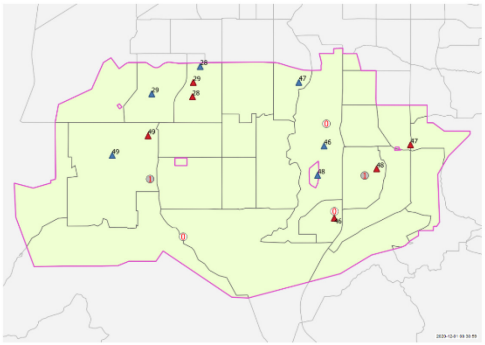
Within-day simulation of the microtransit system is the module located at the lowest level in the entire framework. Due to the limitation of the observation in the actual mobility market, this study uses simulation to estimate users' responses and system performance. The main framework of this simulation (as illustrated in Figure 3.4) is newly extended from Yoon, et al. (2021) to include:

- Virtual stops, meeting points other than actual origin and/or destination of users, and
- Feature of depot assignment, designating a depot of vehicles based on relocation cost and average wait time.



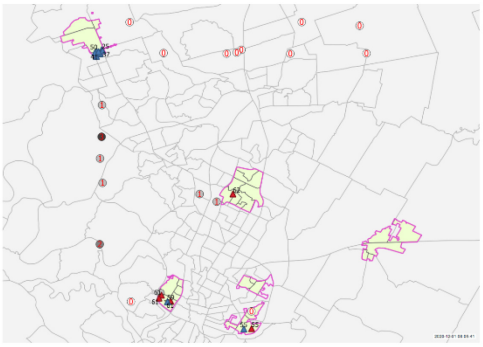
There are four categories of required inputs of this simulation as described below. Simulation parameters control the length and precision of simulations, regulating how long and often to collect generated data. Example simulations conducted for four different cities (Salt Lake City, Austin, Cupertino, and Sacramento) are shown in Figure 3.5.

- Simulation parameters: simulation length, time step length
- Scenario parameter: walking speed, maximum walking distance, average vehicle running speed, weight for passenger in-vehicle/wait/access time, value of time, unit operation cost, weight of operator cost
- System design parameter: vehicle capacity, fleet size, number of depots, average dwell time
- Dataset: passenger request information, passenger arrival data, depot locations, virtual stop locations, vehicle allocation distribution among depots



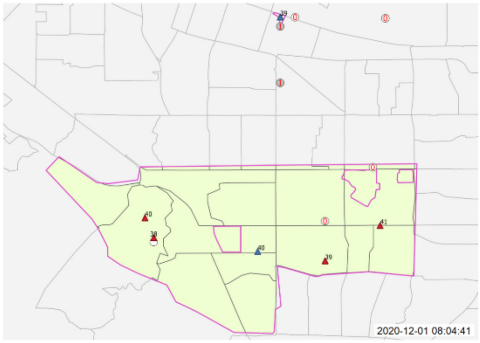
Salt Lake City

Days to convergence: **33**
Simulation time: **6-9 AM**
Fleet Size: **5**
Number of Via Passengers: **80**
Average In-vehicle Time: **10.3 min**
Average Wait Time: **10.4 min**



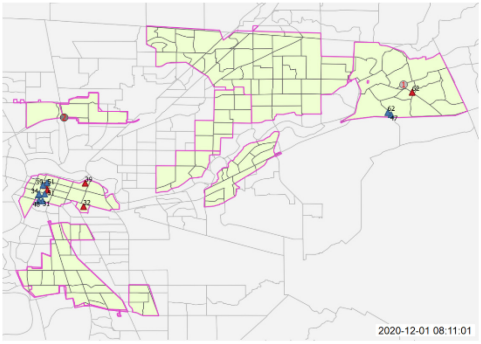
Austin

Days to convergence: **21**
Simulation time: **6-9 AM**
Fleet Size: **20**
Number of Via Passengers: **103**
Average In-vehicle Time: **22.0 min**
Average Wait Time: **8.5 min**



Cupertino

Days to convergence: **51**
Simulation time: **6-9 AM**
Fleet Size: **8**
Vehicle Capacity: **6**
Number of Via Passengers: **69**
Average In-vehicle Time: **10.9 min**
Average Wait Time: **6.1 min**
Average Walk Time: **0.1 min**



Sacramento

Days to convergence: **48**
Simulation time: **6-9 AM**
Fleet Size: **2**
Vehicle Capacity: **6**
Number of Via Passengers: **48**
Average In-vehicle Time: **9.1 min**
Average Wait Time: **32.5 min**
Average Walk Time: **1.7 min**

3.4. Scenario Generator

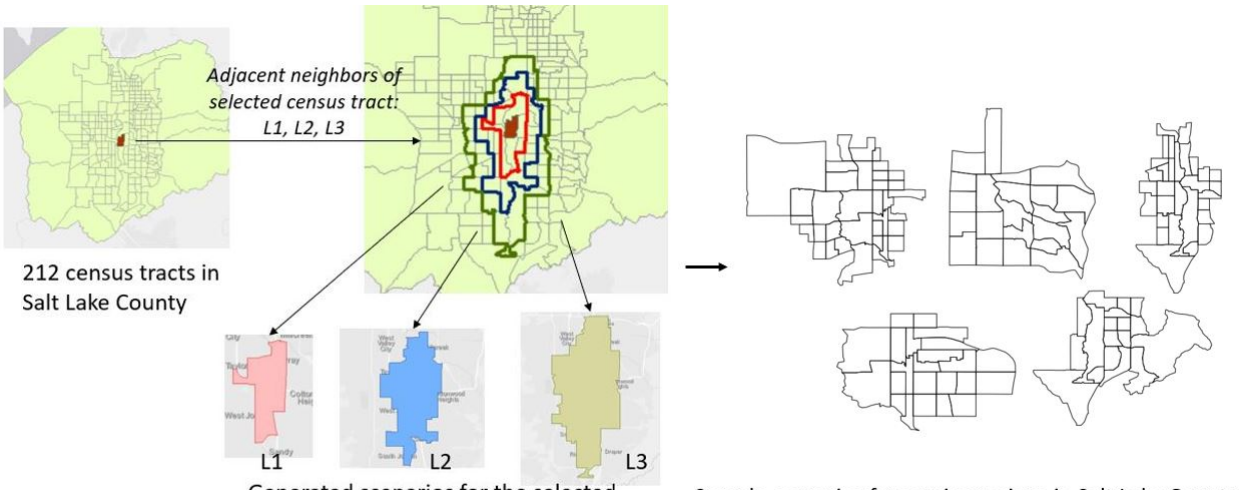
Once the simulation-based market equilibrium model is calibrated, a process is used to generate additional scenarios to upscale the existing data. In each scenario ω, one of the calibrated cities is selected and the service region S is redefined as $S(\omega)$. We define a region as constituting two or more contiguous zones where microtransit service operates both door-to-door and first-last mile service with a specific pricing policy. The scenario is generated using Algorithm 3.2.

Algorithm 3.2. Scenario generation

Given a city where a microtransit service operates (in a specific region), we generate multiple scenarios (regions) with the following steps:

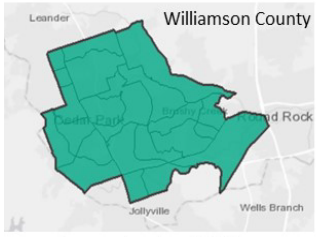
1. Obtain the list of census tracts (zones) and their boundaries for all zones within the county/counties intersecting with the existing service region S.
2. For each census tract (as obtained in step 1), store their neighboring census tracts (i.e., zones sharing common boundaries).
3. Select a zone (let's say x) and generate 3 scenarios: L1, L2, L3, where, L1 constitute the direct neighbors of zone x, then we add the neighbors of each zone in L1 to get L2, and for L3 we add neighbors of all zones in L2 to the existing L2 scenario. Figure 3.6 provides an illustration of service region scenario generation. Randomly select one (from L2 and L3) as a service region $S(\omega)$.
4. Assume a pricing policy as shown below:
    - PP1 = fixed fare for door-to-door services and first last mile rides
    - PP2 = fixed fare for door-to-door services and free first last mile rides
    - PP3 = variable fare

    Apply the most common pricing policies PP1 and PP2 to each of the sample scenarios from step 3.



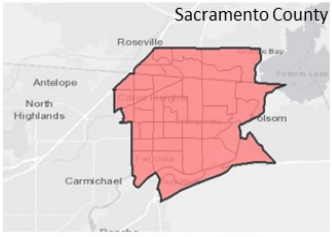
To provide a better idea on how the population data and simulated microtransit performance data for different scenarios look like, Figure 3.7 shows examples of four service regions generated in different U.S. counties using Algorithm 3.2. Each scenario is comprised of a set of census tracts based on which we obtain the aggregate population data (as listed in the figure) for the scenario. Microtransit performance data (e.g., ridership, vehicle miles traveled, fleet size, and others) is obtained using the calibrated simulation model for the associated pricing policy considered in the scenario. To ensure that we cover a diverse set of scenarios with a reasonable range of the population characteristics and ridership for the forecast models, $S(\omega)$ in step 5 of Algorithm 3.2 is randomly selected from different clusters of scenarios. In particular, for multiple

scenarios generated for a city (let's say all L2 scenarios), we divide them into 6-8 clusters based on their population characteristics (using k-means clustering algorithm (MacQueen, 1967)). Then from each of these clusters, we randomly select sample scenarios and apply pricing policies (PP1 and PP2) to obtain simulated microtransit performance data for the selected scenarios. This way we obtain upscaled data that is used for developing forecast models for service portfolio design and deployment planning as demonstrated in the case study in Section 4.



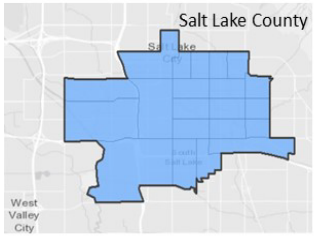
census tracts = 22; L2,
Pricing policy: PP2

Population	145570
Unprotected area (acre)	34577.08
Transit stop density	0.001
Street density	10.963
Population density	4.210
Employment density	2.236
Household (HH) density	1.405
Auto ownership per household	0.914
HH weighted mean income (dollars)	115566
Microtransit ridership	549
Fleet size	45
Vehicle miles traveled	2054.88
Revenue estimated (dollars)	203
Operation cost estimated(dollars)	10292



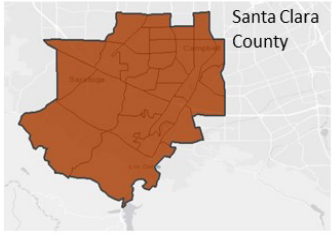
census tracts = 41; L3,
Pricing policy: PP1

Population	154988
Unprotected area (acre)	25247.78
Transit stop density	0.005
Street density	15.691
Population density	6.139
Employment density	2.946
Household (HH) density	2.395
Auto ownership per household	0.980
HH weighted mean income (dollars)	106630
Microtransit ridership	92
Fleet size	9
Vehicle miles traveled	424.52
Revenue estimated (dollars)	322
Operation cost estimated(dollars)	1721



census tracts = 21; L2,
Pricing policy: PP1

Population	85223
Unprotected area (acre)	10630.45
Transit stop density	0.042
Street density	19.542
Population density	8.017
Employment density	4.616
Household (HH) density	3.312
Auto ownership per household	0.875
HH weighted mean income (dollars)	80421
Microtransit ridership	457
Fleet size	19
Vehicle miles traveled	804.98
Revenue estimated (dollars)	1143
Operation cost estimated(dollars)	1914



census tracts = 28; L3,
Pricing policy: PP2

Population	165336
Unprotected area (acre)	22040.15
Transit stop density	0.014
Street density	17.791
Population density	7.502
Employment density	3.795
Household (HH) density	2.768
Auto ownership per household	1.003
HH weighted mean income (dollars)	186384
Microtransit ridership	1085
Fleet size	31
Vehicle miles traveled	1501.18
Revenue estimated (dollars)	294
Operation cost estimated(dollars)	5746

Figure 3.7. Examples of generated scenarios in different U.S. counties with population data and simulation data obtained for the scenarios

4. Portfolio Model for U.S. Microtransit Deployment

The proposed methodology is evaluated as follows. Since the contribution focuses on simulating new data to supplement limited existing data for the purpose of evaluating deployments in different cities, our objective is to show that:

1) A forecast model can be specified (having statistically meaningful relationships between public data and ridership/VMT) using the upscaled data that

2) Adequately fits the limited data that we have.

A case study is conducted in collaboration with Via, who provided aggregate ridership data for six different U.S. cities: Salt Lake City, Austin, Cupertino, Sacramento, Columbus, and Jersey City, as summarized in Table 4.1. The benchmark is a forecast model that is built using only the six cities' data, which is not statistically viable since that would simply be an insufficient sample size.

Table 4.1. Via service regions (obtained from Via) and pricing policies (defined based on (Via, 2021))

City (Via service region)	Counties (transit demand considered for potential first last mile trips)	Number of census tracts within service region boundary	Pricing policy
Salt Lake City, Utah	Salt Lake County	26	PP2
Austin, Texas	Travis and Williamson County	28	PP2
Cupertino, California	Santa Clara County	21	PP2
Sacramento, California	Sacramento and Placer County	148	PP2
Columbus, Ohio	Franklin and Licking County	45	PP1
Jersey City, New Jersey	Hudson and New York County	68	PP3

4.1. Data

Table 4.2 presents the data used in our case study, this includes public data for estimation of mode choice models, simulation model calibration, and design of deployment portfolio forecast model, along with data obtained from Via. For our case study, we focus on ridership during peak period of the day i.e., 6AM-9AM, hence all the commute (demand and time data) and Via operation data are considered for this time period of the day.

Table 4.2. Summary of data and data sources used in the study

Data source	Granularity	Data
Census Transportation Planning Products (2012-2016) (CTPP, 2016)	Census Tracts	Commute flows between census tracts (for various modes including auto, bike, transit, walk, and others)
American Community Survey (2019)	Census Tracts	Demographic, economic and household details
Open Mobility Data (GTFS) (Transitfeeds, 2021)	Transit network	Transit station/stop locations
Smart Location Database (EPA, 2021)	Census Block Groups (aggregated to Census Tract level)	Details on household auto ownership, unprotected area, street network (road density), trip equilibrium index (trip attraction and production)
Open Street Map (OSM, 2021)	Street network	Auto, Walk, and bike travel time between census tracts; walk and bike travel time to and from nearest transit stops in census tracts
Open trip planner (OTP, 2021)	Transit network	Transit commute time including in-vehicle time, wait time, walk time (to and from the nearest stops)
Via data: weekly average during first week of 3/20	Via service regions in Salt Lake City, Cupertino, Austin, Columbus, Sacramento, and Jersey City in the U.S.	Via service region boundaries, average ridership, average wait time, average ride distance and duration, vehicle utilization, pricing policy, fare structure

4.2. Calibration of the Market Equilibrium Model

The calibration of the model involves two major parts. The first is the estimation of the mode choice model, one for each of the six cities with provided data. After the estimation, each model is connected to a day-to-day market equilibrium model that is calibrated to fit the within-day simulation so that the equilibrium ridership is close to the observed aggregate average Via ridership values in the region (combining direct and first/last mile trips). The mode choice model includes one round of feedback from the equilibrium model output to update the Via attributes. Table 4.3 presents the mode choice model estimation results that were calibrated after one round of feedback from the simulation model. The p-values and ρ^2 are based on the initial estimation without microtransit (since there's no OD-level flow data for microtransit).

The bottom part of the table compares the estimated error for each city's model when using the optimal asc_{MT} compared to a model where the $asc_{MT} = 0$, and the error reduction is significant. The travel time and cost coefficients are negative in most cities, with Salt Lake City having a positive coefficient for walk time. Moreover, positive $\mathrm{asc_{MT}}$ values observed for Salt Lake City and Cupertino indicate a positive (average) effect of latent (unincluded) factors on the utility of the microtransit (Via) in these cities, while an opposite effect is noticed in other 4 cities. This observation highlights the effects the city type (among other latent factors) may have on the utility of microtransit in the city.

Table 4.3. Demand model parameter estimated values (for Via cities)

Mode choice model coefficient estimates for Via cities

Coefficient	Units	Salt Lake City, Utah	Austin, Texas	Cupertino, California	Sacramento, California	Columbus, Ohio	Jersey City, New Jersey
asc_{auto}	N/A	0.649***	-0.145*	*Not signif.*	0.231***	0.330***	*Not signif.*
asc_{bike}	N/A	-3.318***	-4.393***	-3.934***	-2.494***	-6.555***	-4.004***
$asc_{transit}$	N/A	-1.510***	-1.956***	-0.707***	-0.682***	-1.329***	*Not signif.*
asc_{MT}	N/A	0.848	-1.096	2.089	-0.689	-7.354	-2.265
asc_{walk}	N/A	-1.973***	-3.909***	-2.363***	-0.312***	-1.839***	0.560***
$\beta_{tt_{auto}}$	1/min	-0.204***	-0.049***	-0.131***	-0.109***	-0.009**	-0.177***
$\beta_{tt_{bike}}$	1/min	-0.129***	-0.051***	-0.098**	-0.105***	*Not signif.*	-0.251***
$\beta_{tt_{walk}}$	1/min	0.033***	-0.006	-0.038	-0.064***	-0.037***	-0.086***
β_{cost}	1/U.S. $	-1.851***	-2.062***	-1.768***	-1.058***	-0.998***	-0.930***
$\beta_{interzone}$	N/A	8.326***	12.895	7.403***	6.987***	6.356***	5.429***
β_{AE}	1/min	-0.005	*Not signif.*	*Not signif.*	-0.021***	-0.016**	*Not signif.*
$\beta_{tt_{transit}}$	1/min	-0.003	*Not signif.*	*Not signif.*	-0.012***	-0.009**	-0.001
β_{wait}	1/min	-0.005	*Not signif.*	*Not signif.*	-0.019***	-0.014**	-0.002
β_{MTwait}	1/min	-0.005	-0.075***	-0.200***	-0.019***	-0.014**	-0.002
Mode choice model calibration performance for Via cities							
ρ^2 (w/o microtransit)		0.78	0.69	0.72	0.78	0.85	0.43
Min. absolute error (pred vs actual Via ridership) with estimated asc_{Via}		0.004	0.008	0.003	0.008	0.001	0.002
Min. absolute error (pred vs actual Via ridership) with asc_{Via}=0		75.56	244.11	42.91	182.83	1167.14	1908.97

, **, * refer to p-values from initial estimation without microtransit less than 0.05, 0.01, and 0.001 respectively.*

The day-to-day adjustment parameters are calibrated as follows. Parameters include the walking limit of micro-transit users, micro-transit dwell time, and user/operator weights for the insertion heuristic in micro-transit within-day simulation. The performance measure for finding the best insertion option is shown as Eq. (4.1), which is a combined measure of the users' loss and the operator's loss balanced by operator's weight α_{oper} and user's weight $(1-\alpha_{oper})$. Average operator cost per mile is estimated from the average operation cost per passenger provided by (Volinski, 2019), and the average trip length data provided by Via (Eq. (4.2)).

$$Performance\ measure = (1-\alpha_{oper}) \times Value\ of\ time \times User\ time\ increment + \alpha_{oper} \times Operator\ cost\ per\ mile \times Distance\ traveled\ increment \quad (4.1)$$

$$Operator\ cost\ per\ mile = \frac{Operator\ cost\ per\ passenger}{Average\ trip\ length\ (mile)} \quad (4.2)$$

For calibration, we designed 3 discrete levels for each parameter:

- Walking limit: 0.5 miles, 0.3 miles, 0.1 mile
- Dwell time: 15 sec, 10 sec, 5 sec
- Operator weight in insertion heuristic: 0.8, 0.5, 0.2 (the corresponding user weight is 0.2, 0.5, 0.8)

Hence, 27 combinations are produced. We run the simulation for each of the combinations to find the best combination for each city, which is the combination with the smallest sum of squared error of in-vehicle time, wait time, and microtransit ridership.

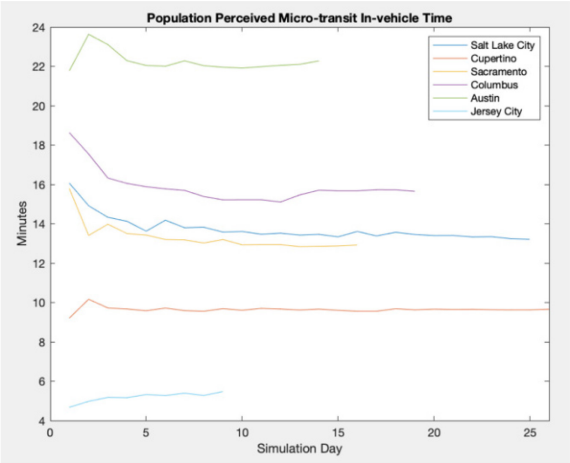
The calibration results are shown in Table 4.4. This shows that the cities can vary in their characteristics. For example, Salt Lake City and Jersey City suggests longer access via walking for travelers, while Austin and Salt Lake City tend to have longer dwell times for their vehicles. Cupertino has the highest weight for operator cost, which suggests that their travelers are the least elastic to the service quality (and hence more weight is assigned to operator cost). Generally, cities with smaller walking limit have smaller operator weight, since when the users are more reluctant to walk, user's weight should be higher. In terms of error, the overall ridership error indicates fits with an average of 18.4% among the six cities. Columbus had a poorer fit. Jersey City also had less data available so that the in-vehicle and wait time errors could not be computed.

Table 4.4. Summary of calibration results

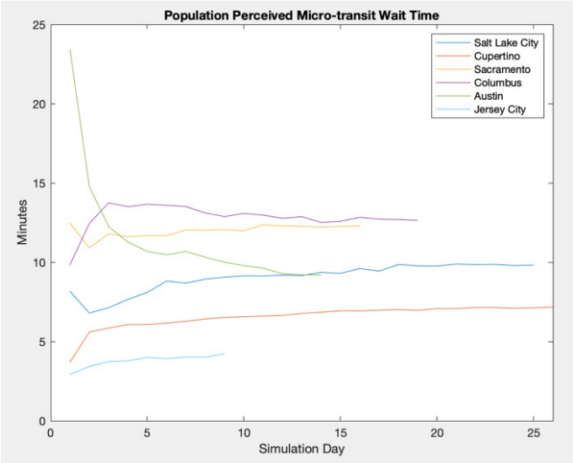
City	Calibrated Parameters			Operator cost per mile ($)	Error					
	Walking limit (mile)	Dwell time (sec)	Operator Weight		In-vehicle Time Error	%	Wait Time Error	%	Ridership Error	%
Salt Lake City	0.5	15	0.5	5.3	2.2	20.8	4.3	32.5	56	41.5
Austin	0.1	15	0.2	9.9	13.4	158.8	0.9	9.1	21	12.1
Cupertino	0.3	5	0.8	8.5	1.6	16.5	5.8	46.7	1	2.0
Sacramento	0.1	5	0.2	7.3	0.2	1.7	15.5	55.2	44	20.0
Columbus	0.1	5	0.2	8.3	7.1	93.2	8.5	136.4	2	33.3
Jersey City	0.5	5	0.5	7.9	-	-	-	-	3	1.2
AVG.										18.4

The process of convergence for the 6 cities with the calibrated parameters are shown in Figure 4.1. The average computation times for one run of Salt Lake City, Cupertino, Sacramento, Columbus, Austin, and Jersey City are respectively 10min 42s, 4min, 6min 24s, 36s, 4min 42s, and 13 min on a laptop with 2.3 GHz Quad-Core Intel Core i7 and 32 GB 3733 MHz LPDDR4X memory. The results indicate that steady states do exist for these cities and that the number of days to convergence can differ from city to city.

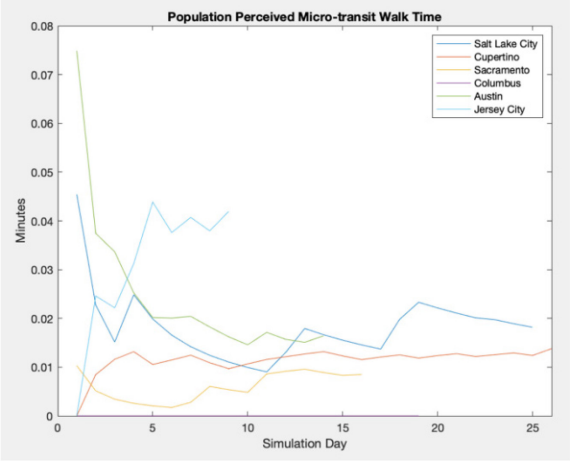
The results indicate that a market equilibrium model can indeed be calibrated to different cities, and the fit to the limited data is on average within 20% which is acceptable. Table 4.5 shows the ridership, VMT, fleet size, and perceived time values of population at convergence for the 6 cities as obtained using the calibrated market equilibrium model. One interesting observation around the Via ridership is that for Salt Lake City and Sacramento, higher proportions of Via ridership are within service region door-to-door trips, while for the other 4 cities, first-last mile Via access trips predominantly contribute to the Via ridership in respective cities. This highlights the variable effects different operation strategies can have on microtransit ridership and consequently on other performance measures (like VMT and fleet size) in different groups of cities.



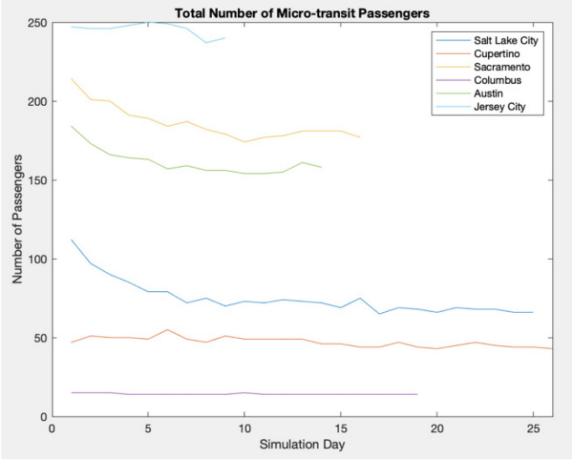
(a) Population Perceived Microtransit In-vehicle Time



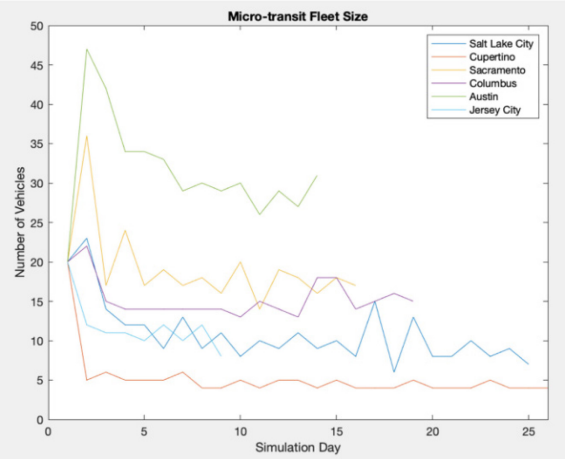
(b) Population Perceived Microtransit Wait Time



(c) Population Perceived Microtransit Walk Time



(d) Total Number of Microtransit Passengers



(e) Microtransit Fleet Size

Figure 4.1. Convergence of day-to-day adjustment for the 6 cities, with (a) in-vehicle time, (b) wait time, (c) walk time, (d) ridership, and (e) fleet size

Table 4.5. Summary of microtransit performance in 6 U.S. cities based on the calibrated market equilibrium model; *obs* is the Via observed data while *sml* refer to the output of the calibrated market equilibrium model

Cities	# days until convergence	Fleet size	VMT	Microtransit (Via) 6-9AM ridership (Mode Share)				Perceived time values of the population at convergence day				
				Total *(obs)*	Total *(sml)*	Within service region *(Mode share %)*	First/ Last-mile access *(Mode share %)*	In-vehicle time (min)		Wait time (min)		Walk time (min)
								obs	*sml*	*obs*	*sml*	*sml*
Salt Lake City	43	9	405.28	135	79	51 *(0.23)*	28 *(2.14)*	10.45	12.69	13.24	8.91	0.03
Cupertino	23	6	246.58	50	49	9 *(0.16)*	40 *(3.97)*	9.71	11.32	12.42	6.61	0.01
Sacra mento	22	18	883.19	220	176	152 *(0.25)*	24 *(0.47)*	12.03	12.24	28.09	12.62	0.01
Columbus	10	7	298.43	6	8	0 *(0)*	8 *(0.85)*	7.62	14.7	6.23	14.69	0
Austin	15	27	1220.88	174	153	18 *(0.40)*	135 *(12.10)*	8.44	21.81	9.94	9.00	0.02
Jersey City	11	11	519.73	245	242	79 *(0.48)*	163 *(0.88)*	n/a	5.38	12.9	4.02	0.02

4.3. Microtransit Deployment Forecast Portfolio Model

First, we wish to see if we can use the well-fitted market equilibrium (as shown in Section 4.2) to generate new scenarios to use to upscale the data for inferring new insights for microtransit deployment. This can be effectively demonstrated by showing that the scenario data can be related to public data to find statistically significant relationships.

4.3.1 Forecast Model Estimation and Validation

Two sets of models are estimated: one for predicting ridership and one for predicting fleet VMT. For the scenario generation, because Jersey City operates under a very different operation and Columbus is such an outlier, those two cities are removed from this section's demonstration effort. In future research, with more city data available one should ideally classify clusters of city types (like in (Oke, et al., 2019)) that can be fitted to different forecast models.

These scenarios are assumed to cover a reasonable range of ridership and pricing policies. A set of 326 scenarios are generated, with characteristics shown in Table 4.6. Based on the peak period average ridership and VMT values derived from the market equilibrium of those 326 scenarios used as surrogate data, we develop microtransit portfolio forecast models using multiple linear regression with second order polynomial (interacting) features. Here the dependent (target) variables for the ridership and VMT models include ridership per region population (in thousands) and VMT per region area in acres (in hundreds) respectively. We fit this model using the method of least squares and apply lasso regularization for feature elimination.

Table 4.6. Summary of data samples from scenario generation process used in forecast models

Number of scenarios	326
Breakdown by city	Salt Lake City: 71, Austin: 79, Sacramento: 100, Cupertino: 76
Breakdown of PP1/PP2	PP1: 174, PP2: 152
Breakdown of L2/L3 scenarios	LL2: 178, LL3:148
Range of number of riders	[0,2217]
Breakdown of direct trips versus first/last mile	direct: [1% - 88%]; first/last mile: [12%-99%] of total ridership

We consider the following independent variables (pertaining to each service region) in our models, where the feature (variable) values of a region are computed as the aggregate of all census tracts in the region (details below):

- Employment density (Total employed population in the region over total unprotected region area in acres)
- Household density (Total households in the region over total unprotected region area in acres). This is highly correlated to total population, and male/female population density features, hence we consider only one of these in our models.
- Mean income (Household weighted mean income in the region in U.S. dollars)
- Street density (Total road network in the region in miles over total unprotected region area in acres)
- Transit stop density (Total number of transit stops in the region over total unprotected region area in acres)
- Ratio of households with one or more auto (Sum of households with 1 or more auto ownership with respect to total households in the region)
- Trip equilibrium index (mean trip productions and trip attractions equilibrium index in the region; the closer to one, the more balanced the trip making)
- PP1 (if pricing policy in the region is PP1 then 1 else 0)
- PP2 (if pricing policy in the region is PP2 then 1 else 0)
- Via fix fare (value of fixed fare in the region in U.S. dollars)

We use the 326 sample scenarios surrogate data for training the forecast models. The evaluation (i.e., test set) is mainly done over the four data points (i.e., Via operated service regions) for which we have the actual Via ridership data and corresponding VMT values from the simulation. We consider the coefficient of variation (CV) as an evaluation metric, where CV is calculated using Eq. (4.3).

$$CV = \frac{\sqrt{\frac{\sum_{i=1}^{N} y_i - y'_i}{N}}}{\bar{Y}} = \frac{RMSE}{\bar{Y}} \qquad (4.3)$$

where y_i is the actual value and y'_i is the predicted value of the target variable for a sample i (in sample size N); $\bar{Y}$ is the mean of the actual values of the target variable across all samples. Table 4.7 shows the estimation performance of the forecast models.

Table 4.7. Estimation results for the ridership and VMT model

Model estimation				
City	**Ridership model**		**VMT model**	
	Observed	**Predicted**	**Observed**	**Predicted**
Austin	174	277	1220.88	1505.46
Cupertino	50	19	246.58	152.94
Sacramento	220	225	883.19	1068.07
Salt Lake City	135	211	405.28	778.89
Model performance				
	Ridership model		**VMT model**	
Train set R^2	0.72		0.9	
Number of features (including intercept)	47		55	
Via cities RMSE	65.92		256.67	
Via cities mean	144.75		688.98	
Via cities CV (%)	45.54		37.25	

The model suggests the ridership and VMT are indeed dependent on employment density, household density, mean income, street density, transit station density, and car ownership (estimated feature coefficient values are included in Appendix A). In addition, the models show sensitivity to the pricing policy. While the model outputs the statistically significant features based on the Lasso regularization, this is still done using upscaled data, so it is not feasible to compute an elasticity with respect to these features. Nonetheless, this estimation effort demonstrates that upscaling data from just four cities, we can fit models well (R^2 values fit quite well).

The key question is whether upscaling improves over having no upscaling at all. Without upscaling, data from only the four cities would not allow for even a forecast model to be estimated in the first place. When the model's predictions are compared to the four data samples, the CV of ~ 45% based on only four observations is rather adequate. While this is not within an accurate forecast range, it demonstrates that upscaling can result in more informative insights than relying only on the original data from four the cities alone.

4.3.2 Application of Forecast Models for Deployment Planning

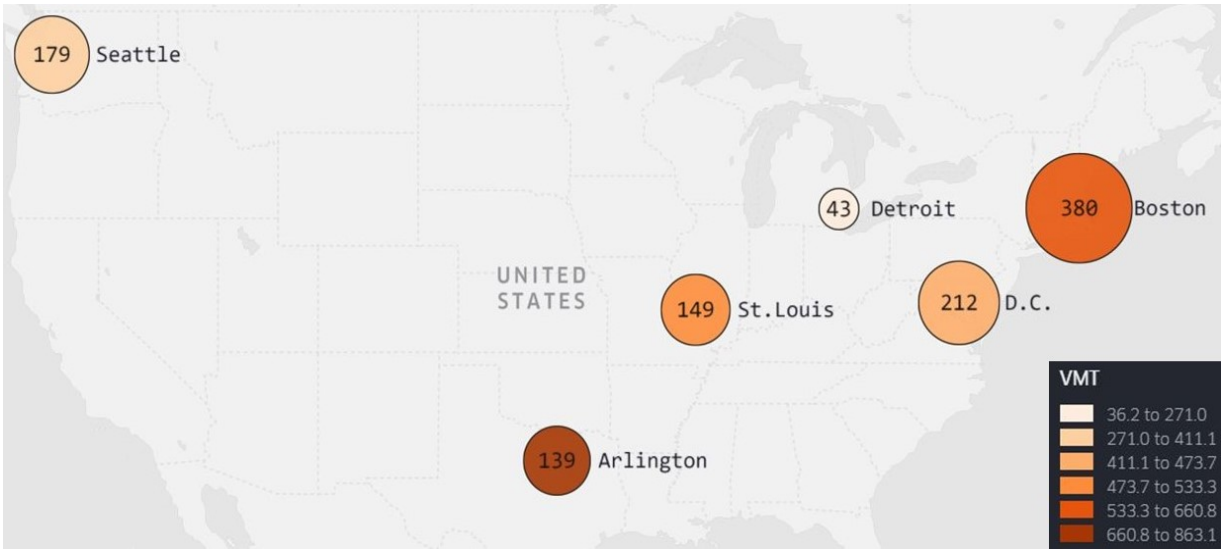
To provide a better idea of how the forecast models can be used for microtransit service portfolio design in different cities, we consider eight new cities in the U.S. (other than the cities considered in our study) i.e., Arlington (Texas), Birmingham (Alabama), Boston (Massachusetts), Chicago (Illinois), Detroit (Michigan), Seattle (Washington), St. Louis (Missouri), and Washington D.C. Assuming a constraint on total VMT (i.e., a budget constraint around the same value as the total VMT observed for the four Via cities considered in the case study), we present alternative portfolios for service deployment in different cities. For each of the eight cities, we generate various L2 scenarios (service regions) and get their population and built

environment characteristics. We apply PP1 pricing policies to the cities with fares based on the transit fares in respective cities. We use the VMT forecast model and select service regions from different cities such that the total forecasted VMT matches the budget considered. We design two alternative service portfolios as shown in Figure 4.2 and 4.3 and use the ridership forecast model to estimate the peak period ridership for the two portfolios.

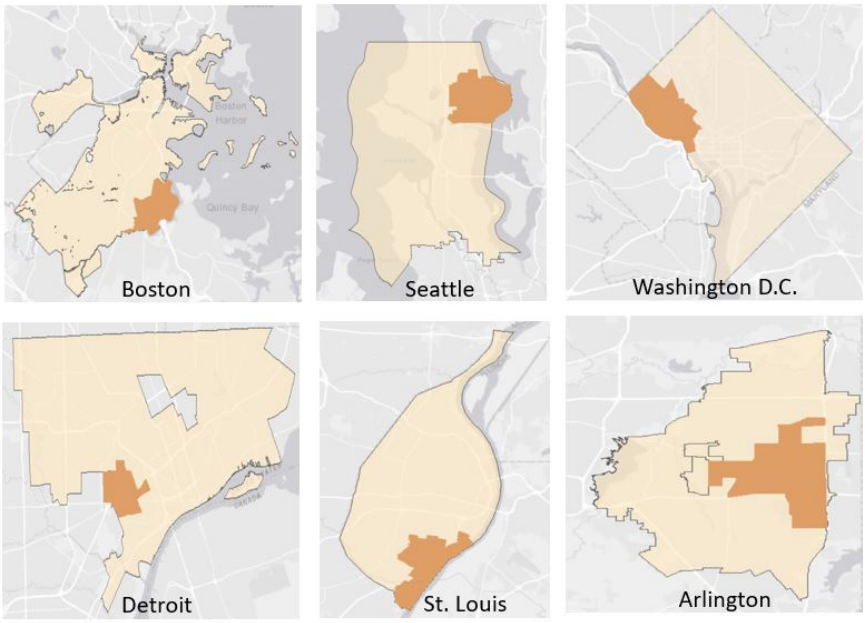
(a)

(b)

Figure 4.2. Portfolio design #1 for microtransit service deployment in 4 U.S. cities (a) estimated ridership and VMT in each city; circle radius is by ridership (values labeled in the figure), and circle sequential colors is by VMT (in the legend) (b) microtransit service regions in cities



(a)



(b)

Figure 4.3. Portfolio design #2 for microtransit service deployment in 6 U.S. cities (a) estimated ridership and VMT in each city; circle radius is by ridership (values labeled in the figure), and circle sequential colors is by VMT (in the legend) (b) microtransit service regions in cities

The total forecasted ridership values in portfolios #1 and #2 are 1.4 times and 1.9 times higher than the total ridership of the four Via cities for the same value of total VMT. Although we have presented only two alternative portfolio designs assuming one pricing policy in all cities and a VMT constraint, microtransit

operators can use such forecast models to help public agencies compare across multiple service portfolios by optimizing for ridership, considering additional operation cost constraint, applying different pricing and operating policies to specific cities (e.g., based on city types), etc. Hence, this can be used as an effective decision-support tool for microtransit service deployment planning for strategizing resource-allocation and investment decisions.

4.4. Discussion

Transportation technologies are not "one-size-fits-all" solutions; this point is clearly demonstrated by the 67%/23%/50% failure rates of demand-responsive transport services implemented in UK/Europe/North America. Emerging technologies like microtransit need to have effective decision-support tools, which are limited by the complexity of the decisions that need to be made, the limited availability due to the "emerging technology" aspect, and due to the myriad of operations that expand the dimensionality of the problem further. For example, even a success story like Via only operates in less than 40 U.S. cities while there are over 3000 cities with populations of 10,000 or more.

Identifying the right cities to enter their markets requires having some understanding of the typology of these cities. We propose a supervised machine learning approach (details in Section 5) to predict a city's typology given the information in its Wikipedia page. Our proposed method leverages recent breakthroughs in natural language processing, namely sentence-BERT (Reimers & Gurevych, 2019), and shows how the text-based information from Wikipedia can be effectively used as a data source for city typology prediction tasks that can be applied to over 2000 cities worldwide.

5. City Typology Prediction using Wikipedia

City typologies or profiles based on the dynamics of mobility in cities can allow easy identification of comparable cities for learning best practices and policies in the urban mobility planning context. City typologies have been studied in the past based on broad economic and geo-graphic forms. For example, (Harris, 1943) classified cities by economic functions like manufacturing, retail, education, etc., while (Creutzig, et al., 2015) proposed 8 typologies oriented around socioeconomic and environmental indicators to classify 274 cities. In terms of transportation metrics, studies have found that cities do exhibit commonalities, whether it is in road networks (Louf & Barthelemy, 2014) or public transit services (Derrible & Kennedy, 2010; Fielbaum, et al., 2017). A number of research studies have also identified typologies for cities (Thomson, 1978; Cervero, 1998; Priester, et al., 2013) focusing on the transportation aspects. A recent study by Oke et al. (2019) uses hierarchical clustering to present a mobility-based typologization covering 331 cities (across the globe) using factors related to transportation, demographic, geographic, economic, and environmental dimensions of cities. In their study, authors present 12 typologies grouped into 6 high level categories as shown in Table 5.1.

Table 5.1. Summary of the high-level city typologies based on (Oke, et al., 2019)

City typology	Description	Example cities
Auto	Auto dependent	Washington DC, Toronto, Raleigh, Kuwait City
Bus transit	High usage of bus transit	Rio de Janeiro, Jakarta, Tehran, Mecca
Congested	Congestion in cities	Bangalore, Lagos, Manila, Port-au-Prince
Metro bike	High bike share and metro	Ningbo, Zhengzhou, Shenzhen, Chongqing
Mass transit	High usage of mass transit	Singapore, Seoul, Tel Aviv, London
Hybrid	Mix of mode choices	Busan, Lisbon, Santiago, Johannesburg

For the most part, researchers have restricted the scope of their analysis to a limited number of cities due to data scarcity in many cities (especially in developing countries) mainly relying on datasets made available by city agencies, private sector companies, transportation operators, and universities. As such, application of these typologies to cities at a large, global-wide scale, e.g., on the order of thousands of incorporated places around the world, is not currently feasible, especially across multiple countries. We explore the usage of Wikipedia as such a source of data for identifying city typologies.

Wikipedia is unique in many aspects; it is essentially the largest digital encyclopedia worldwide that is powered by millions of crowd-sourced content editors and moderators. Enhanced at a rate of over 1.9 edits every second (Wikimedia statistics, 2021), Wikipedia serves as a reliable and inexpensive source of information. Wikipedia not only provides a wide range of information on various aspects of cities (such as transportation, demography, geography, economy, environment, education, culture, and others) in a consolidated manner, but also does it at an unparalleled scale (i.e., covering thousands of cities across the world for which detailed data may not be available otherwise).

For example, in the study by (Oke, et al., 2019), New York City (NYC) was assigned *the transit-heavy* city typology label based on multiple data sources excluding Wikipedia. Assuming, the Wikipedia page on NYC has supporting evidence that the city is indeed transit-heavy, if a human reader was tasked with identifying supporting lines, it would be an easy task to surface the lines highlighted in Figure 5.1 (which strongly support the transit-heavy label). Furthermore, the Wikipedia page may also contain information on different aspects of the city including recent demographic estimates and numbers from the infobox fields which may be latent factors influencing the typology.

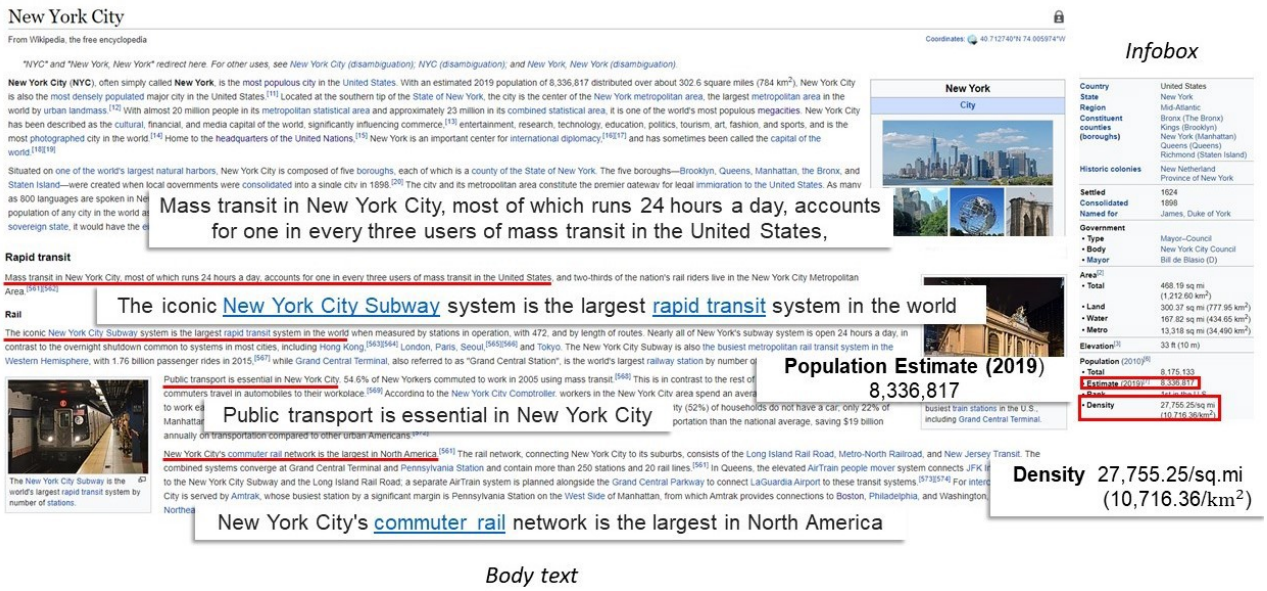


Figure 5.1. New York City's Wikipedia page: The infobox is structured with fields and theirvalues, whereas the article body text is unstructured. Lines indicative of the city typology (transit-heavy as per (Oke, et al., 2019)) have been highlighted

Similar information from Wikipedia on transportation scenarios in other cities could plausibly paint a picture of the typology of respective cities. But extracting such information in an automatic manner, e.g., via a model which *understands* the city's Wikipedia page from the perspective typology prediction is a challenging task.

Due to challenges in extracting useful information from long unstructured text, Wikipedia has remained a largely unexplored data source for transportation related studies. In the larger context of text understanding, the task of understanding and representing sentences (and paragraphs)for downstream prediction tasks has been an area of active research in the natural language processing (NLP) research community for several decades.

We propose a novel method to algorithmically extract lines from a city's Wikipedia page which semantically match a known set of possible typologies (*e.g.*, congested, transit-heavy, auto- heavy or bike-friendly), and use the typology-wise match scores to form a 4-dimensional vector representation (feature vector) for a city's Wikipedia page. In addition, we use information from structured components like the infobox (*e.g.*, population density) as an additional numeric feature for a city and use the resultant low dimensional feature vector for training a logistic regression model for city typology prediction. In particular, we use the labels in (Oke, et al., 2019) as ground truth labels for ~300 cities, and we adopt a one-versus-all approach for multi-class classification (*i.e.*, train binary classifiers for 4 different city typologies, and study their prediction accuracy). With such trained models, we can easily propagate the labels in ~300 cities in (Oke, et al., 2019), to over 2,000 cities in Wikipedia. Our main contributions can be listed as follows:

1. We propose a low dimensional representation of a city's Wikipedia page for the task of city typology prediction. The representation is based on algorithmically identifying lines in the Wikipedia page which semantically match (via SBERT) a known typology and use their match scores to form features.
2. We propose an iterative *keyline* expansion method which finds a set of representative lines (which we refer to as keylines) from Wikipedia pages of cities; these representative (key)lines allude to a known city typology and are crucial for identifying similar linesin cities beyond the training data set.
3. Using our trained model, we predict city typology scores (for different typologies) for over 2,000 cities present in Wikipedia. To the best of our knowledge, this is the largestdataset/analyses in scope on mobility-based city typology inferences.

We provide a brief literature review in Section 5.1, discuss the problem formulation and objectives in Section 5.2. We describe our proposed methodology in Section 5.3, followed by experiments in Section 5.4 and results in Section 5.5.

5.1. Literature Review

5.1.1. Usage of Crowd-sourced Data in Transportation and Urban Planning

The advent of user-generated information (including data from social media accounts, GPS, smart cards, and mobile phones) and the ease of data access opened various opportunities for multidisciplinary research. In the context of city understanding (particularly focusing on the transportation aspect), several studies have utilized location-based user-generated data (*e.g.*, social media check-ins, geo-tags, GPS, and smartphone data) (Zhan, et al., 2014). In this setting, the livelihoods project by (Cranshaw, et al., 2012) classifies the livelihood dispersion patterns in a city using geo-location data. (Hasan, et al., 2013) characterize human patterns in a city based on purpose-specific activities using location based social media data. Similarly, (Louail, et al., 2014) study the morphological patterns in 31 Spanish cities. Lenormand et al. (2015) perform a systematic comparison between five cities in Spain based on the land use patterns from mobile phone records. Calafiore, et al. (2021) model cities as series of global urban networks to obtain functional neighborhoods based on human dynamics and their contexts, across a sample of 10 global cities. A detailed overview of big data analysis for the systematic study of cities and urban phenomena can be found in (Lenormand & Ramasco, 2016) and (Martí, et al., 2019).

5.1.2. Text Understanding

Wikipedia is a rich source of crowd-sourced information on cities. The information in Wikipedia isgenerated at no cost (updated by numerous contributors worldwide) and verified regularly by moderators. So far, it has been a free online service, and is also freely available for off-line analysis. However, much of the useful (qualitative) information is in a textual format (in unstructured article bodies), and extracting such information automatically is difficult. A few recent studies have investigated Wikipedia as a potential indicator of city characteristics (*e.g.*, smart city related expressions (Cronemberger, et al., 2018), and for

understanding poverty and education across sub-Saharan African nations (Sheehan, et al., 2019). However, to the best of our knowledge, there is no prior work on predicting a city's typology usingits Wikipedia page; we believe that the hardness in understanding unstructured Wikipedia text and limited labeled data on city typologies (Oke, et al., 2019) may have contributed to the lack of prior work in this direction.

Fundamentally, in our supervised learning setup, the objective is to map a Wikipedia page (on a city) to a label (*e.g.*, binary label indicating whether it is transit-heavy or not). However, in itsraw form, the Wikipedia page is a collection of lines (with multiple words in a line). To representsuch a page numerically (for training a machine learning model and eventually mapping it to a label), one of the earliest approaches was to simply encode it in a bag-of-words fashion, *i.e.*, by using a vector of size equal to the vocabulary, and populating it with weighted word counts (TF-IDF (Jones, 1974; Salton, 1968)). However, such an approach led to high feature dimensions (English documents can easily lead to a vocabulary size in the order of tens of thousands), and the bag-of-words representation didnot capture the sequencing of words in a document (the sequence can easily alter the semantics).

A subsequent approach, Doc2Vec (Le & Mikolov, 2014), made progress towards low dimensional document representation (embedding), but was still limited in its ability to capture sequences and under- stand context. Around the same time, for short text (*e.g.*, a sentence), using the average of low- dimensional word embeddings (of words in the sentence) also became a popular method for sen- tence representation; Word2Vec (Markov, et al., 2021), FastText (Joulin, et al., 2016), and GloVe (Pennington, et al., 2014) are examples of such low dimensional word embedding methods. However, since the average does not capture the exact sequence of words, such methods are still limited in capturing the semantics.

A major breakthrough in the sentence representation problem came with the introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) which uses a neural network architecture called Transformer (Vaswani et al., 2017) to capture the exact sequence of words and learn thecontext (by utilizing both the forward sequence and the backward sequence, and hence becoming bidirectional in nature). BERT and its variants for text representation led to the state-of-the-art results for many natural language processing tasks (*e.g.*, text classification, and sentiment analysis). Furthermore, sentence-BERT (Reimers & Gurevych, 2019) leveraged a pre-trained BERT model (trained on millions of English examples) and was fine tuned for textual similarity. In other words, sentence-BERT can be used to get a 768-dimensional embedding for a sentence, and the semantic similarity of two sentences can be effectively gauged by the cosine similarity between their embeddings. We leverage this property of sentence-BERT embeddings in our proposed approach.

Finally, it should be noted that simply using sentence-BERT does not solve the problem of city typology prediction using Wikipedia; the remaining bottleneck comes from the low volume of labeled data (even lower than 768, which is the sentence-BERT dimension size), and the restriction that sentence-BERT can only process a sentence and not Wikipedia pages comprising of hundreds of lines. As described later, we propose a novel method to leverage sentence-BERT embeddings of lines in a Wikipedia page, to form a low

dimensional representation of the page such that supervised training with a few hundred labeled cities is sufficient.

5.2. Problem Formulation

5.2.1. Setup

We assume a set W of Wikipedia pages (size of the set denoted by ($|W|$). Each page $W_i \subset W$ corresponds to a unique city (anywhere in the globe). We focus on two components of each page in W: (i) unstructured text from different sections (main body), and (ii) structured data from the infobox (like the one illustrated in Figure 5.1). Sections commonly found in city Wikipedia pages include demography, geography, history, economy, education, and transportation (not necessarily with the same section titles), along with a general description of the city in the introduction paragraph. In our setup, for simplicity, we ignore text from the footnotes and references, as their mentions in the main body are preserved and typically provide enough context. In addition, from the infobox, we extract demographic information on population densities of cities. For cities without density data, we compute the missing values using population and area data extracted from the infobox.

As ground truth (typology) label for each city, we leverage the transportation related city typology labels provided by Oke et al. (2019) for 331 cities (worldwide). In the highest level of their hierarchical typologization, Oke et al. (2019) have six city typologies. For example, as shown in Table 5.1, typologies such as auto or mass transit mainly denote high usage of respective modes in cities, whereas congestion typology indicates high level of congestion in the city. Each of the 331 cities is assigned one of the six labels. Based on these typologies, we define four distinct city categories in our study *i.e.*, 'auto-heavy' (auto-dependent), 'transit-heavy' (high usage of public transit), 'bike-friendly' (high share of bike usage) and 'congestion'. For simplicity, we combine the mass transit and bus transit labels to form the 'transit-heavy' label (*i.e.*, a city is considered transit-heavy in our study if it is either labeled mass transit or bus transit in Oke et al. (2019)) and discard cities with the 'hybrid' label in Oke et al. (2019). This leaves us with 282 cities, with each city having one of the four possible labels: congestion, auto-heavy, transit-heavy, and bike-friendly (metrobike). We adopt a one-versus-all approach where we focus on separately classifying a city for auto-heavy, transit-heavy, bike-friendly, and congestion labels (as explained in Section 5.2.2 below).

5.2.2. Objective

The objective of our study is to automatically answer the following questions about a city given its Wikipedia page.

1. Congestion prediction: is the city congested?
2. Auto-heavy prediction: are automobiles the major mode of transport for this city?

3. Transit-heavy prediction: is public transit the major mode of transport for this city?
4. Bike-friendly prediction: is bike a common form of transportation in this city?

With the one-versus-all approach, we formulate four binary classification problems leadingto the following conditional probability estimates for a city i as Eqs. (5.1)-(5.4).

$\hat{p}_i^{(c)}$=*P*(city is congested | city's Wikipedia page) (5.1)

$\hat{p}_i^{(a)}$=*P*(city is auto-heavy | city's Wikipedia page) (5.2)

$\hat{p}_i^{(t)}$=*P*(city is transit-heavy | city's Wikipedia page) (5.3)

$\hat{p}_i^{(b)}$=*P*(city is bike-friendly | city's Wikipedia page) (5.4)

The training objective for each of the four binary classifiers is the minimization of the binary-cross-entropy loss (log-loss) across all training samples (Murphy, 2012). Hence, the objective for the congestion prediction classifier can be stated as Eq. (5.5).

$$\min\left[-\sum_{i=1}^{n}\left(label_i^{(c)}\right)\ln\left(\hat{p}_i^{(c)}\right)+\left(1-label_i^{(c)}\right)\ln\left(1-\hat{p}_i^{(c)}\right)\right] \quad (5.5)$$

where the sum is across all training samples (of size n), $label_i^{(c)} \in \{0,1\}$ is the congestion (binary) label for a city i (*i.e.*, the label is 1 if city is congested, 0 otherwise), and $\hat{p}_i^{(c)}$ is the probability estimate for the city being congested given the information in its Wikipedia page. The objectives for the auto-heavy, transit-heavy, and bike-friendly classifiers can be written in a similar manner.

5.3. Methodology

5.3.1. High-level Overview

In our supervised learning approach, we first represent a city i via a 5-dimensional feature vector. The feature vector includes congestion, auto-heavy, transit-heavy, and bike-friendly keyline features and a numeric feature (*i.e.*, population density). Intuitively, the congestion keyline featureindicates the presence of a line (text) in the city's Wikipedia page (main body) which strongly indicates that the city is congested; the term keyline is used since each line in the city's Wikipedia pageis checked for semantic similarity with a pre-determined set of representative (key)lines indicating congestion. The keyline features for auto-heavy, transit-heavy, and bike-friendly are designed in the same spirit. Eq. (5.6) represents the 5-dimensional feature vector f_i for a city i.

$$f_i = \begin{Bmatrix} f_i^{congestion} \\ f_i^{auto} \\ f_i^{transit} \\ f_i^{bike} \\ f_i^{density} \end{Bmatrix} = \begin{Bmatrix} f_i^{(c)} \\ f_i^{(a)} \\ f_i^{(t)} \\ f_i^{(b)} \\ f_i^{(density)} \end{Bmatrix} \quad (5.6)$$

where $\boldsymbol{f_i} \in R^5, f_i^{(c)} \in [-1,1]$ denotes the congestion keyline feature, $f_i^{(a)} \in [-1,1]$ denotes the auto keyline feature, $f_i^{(t)} \in [-1,1]$ denotes the transit keyline feature, and $f_i^{(b)} \in [-1,1]$ denotes the bike keyline feature (extracted from unstructured text in Wikipedia main body). Given the above city representation, our approach is to train four logistic regression models to come up with the estimates $\hat{p_i^{(c)}}, \hat{p_i^{(a)}}, \hat{p_i^{(t)}}\ \hat{p_i^{(b)}}$ for a city i as formulated in Eqs. (5.1)-(5.4). An illustration of the above process is shown in Figure 5.2.

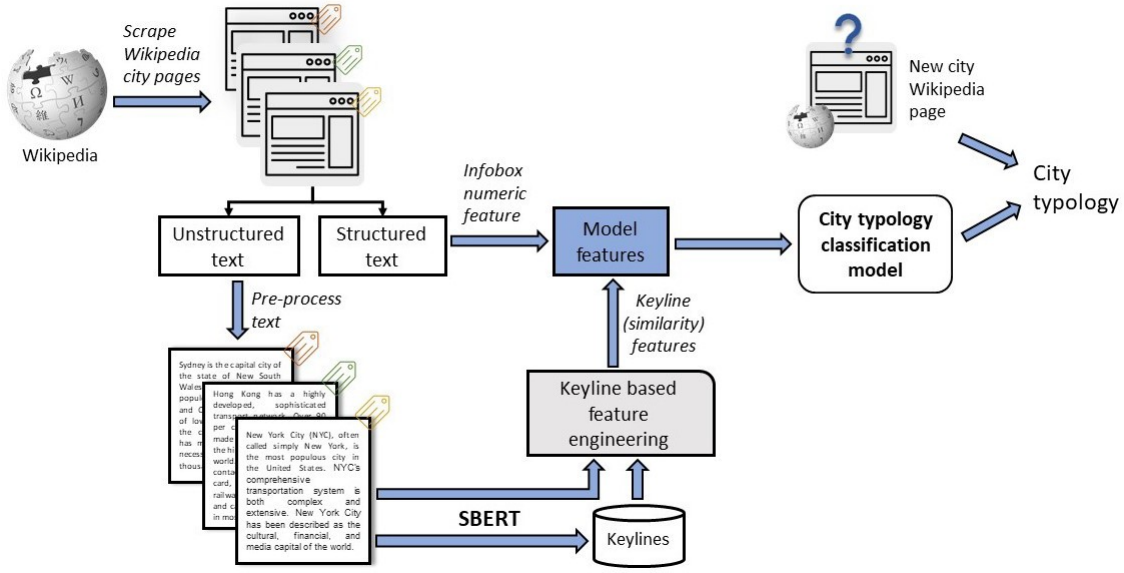


Figure 5.2. High-level overview of our proposed method for city typology classification using Wikipedia

5.3.2. Supervised Learning using Logistic Regression

For brevity, we describe the logistic regression model in the context of estimating $\hat{p_i}^{(c)}$ for city i, i.e., the chances of city i being congested given its Wikipedia page (the models for estimating $\hat{p}^{(a)}, \hat{p}^{(t)}, and\ \hat{p}^{(b)}$ can be described in a similar manner). For estimating $\hat{p_i}^{(c)}$ via logistic regression, we consider the following parametric form as in Eq. (5.7).

$$\hat{p_i}^{(c)} = \frac{1}{1+e^{-(w_T f_i + b)}} \tag{5.7}$$

where the learning parameters $\boldsymbol{w} \in R^5$, and bias $b \in R$ are optimized by minimizing the binary cross-entropy loss function as defined in Eq. (25). Note that for estimating $\hat{p_i}^{(c)}, \hat{p_i}^{(a)}, \hat{p_i}^{(t)}$ and $\hat{p_i}^{(b)}$, we employ the same set of features $\boldsymbol{f_i}$ (as used for $\hat{p_i}^{(c)}$ above). In the following section, we describe the keyline features that occupy the first 4 dimensions of $\boldsymbol{f_i}$.

5.3.3. Keyline-based Feature Engineering

5.3.3.1. Semantic Textual Similarity

The notion of semantic textual similarity (Reimers & Gurevych, 2019) originated in the field of NLP, where the underlying task was to automatically identify if two sentences have similar meaning (i.e., one sentence is a paraphrased version of the other). For example, a perfect model for semantic textual similarity is expected to identify the following similar and dissimilar sentences:

- Example 1: (this city is congested, this city suffers from traffic jams) → similar (score of 1),
- Example 2: (this city is congested, this city does not experience traffic jams) → dissimilar (score of -1)

Note that semantic textual similarity is a very challenging task; even in the above example,a model needs to be intelligent enough to understand that *congestion* and *traffic jams* are related, while *does not experience traffic jams* means there is no *congestion*. Due to the introduction of SBERT, the state-of-the-art performance for semantic textual similarity tasks has seen a step-jump (thereby encouraging downstream applications like the one we propose in this study). In our setup,we directly use the SBERT model fine-tuned for the semantic-textual-similarity task. SBERT can be used in the following manner to estimate the semantic similarity between apair of sentences (lines) l_i and l_j in Eq. (5.8).

$$similarity\left(l_i, l_j\right) = cosine\ similarity\ (\phi(l_i), \phi(l_j))$$

$$\Rightarrow \quad similarity\left(l_i, l_j\right) = \frac{\phi(l_i).\phi(l_j)}{norm(\phi(l_i)) \times norm\left(\phi(l_j)\right)}$$

$$\Rightarrow \quad similarity\left(l_i, l_j\right) = \bar{\varphi}(l_i)^T \times \bar{\varphi}(l_j) \tag{5.8}$$

where $\phi(l) \in R^{768}$ denotes the 768-dimensional sentence-embedding for sentence l using the trained sentence-BERT model in (Reimers & Gurevych, 2019). Eq. (5.8) is essentially the cosine-similarity between the embeddings of the two sentences (and is in the range [-1,1]). The cosine similarity between two vectors can be defined as the inner product of the same vectors normalized to both have length 1. The sentence embedding $\phi(l)$ is normalized to have unit l_2 norm; $\bar{\varphi}(l)$ denote the normalized vector and T denotes the transpose operation.

Essentially, the cosine similarity between two lines (l_i and l_j) can be calculated using matrix multiplication on the two normalized vectors ($\bar{\varphi}(l_i)^T$ and $\bar{\varphi}(l_j)$). Based on this calculation, two sentences with similar meaning will have higher scores than two non-similar sentences.

5.3.3.2. Keyline Similarity Features

Assuming the presence of a semantic-textual-similarity model as described above, we focus on the following idea. Consider the congestion prediction task, where one is given the Wikipedia page of a city and has to estimate the chances of the city being congested. Intuitively, if the Wikipedia page has line(s) which are semantically similar to '*the city suffers from traffic congestion*', there may be a good chance that the city is indeed congested. However, there may be other ways in which the city's congestion problem may be expressed in the Wikipedia page. For example, '*cars are stuck for hours on the main roads of the city on weekdays*' is another plausible (key)line representing congestion.

Building on this intuition, if we can construct a small yet representative set of keylines indicating congestion in a city, we can check each line in a city's page (as shown in the bi-partite graph in Figure 5.3 with the representative keylines to see if there is a high semantic similarity; having a set of keylines just casts a wider net compared to having just one keyline. The highest semantic similarity score across all possible pairs is derived from the keylines and the city's Wikipedia page lines can then serve as a *congestion* keyline feature for the city (i.e, $f_i^{(c)}$ for a city i), as illustrated in Figure 5.3. This is precisely the idea behind the congestion keyline feature proposed in this study and we give a formal description below.

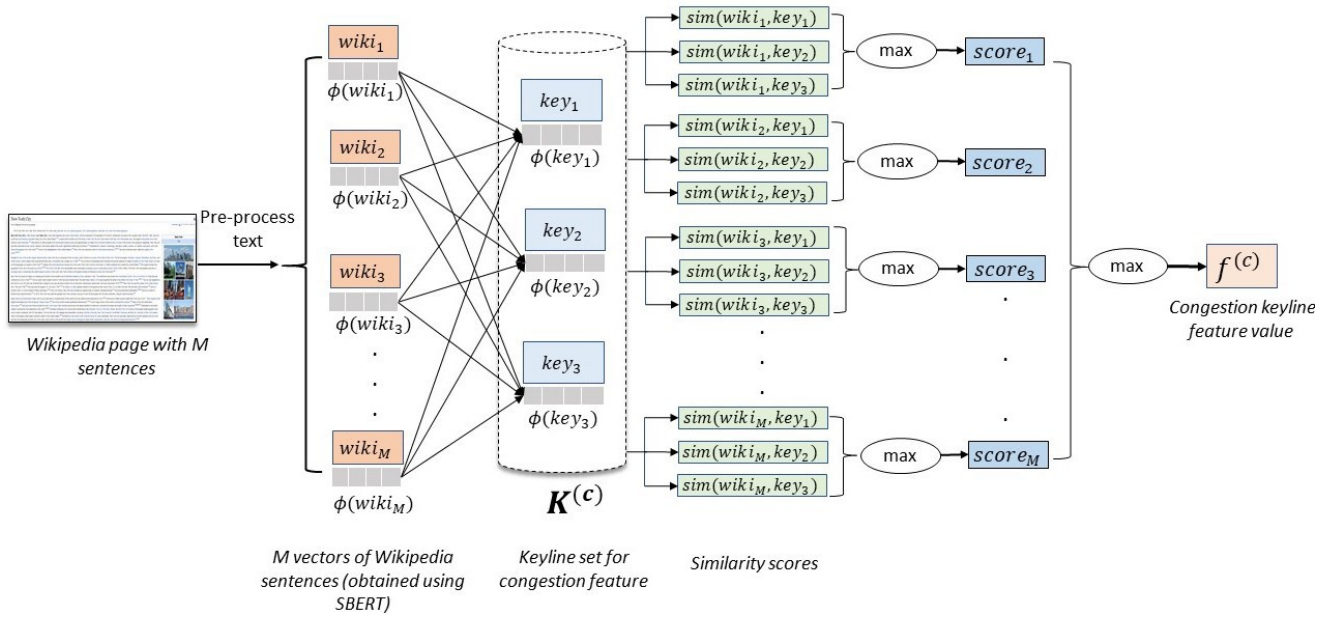


Figure 5.3. Illustration of congestion keyline similarity feature extraction from a city'sWikipedia page

Consider the Wikipedia page of city i and the indexed set of lines $\{wiki_{i1}, wiki_{i2}, \ldots\ldots, wiki_{iMi}\}$ in the main body of the page (assuming the page i has M_i lines of text). Using Eq. (26) to compute semantic similarity between two sentences, the congestion keyline feature can be computed as Eq. (5.9).

$$f_i^{(c)} = \max_{1\le j\le M_i} \max_{1\le k\le |\mathbf{K}^{(c)}|} similarity\left(wiki_{ij}, key_k^{(c)}\right) \quad (5.9)$$

where $wiki_{ij}$ is the jth line in the city i's Wikipedia page, $key_k^{(c)}$ is the kth keyline in the set of keylines for congestion (set denoted by $\mathbf{K}^{(c)}$), and the feature value $f_i^{(c)}$ is the maximum cosine similarity between all Wikipedia page lines of the city, and all the keylines in the congestion keyline set $\mathbf{K}^{(c)}$ (similarity calculated using Eq. (5.8)).

The auto, transit, and bike keyline features can be stated in a similar manner as shown in Eqs. (5.10)-(5.12).

$$f_i^{(a)} = \max_{1\le j\le M_i} \max_{1\le k\le |\mathbf{K}^{(a)}|} similarity\left(wiki_{ij}, key_k^{(a)}\right) \quad (5.10)$$

$$f_i^{(t)} = \max_{1\le j\le M_i} \max_{1\le k\le |\mathbf{K}^{(t)}|} similarity\left(wiki_{ij}, key_k^{(t)}\right) \quad (5.11)$$

$$f_i^{(b)} = \max_{1\le j\le M_i} \max_{1\le k\le |\mathbf{K}^{(b)}|} similarity\left(wiki_{ij}, key_k^{(b)}\right) \quad (5.12)$$

where $\mathbf{K}^{(a)}$, $\mathbf{K}^{(t)}$, and $\mathbf{K}^{(b)}$denote the auto, transit and bike keyline sets respectively. For computational efficiency, the above features can be computed via a matrix multiplication (between a matrix of stacked

embeddings of the Wikipedia page lines and matrix of stacked keyline embeddings). The above feature computations assume a set of keylines for congestion, auto, transit and bike. We describe in the following section an algorithm to obtain such keyline sets starting from an initial guess and iteratively extracting lines from Wikipedia pages of cities in the training dataset.

5.3.3.3. Keyline Sets (Initial Guess and Set Expansion)

We describe below our proposed method for constructing the set $\mathbf{K}^{(c)}$, i.e., keylines for congestion (the method for $\mathbf{K}^{(a)}$, $\mathbf{K}^{(t)}$, and $\mathbf{K}^{(b)}$ is similar and their description has been skipped for brevity). To construct the keyline sets $\mathbf{K}^{(c)}$, $\mathbf{K}^{(a)}$, $\mathbf{K}^{(t)}$, and $\mathbf{K}^{(b)}$ we start with initial guesses for these sets (one line for each set). These initial keyline guesses will be referred to as anchor text, and our choices in this study are listed in Table 5.2.

Table 5.2. Initial keylines (anchor text) for the city typology prediction tasks considered in this study

Keyline feature type	Initial keyline or anchor text	Notation
congestion	'the city has heavy traffic congestion'	$anchor^{(c)}$
auto	'most people in the city use cars'	$anchor^{(a)}$
transit	'most people in the city use public transit like bus and metro'	$anchor^{(t)}$
bike	'many people in the city use bike or cycle'	$anchor^{(b)}$

The keyline set expansion algorithm assumes that we have a dataset consisting of Wikipedia cities, and their congestion labels (i.e., congested or not congested). We randomly divide the dataset into train and test sets and use only the train data to extract additional keylines. Specifically, we focus on all cities in the train set which are labeled positive (i.e., congested). From each positively labeled city i in the train set, we extract a *candidate* congestion keyline as shown in Eq. (5.13).

$$candidate\ keyline_i^{(c)} = arg \max_{1 \le j \le M_i} similarity\left(wiki_{ij}, anchor^{(c)}\right) \qquad (5.13)$$

where $anchor^{(c)}$ is the anchor text for the congestion feature as listed in Table 5.2. $candidate\ keyline_i^{(c)}$ is the semantically closest line in the page of city i compared to the anchor text for congestion. For example, a candidate keyline obtained from the Wikipedia page of the city Manila is *'Manila is notorious for its frequent traffic jams and high densities'*.

Using similar notations as in Eq. (5.13), candidate keylines for auto, transit, and bike are extracted using Eq. (5.14) – (5.16) respectively.

$$candidate\ keyline_i^{(a)} = arg\max_{1 \le j \le M_i} similarity\left(wiki_{ij}, anchor^{(a)}\right) \quad (5.14)$$

$$candidate\ keyline_i^{(t)} = arg\max_{1 \le j \le M_i} similarity\left(wiki_{ij}, anchor^{(t)}\right) \quad (5.15)$$

$$candidate\ keyline_i^{(b)} = arg\max_{1 \le j \le M_i} similarity\left(wiki_{ij}, anchor^{(b)}\right) \quad (5.16)$$

For each feature, we collect all such candidate keylines from the training dataset (as shown in Figure 5.4) and sort them in decreasing order of similarity score with their anchor texts. For congestion we obtain the list of sorted candidate keylines as $sorted\_candidates^{(c)}$.

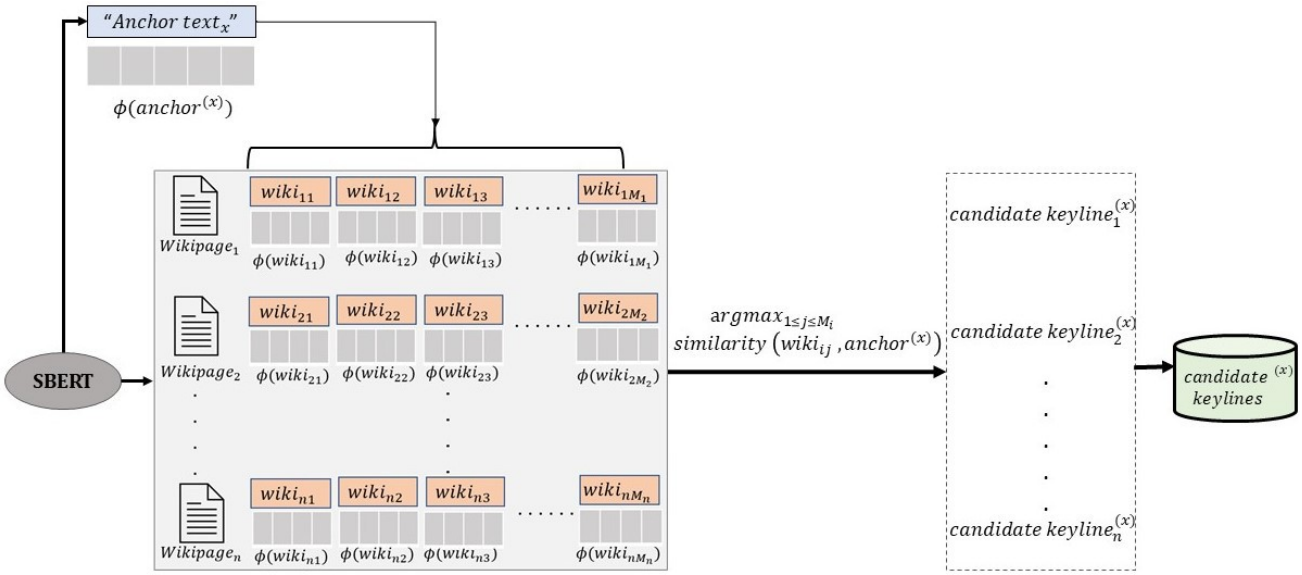


Figure 5.4. Selection of candidate keylines for a feature type x from n Wikipedia pages

The keyline set expansion method for the feature *congestion* is outlined in Algorithm 5.1. Wefirst initialize the feature keyline sets with the respective anchor texts (singleton keyline sets). We then go over all candidates in the ordered $sorted\_candidates^{(c)}$ list (in the train set), and greedily add one candidate at a time to the keyline set $\mathbf{K}^{(c)}$. In each iteration *e* (where we add a candidate), we derive the keyline similarity feature $f^{(c)}$ from the current iteration's keyline set $\mathbf{K}^{(c)}$. (using Eq. (5.9)). The $f^{(a)}$, $f^{(t)}$ and $f^{(b)}$ features for training the logistic regression model in each iteration are derived from only the anchor texts in the corresponding singleton keyline sets $\mathbf{K}^{(a)}$, $\mathbf{K}^{(t)}$, and $\mathbf{K}^{(b)}$ (using Eqs. (5.10)-(5.12)). Using the computed features, we train the logistic regression model for the congestion prediction task. Using the trained logistic regression model on the validation set, we record the validation set performance metric (AUC, *i.e.*, area under the receiver operating characteristic curve for binary classification predictions). We keep track of the

validation metric across all iterations, and finally select the iteration (and the corresponding keyline set) with the best validation set performance to arrive at the optimized (and expanded) keyline set for congestion.

Algorithm 5.1. Greedy keyline set expansion for congestion keyline feature (using validation set performance)

Input: train set, validation set, $\mathbf{K}^{(a)}$ = {$anchor^{(a)}$}, $\mathbf{K}^{(t)}$, = {$anchor^{(t)}$}, $\mathbf{K}^{(b)}$= {$anchor^{(b)}$}, $num\_candidates$ = $|sorted\_candidates^{(c)}|$ in train set

Output: $max\_expansion$, optimized expanded keyline set $\mathbf{K}_{opt}^{(c)}$

1. **Initialization**: max_AUC = 0, candidate_index = 1, $max\_expansion$ = 0, $\mathbf{K}^{(c)}$ = {$anchor^{(c)}$}
2. while candidate_index ≤ *num_candidates do*
3. ADD candidate from $sorted\_candidates^{(c)}|$ to congestion keyline set $\mathbf{K}^{(c)}$,
4. COMPUTE keyline similarity features: $f^{(c)}$, $f^{(a)}$, $f^{(t)}$ and $f^{(b)}$ from $\mathbf{K}^{(c)}$, $\mathbf{K}^{(a)}$, $\mathbf{K}^{(t)}$, and $\mathbf{K}^{(b)}$ respectively (using Eqs. (5.9)-(5.12)),
5. TRAIN congestion classification model using train set, compute validation AUC (using validation set),
6. if validation AUC > max_AUC do
   - max_AUC ← validation AUC,
   - max_expansion ← candidate_index,
   - **continue**
7. candidate_index ← candidate_index + 1

   $\mathbf{K}_{\text{opt}}^{(c)} = \mathbf{K}^{(c)}[:max\_expansion + 1]$

Note that the (typology) classification model used in the keyline expansion method is based on the feature whose keyline set is to be expanded. For example, if the method is applied to the congestion feature in order to get a representative set of keylines indicating congestion ($\mathbf{K}^{(c)}$), then the congestion typology classification model is used (as shown in Algorithm 5.1). Similarly, we use auto-heavy typology classifier for computing $\mathbf{K}^{(a)}$, transit-heavy typology classifier for $\mathbf{K}^{(t)}$, and bike-friendly typology classifier for $\mathbf{K}^{(b)}$ along with the features computed using these sets ($f^{(c)}$, $f^{(a)}$, $f^{(t)}$ and $f^{(b)}$). These features are used as initial input variables in the congestion typology classification model.

To assess performance of the LR models used in the keyline set expansion method (Algorithm 5.1 step 5), we perform 3-fold cross-validation. Cross-validation gives an idea about how well the trained model will generalize for an unseen dataset and avoids fitting the model to just the training dataset. To do this, at each iteration e,we split our training dataset into three equal parts; for each instance (part) in our dataset, we build a logistic regression model using all other instances and then validate it on the selected instance (i.e., validation set). In our setup, the 3-fold cross validation process is repeated three times for each iteration,

and the mean AUC value across all folds from all runs is considered as the validation AUC at that iteration (step 5 in Algorithm 5.1). This is done to reduce error in the estimate of mean performance of the model.

Note that the test data set is kept aside for final evaluation, and only the validation data set is used to optimize keyline set expansion. The keyline set expansion algorithm for auto, transit, and bike keyline features is similar, and we skip their description for brevity. Also, one may suggest considering all the candidate keylines from the training set (say, $\mathbf{K}_{\text{all}}^{(c)}$ for congestion feature) instead of the optimized feature keyline set ($\mathbf{K}_{\text{opt}}^{(c)}$ obtained using Algorithm 5.1 for congestion). However, using $\mathbf{K}_{\text{all}}^{(c)}$ for feature definition is not only computationally costly but also results in sub-optimal performance on the test set compared to the optimal keyline set.

5.4. Experiments

5.4.1. Data

In our study we focus on four binary classification tasks centered around: congestion, auto-heavy, transit-heavy, and bike-friendly typology prediction for a city. For training each of the above binary classifiers in a supervised manner, we need ground truth labels in the form of $(city_i, label_i^{task})$ where binary $label_i^{task} \in \{0,1\}$, and $task$ specifies the classification task (congestion, auto-heavy, transit-heavy, and bike-friendly). We obtain the task specific binary labels (ground truth) for a city following a recent study (Oke, et al., 2019) as described below.

We obtain 282 cities with each city having one of the four possible labels: congestion, auto-heavy, transit-heavy, and bike-friendly. We will refer to this dataset as the 'typology' dataset. The typology label distribution across these 282 cities is as follows: 27% congestion, 23% auto-heavy, 39% transit-heavy, and 11% bike-friendly. Using these four labels, we compute the task-specific binary labels for a city in a one-versus-all fashion, e.g., for the auto-heavy prediction task, all cities with the auto-heavy label are assigned a label of 1, and the remaining cities (from the remaining 3 typologies) are assigned a label of 0.

Once the city typology dataset is finalized as described above, we collect their Wikipedia addresses (URLs) using the (Wikipedia API, 2014). Using the city URLs, we automatically crawl the Wikipedia pages associated with the URLs and collect data from the main bodies and infoboxes. We use web scraping tools in Python (such as Beautiful Soup, Wikipedia API, and Pandas) for data cleaning and processing as described below.

5.4.1.1 Data from Unstructured Main Body

The qualitative information on each city in the typology dataset are collected from different sections in the respective Wikipedia pages (including sections like demography, geography, economy, transportation, infrastructure, education). Although we focus on the transportation aspect of cities, we chose to collect data

from all sections in a Wikipedia page, and not just the transportation section. This is because, in some cases, information regarding mobility scenario in a city is present in multiple sections such as infrastructure, and economy. The textual data extracted for each city are pre-processed (such as removing section titles, footnotes and references) and stored in a format such that given a city name, we can get a list of sentences from the city's Wikipedia page. Each sentence l in this list is converted to a 768-dimensional real-valued vector ($\phi(l) \in R^{768}$) using a pre-trained SBERT model (using the version stsb-distilbert-base (Reimers & Gurevych, 2019) pre-trained for semantic textual similarity). In this manner, for each city i, we obtain a list of M_i vectors (with each vector being 768 dimensional); here M_i denotes the number of sentences extracted from the Wikipedia page of the city i. Figure 5.5 provides an illustration of this process for New York City.

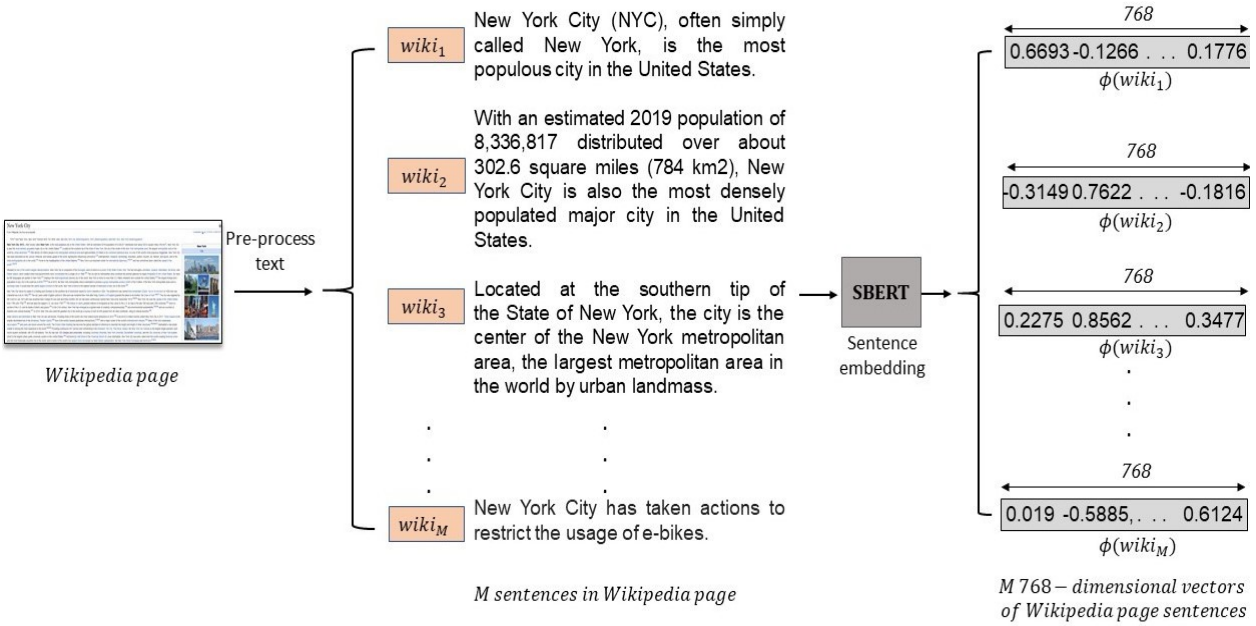


Figure 5.5. Textual data extraction from New York City Wikipedia page main body with M sentences and vector representation of these sentences using pre-trained SBERT model

5.4.1.2 Data from Structured Infobox

The quantitative information pertaining to each city's demography are collected from the corresponding Wikipedia infobox. We extract population density, population and area estimates (all in text format), pre-process the data, and convert them to numeric values. The density and area values in Wikipedia articles are available in both mile and kilometer units. For cities with no density information in their Wikipedia pages, we compute the missing values using their population and area estimates. The derived population densities (in sq.mi) for each city are first normalized and then used as numeric features in the typology classification models in our study. Additionally, the city coordinates are also obtained from their respective Wikipedia pages for visualization purposes.

5.4.1.3 Train and Test Datasets

For our typology prediction tasks, we use logistic regression (LR) model for each of the typologyclassifiers. Such models are trained (supervised) based on a training dataset so that the model learns the relationship between the input and output variables; the performance of the model is then assessed using a test dataset (different from training set yet representative of the dataset as a whole). Therefore, in our case we considered 70% of the city typology data for training (*i.e.*, 197 cities with 28% congestion, 23% auto-heavy, 36% transit-heavy, and 13% bike-friendly cities) and 30% for testing purpose (*i.e.*, 85 cities with 23% congestion, 23% auto-heavy, 46% transit-heavy, and 8% bike-friendly cities). For both the train and test datasets, we generate positive and negative labels (1s and 0s) for each typology classification model (binary classifier) based on the respectivetypology. For example, for *auto-heavy* model, we label cities with category *auto* as 1 while cities with categories other than *auto* (*e.g.,* congestion, transit, or bike) are labeled 0. Therefore, the trainset (197 cities) and the test set (85 cities) used in each classification model remain the same. However, the city typology labels (output variables) are modified based on the prediction task of the LR models.

5.4.2. Evaluation Metrics

For quantitatively measuring the performance of the typology classifiers in our study, we consider the metric widely used for evaluating binary classification models *i.e.,* AUC (Area under the curve) ROC (Receiver Operating Characteristics) curve (Murphy, 2012). The ROC curve is a probability curve that plots the true positive rate against false positive rate at various threshold values. AUC provides the summary of the ROC curve and can be used to compare classifiers directly without specific decision thresholds. In simple terms, the AUC score (AUC $\in [0,1]$) tells us how well the model is able to distinguish between positive and negative classes. For example, consider the congestion classification task: the higher the AUC, the better the model is at classifying between whether the city is congested or not. In addition, for the models with features obtained from optimal expanded keyline sets,we report the accuracy (fraction of samples where predicted label matches ground truth label), precision (true positives over the sum of true positives and false positives), recall (true positives over the sum of true positives and false negatives), and F-1 score (harmonic mean of precision and recall).

5.4.3. Initial Keyline Features

For each city in our data we obtain a list of 768-dimension vectors (the size of the list varies basedon the number of lines in each city Wikipedia page). Our proposed method of algorithmically extracting lines from a city's Wikipedia page (Algorithm 5.1) to semantically match the typology of interest provides a 4-dimensional keyline based feature vector for each city (as mentioned in Eq. (5.6). As outlined in Algorithm 5.1, the feature keyline sets $\mathbf{K}^{(\mathrm{c})}$, $\mathbf{K}^{(\mathrm{a})}$, $\mathbf{K}^{(\mathrm{t})}$, and $\mathbf{K}^{(\mathrm{b})}$ are initialized using their corresponding anchor texts. For clarity, we denote these singleton keyline sets as $\mathbf{K}^{(\mathrm{c})}_{\mathrm{intial}}$, $\mathbf{K}^{(\mathrm{a})}_{\mathrm{intial}}$, $\mathbf{K}^{(\mathrm{t})}_{\mathrm{intial}}$, and $\mathbf{K}^{(\mathrm{b})}_{\mathrm{intial}}$. The associated

keyline features $f^{(c)}$, $f^{(a)}$, $f^{(t)}$ and $f^{(b)}$ are computed using the keyline similarity formulae (Eqs. (5.9)-(5.12)). We denote these features as $f^{(c)}_{initial}$, $f^{(a)}_{initial}$, $f^{(t)}_{initial}$ and $f^{(b)}_{initial}$; values of these features for example cities in the typology data are shown in Table 5.3.

Table 5.3. Keyline feature values (obtained using the anchor texts) for example cities in the typology data

City name	$f^{(c)}_{initial}$	$f^{(a)}_{initial}$	$f^{(t)}_{initial}$	$f^{(b)}_{initial}$
Dhaka, Bangladesh	0.681	0.500	0.525	0.550
Dubai, United Arab Emirates	0.475	0.429	0.504	0.441
Amsterdam, Netherlands	0.533	0.493	0.585	0.776
Changchun, China	0.510	0.485	0.449	0.427

The typology classification models trained using these (anchor text based) keyline features serve as base models in our study. Note that since we have at most 282 labeled samples for traininga classifier, we cannot directly use a 768-dimensional representation of the Wikipedia page by simply averaging the vectors across lines in the Wikipedia page.

5.5. Results

5.5.1. Generated Candidate Keylines

Our proposed method considers using a set of representative keylines pertaining to a city feature (in addition to its anchor text). First, we obtain the candidate keylines related to each feature. Using the train data prepared for each typology classifier and the feature anchor texts ($anchor^{(c)}, anchor^{(a)}, anchor^{(t)}$, and $anchor^{(b)}$) we extract candidate keylines pertaining to congestion, auto, transit, and bike features using Eqs. (5.14)-(5.16) respectively.

The number of candidate keylines obtained for congestion, auto, transit, and bike features are 56, 45, 71, and 25 respectively; these numbers correspond to the typology distribution in the train set (e.g., for auto we have 45 positive samples with auto-heavy typology in the train set, and one candidate keyline is derived from the Wikipedia page corresponding to each of the positive samples). An example candidate auto keyline is '*many of these auto routes are frequently congested at rush hour'*; this is obtained from Montreal (Canada) Wikipedia page (having maximum similarity with the auto anchor text i.e., '*most people in the city use cars'*). The similarity scores of the candidate keylines (i.e., with anchor texts) range between 0.31 to 0.79. Our experiments are carried out on a computer with Intel i7 processor with 2 cores, 4 logical processors and 16 GB RAM. The computation time noticed for the above-mentioned sets of candidate keylines is 22 minutes for congestion, 25 minutes for auto, 28 minutes for transit, and 13 minutes for bike. Therefore, average

computation time for finding a keyline (indicative of a typology) on a Wikipedia page using our method is around 30 seconds.

Candidate keylines are able to effectively capture relevant signals indicative of the typologies from the Wikipedia pages; some variability is noticed in keylines with lower scores which may be attributed to the amount and type of information present in corresponding Wikipedia pages. The effect of such noise in candidate keylines on the typology prediction tasks is reduced with the use of optimal keyline sets derived from these candidate keylines (using our proposed feature keyline set expansion method). Having this set of representative keylines indicative of the typology casts a wider net for retrieving relevant signals from the city Wikipedia pages and reduces the dependency on a single keyline. We sort the candidate keylines (in descending order by similarity score) and store them as $sorted\_candidates^{(c)}$, $sorted\_candidates^{(a)}$, $sorted\_candidates^{(t)}$, and $sorted\_candidates^{(b)}$ for feature keyline set expansion.

5.5.2. Feature Keyline Set Expansion Results

The keyline set expansion method for a feature (as outlined in Algorithm 5.1) requires that at each expansion iteration (e), we expand the selected feature keyline set (step 3), compute associated feature using the expanded feature keyline set (step 4), then train and measure the classifier's performance based on the updated feature vectors (step 5). The number of iterations in the feature keyline set expansion algorithm depends on the number of feature candidate keylines in the train sets (considered in multiple runs of the 3-fold cross validation process. In other words, for a particular instantiation of the cross-validation step, the number of iterations depends on the number ofpositive samples in the train set (which is randomly sampled for each instance of cross-validation). Due to such variability, for Algorithm 5.1 we observed 30-33 iterations for congestion, 23-25 for auto, 38-40 for transit, and 16-18 for bike feature respectively.

At each $e(\geq 1)$, the percentage increment in the model performance metric (i.e., average validation AUC) is computed with respect to $e = 0$ (i.e., where the model trained with only anchor text based features $f_{initial}^{(c)}$, $f_{initial}^{(a)}$, $f_{initial}^{(t)}$ and $f_{initial}^{(b)}$. Figure 5.6 plots the performance metric increment graph with incremental expansion of the feature keyline set. As highlighted on the graphs in the figure, it is observed that the congestion typology classifier performs best at $e = 2$. This means, three congestion related keylines added from $sorted\_candidates^{(c)}$ (including $anchor^{(c)}$) in $\boldsymbol{K}_{opt}^{(c)}$, (as explained in Algorithm 5.1) is found optimal for representing the congestion feature $f^{(c)}$. Similarly, for auto, transit, and bike, best performances are noticed at $e = 14, 34$, and 11 respectively. The optimal feature keyline sets for auto, transit, and bike are denoted as $\boldsymbol{K}_{opt}^{(a)}, \boldsymbol{K}_{opt}^{(t)}$, and $\boldsymbol{K}_{opt}^{(b)}$ respectively; hence$|\boldsymbol{K}_{opt}^{(c)}|$= 3, $|\boldsymbol{K}_{opt}^{(a)}|$= 15, $|\boldsymbol{K}_{opt}^{(t)}|$ = 35, and $|\boldsymbol{K}_{opt}^{(b)}|$ = 12. Similarly, features computed based on respective optimal feature keyline sets are denoted as $f_{opt}^{(c)}$, $f_{opt}^{(a)}, f_{opt}^{(t)}$ and $f_{opt}^{(b)}$; as examples Table 5.4 provides values of these features for selected cities (as listed in Table 5.3).

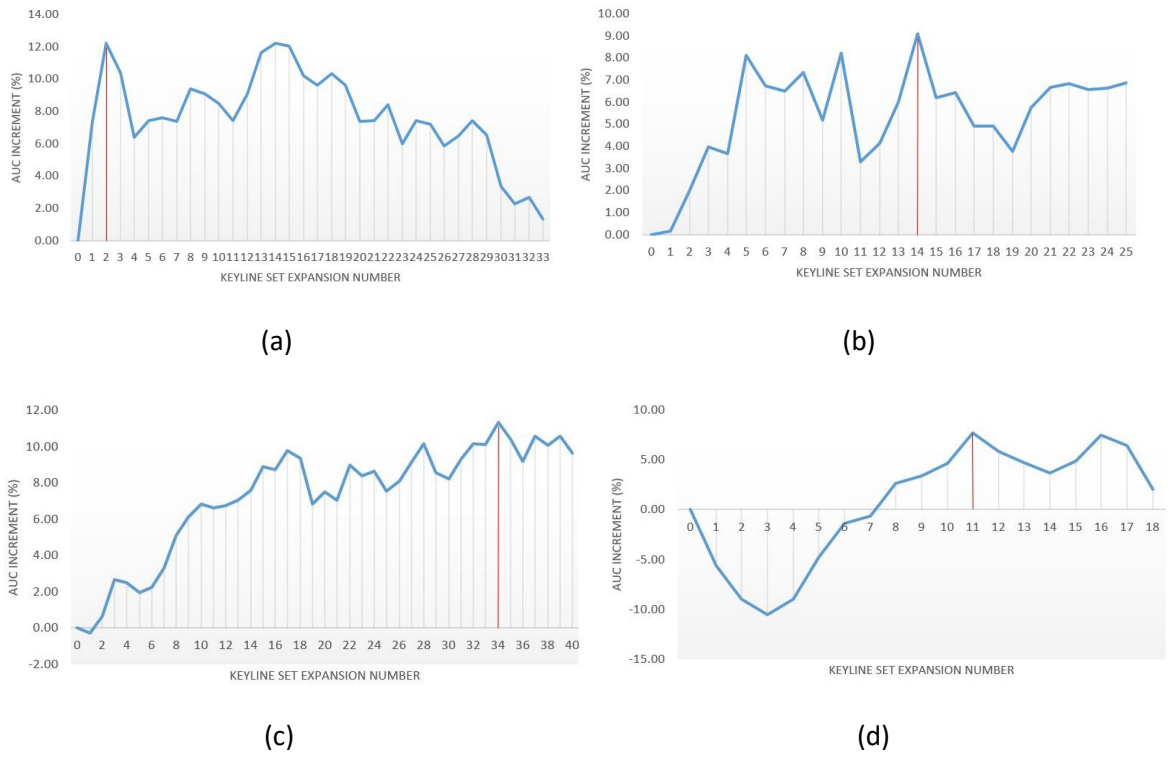


Figure 5.6. Optimal keyline set expansion for: (a) congestion, (b) auto, (c) transit, and (d) bike

Table 5.4. Keyline feature values (obtained using the optimal feature keyline sets) for example cities in the typology data

City name	$f_{opt}^{(c)}$	$f_{opt}^{(a)}$	$f_{opt}^{(t)}$	$f_{opt}^{(b)}$	City typology
Dhaka, Bangladesh	0.781	0.666	0.649	0.658	Congestion
Dubai, United Arab Emirates	0.552	0.716	0.621	0.615	Auto-heavy
Amsterdam, Netherlands	0.554	0.601	0.911	0.776	Transit-heavy
Changchun, China	0.510	0.671	0.609	0.717	Bike-friendly

Table 5.5 provides some examples of expanded keylines from each optimal feature keyline sets. The relevance of most of the expanded feature keylines in indicating the underlying meaning of the typology is worthy of note. For example, the transit keyline *'every major street in the city is served by at least one bus route*' indicates the extensive use of bus transit in the city. In addition, the textual diversity in the expanded feature keylines is a positive indication that our model can find relevant keylines given the anchor text despite the paraphrasing complexity in hundreds of Wikipedia pages (spanning multiple countries and human editors). Moreover, while processing lines from a city's Wikipedia page for predicting the typology, we compute the maximum similarity across a set of keylines; this limits the chances of incorrect inference even when a small fraction of keylines is noisy or not relevant.

Based on the graphs plotted in Figure 5.6, the model performance improves when a representative set of keylines (indicating the typology-based feature) is considered compared to anchor text alone, even though the additional keylines are based on the same underlying meaning as the anchor text. This is because various keylines that are extracted from city Wikipedia pages are not just semantically similar to the intuitive meaning of the typology, they represent city-based factors and elements associated with the typology as well (as can be seen in Table 5.5). Examples include inadequate city infrastructure leading to congestion, availability of sufficient freeways facilitating auto use, or emergence of app-based bicycle sharing services in popularizing cycling in the city. Having a representative set of feature keylines casts a wider net for retrieving useful signals from the city Wikipedia pages that indicate or represent the typology of interest. Moreover, addition of keylines beyond optimal expansion (highlighted in the graphs in Figure 5.6) does not necessarily improve the performance further; this is due to introduction of irrelevant or redundant lines in the keyline set.

Table 5.5. Examples of keylines (extracted from city Wikipedia pages) in the optimal feature keyline sets

	Auto keylines	Congestion keylines	Bike keylines	Transit keylines
Anchor text	Most people in the city use cars	The city has heavy traffic congestion	Many people in the cityuse bike or cycle	Most people in the city use public transit like bus and metro
Expanded keylines	The area has a number of freeways to transport people by car	Uncontrolled urban sprawl has challenged the city infrastructure, producing heavy traffic congestion	It is possible to cycle to most parts of the city	Almost half of all journeys in the metropolitan area are made on public transport
	Car sharing is available to residents of the city and some inner suburbs	Chronic traffic congestion, and a sudden and prolonged surge in crime have become perennial problems	Cycling has seen a resurgence in popularity due to the emergence of a large number of dock- less app-based bicycle sharing systems	Every major street in the city is served by at least one bus route

5.5.3. Typology Classification Results

We discuss the performance of the four typology classifiers developed in our study i.e., congestion, auto-heavy, transit-heavy and bike-friendly. Since we have multiple types of features (e.g., keyline features from

Wikipedia textual components and infobox numeric feature), we examine the model performance for various choices of the feature combinations. The input features used in the typology classification models constitute the following: features from Wikipedia textual component only ($f^{(c)}$, $f^{(a)}$, $f^{(t)}$ and $f^{(b)}$); features from Wikipedia numeric component only ($f^{density}$), and both textual and numeric components. In addition to these sets of input features, we use different combinations of feature keyline sets to define the model features. The keyline sets for congestion, auto, transit, and bike include: singleton sets with anchor texts only ($\mathbf{K}^{(c)}_{intial}$, $\mathbf{K}^{(a)}_{intial}$, $\mathbf{K}^{(t)}_{intial}$, and $\mathbf{K}^{(b)}_{intial}$); optimal sets where keylines are optimally expanded using keyline set expansion method as $\boldsymbol{K}^{(c)}_{opt}, \boldsymbol{K}^{(a)}_{opt}, \boldsymbol{K}^{(t)}_{opt}$, and $\boldsymbol{K}^{(b)}_{opt}$; and full sets where keylines are expanded to the fullest i.e., using all candidate keylines (we denote them as $\boldsymbol{K}^{(c)}_{all}, \boldsymbol{K}^{(a)}_{all}, \boldsymbol{K}^{(t)}_{all}$, and $\boldsymbol{K}^{(b)}_{all}$). The notations used for features are based on the notations of feature keyline sets. For example, the congestion feature $f^{(c)}$, computed using $\mathbf{K}^{(c)}_{intial}$, $\boldsymbol{K}^{(c)}_{opt}$, and $\boldsymbol{K}^{(c)}_{all}$, is denoted as $f^{(c)}_{initial}$, $f^{(c)}_{opt}$,, and $f^{(c)}_{all}$ respectively. Similar notations are used to denote auto, transit, and bike features.

Using multiple combinations of the above-mentioned features, we train classifiers (LR models) using the train set (197 cities) and we study the test set metrics (using 85 cities in the test set) for the choice of feature combinations. Based on the results of the typology classifiers on the test set, it is observed that compared to using only anchor text-based features, the models for congestion, auto-heavy, and bike-friendly typology prediction show significant improvement (56-80% lift in AUC scores) when features from both numeric and textual components are used, where the textual features are computed based on their optimal feature keylines. Congestion, auto-heavy, and bike-friendly classifiers have the highest test set AUC scores (0.85, 0.86, and 0.94 respectively) with features $f^{(c)}_{opt}, f^{(a)}_{opt}, f^{(t)}_{opt}$, $f^{(b)}_{opt}$ and $f^{density}$; for transit-heavy classifier, the highest test set AUC (0.61) is observed with features $f^{(t)}_{opt}, f^{(a)}_{initial}, f^{(c)}_{initial}$,and $f^{(b)}_{initial}$. The transit-heavy prediction model can be improved further with inclusion of additional city related features. This applies to other typology classification models as well, where further enhancing the model performance with extra features (from Wikipedia and/or other data sources) can be a promising future research direction.

For the models with the highest test set AUC scores, we get the input feature coefficients (including intercepts) and calculate additional performance metrics (*i.e.*, classification scores); values are reported in Table 5.6. Features with positive coefficient influence the probability of the event (typology prediction in our case) in a positive way and vice-versa. It is worth highlighting that in each typology classification model, features corresponding to the typology prediction have positive coefficients. For example, the auto-heavy classifier has a positive coefficient for $f^{(a)}$; this implies if there is indicative information in Wikipedia regarding high usage of automobiles in a city, there is a high chance that the city is predicted as an auto-heavy type city. Moreover, the population density feature in the congestion classification model has a positive coefficient, this justifies the relation between population growth and traffic congestion in cities.

Table 5.6. Feature coefficients and classification scores of the best performing city typology prediction models

Model (task)	Intercept	$f^{(c)}$	$f^{(a)}$	$f^{(t)}$	$f^{(b)}$	$f^{(density)}$	Accuracy	Precision	Recall	F1-Score
Congestion prediction	-0.687	0.116	-0.185	-0.274	-0.066	0.537	0.80	0.84	0.80	0.81
Auto-heavy prediction	-1.208	-0.001	0.047	-0.024	-0.013	-0.033	0.85	0.86	0.85	0.85
Transit-heavy prediction	-5.355	-0.021	-0.006	8.553	-2.836	0.000	0.62	0.62	0.62	0.62
Bike-friendly prediction	-1.913	-0.104	-0.019	-0.146	0.278	-0.190	0.76	0.92	0.76	0.81

It is interesting to note the reasonably high classification scores for congestion, auto-heavy, and bike-friendly typology prediction models reflecting their generalization capabilities; this is important when the models are beused on new and unseen data. However, we must also note that the performance of the model for transit-heavy typology prediction can be further enhanced with supplementary features supporting the typology, and additional data for training the model. Nonetheless, the classification results are rather encouraging indicating the effectiveness of Wikipedia as a data source for predicting city typologies.

5.5.4. Insights and Applications

The models developed for city typology predictions in our study can be applied to any city in the world whose details are available on Wikipedia. For the purpose of building the typology classification models, we use the Wikipedia data for 282 cities (including train and test set). Figure 5.7 shows the locations of these cities on the world map. To extend the analysis to other cities across the globe, we collect the list of cities (and urban towns) with 100,000 or more inhabitants from (Wikipedia, 2021) (many other cities are available on Wikipedia based on different criteria). We select around 2102 cities by web crawling the Wikipedia list of cities pages and fetch the city Wikipedia URLs. Using these URLs, for each city, we obtain the necessary data. Based on the input variables used in the best performing typology classification models, we compute the feature vector for these cities. To provide a sense of how a city typology study on limited samples can be propagated to a larger scale using our proposed method, Figure 5.8 shows the application of one of the typology classification models from our study. The figure includes ~ 2100 city locations (exclusive of the city data by Oke et al. (2019)) and the choropleth map of congestion probability scores obtained using the congestion classification model.

Figure 5.7. Cities selected from (Oke, et al., 2019) for developing the typology classification models in our study

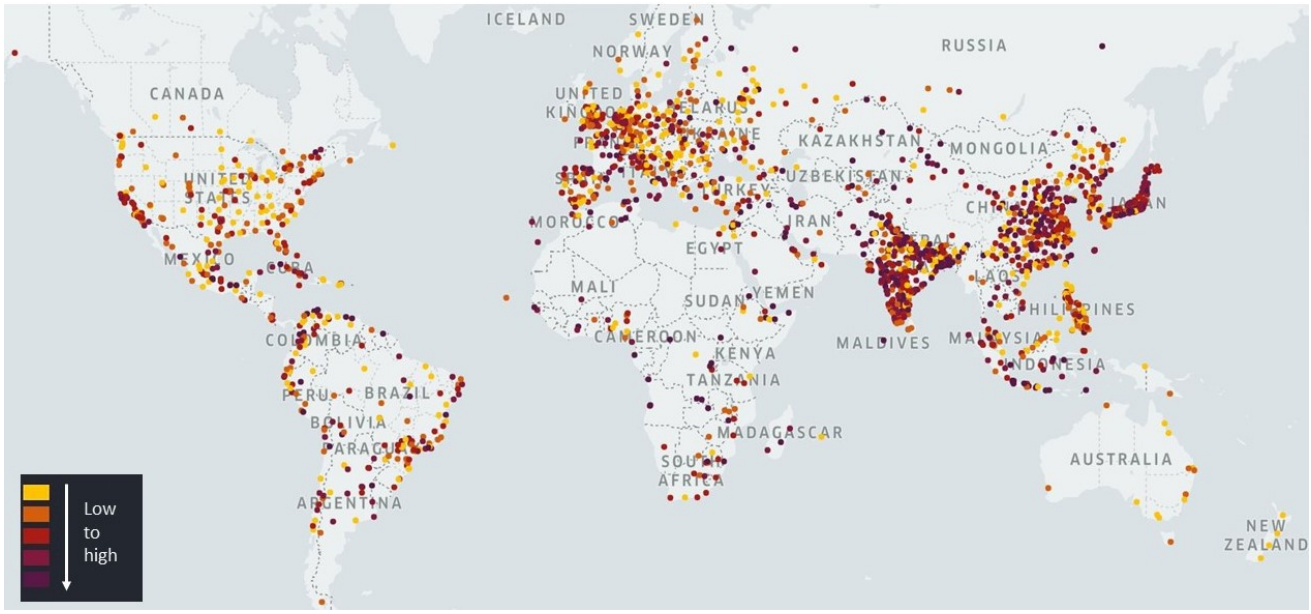


Figure 5.8. 2,102 cities from Wikipedia and their congestion probability scores, the sequential color palette represent low to high range of probability values

Wikipedia contains unprecedented volume and variety of information on different cities across the globe. Wikipedia not only contains detailed information on various aspects of cities, it also offers information on the intangible aspects such as people's preferences and opinions; the way people express these details vary based on locations, regions, and countries. To capture such variety of information pertaining to a specific typology, we present an iterative keyline expansion method that selects a representative set of keylines from city Wikipedia pages that allude to the city typology. The optimal feature keyline sets obtained using our proposed algorithm can be used to identify similar lines indicative of respective typologies for any city in the world that has a Wikipedia page with relevant transportation information in it.

Our methodology permits integration of information from textual and numeric components (on Wikipedia) in the typology prediction models and provides sufficient flexibility for expansion of the model feature vector allowing incorporation of additional variables. As our approach is based on understanding the intuitive meaning of the typology to consequently extract semantically similar and relevant textual information for defining city characteristics, it can be easily extended to different typology prediction tasks. As long as there is sufficient information pertainingto such typologies available in city Wikipedia pages, this holds true for both transportation-based typologies (*e.g.*, pedestrian-friendly or walkable cities, paratransit accommodating cities) and non-transportation-based ones as well (*e.g.*, climate-friendly cities).

6. Conclusion

Three dimensions of multi-modal transportation were explored in connection with ride sharing technologies. These were characteristics of connectivity, the role of human behavior, and new markets associated with food chains.

Connectivity. Multi-modal usage can be promising for the introduction of new modes of travel, such as ride sharing (National Academies of Sciences, Engineering, and Medicine 2021) that VIA provides. Proximity of the user to origins and destinations are important characteristics for connecting to other modes. Ride sharing modes can adopt other technologies such as electric vehicles and autonomous vehicles in a multimodal connectivity framework though conditions and constraints associated with those technologies exist.

Human Behavior. The attitudes and perceptions of users of multiple modes and their connections shape their behavior in the form of acceptance of such modes. The dimensions of behavior incorporate many factors ranging from views about safety and security, environmental compatibility, cost, comfort, and convenience (Zimmerman, 2019a). These can dramatically influence the viability of physical or technological characteristics of multiple modes.

New Markets. In normal times as well as in extreme events such as the pandemic and severe weather, pathways that define how food services connect food sources and consumers are strongly dependent on transportation systems (Zimmerman, 2021a, b). Ride sharing technologies can economize on those sections of the distribution system that bring raw materials to producers and finished products to consumers either by bringing consumers to the food sources or food to the consumers.

We propose a methodology to upscale data from the limited data available to microtransit operators (and to public agencies like the Federal Transit Administration in overseeing deployment regulations at the federal level). The method uses simulation to fit market equilibrium models to the limited data so that those models can be used to generate scenario data at low cost. The overall methodology contributes to the literature by parameterizing the within-day simulator from Yoon et al. (2021), extending the day-to-day market equilibrium model from Djavadian and Chow (2017a,b) to consider travelers with first/last mile access trips as well as direct trips, and providing a scenario generation algorithm for feeding the market equilibrium simulation model.

The new surrogate data proves to be useful; models fit to the data are adequately accurate compared to the original limited data set (CVs $\sim$ 45% for 4 observations) while illuminating statistically meaningful relationships between various public data with deployment portfolio measures like ridership and VMT.

Identifying the right cities to enter their markets requires having some understanding of the typology of these cities. City typologies based on the dynamics of mobility in cities can allow easy identification of comparable cities that can assist the decision-making process for emerging technologies like microtransit service deployment and planning. We propose a novel method for the utilization of Wikipedia articles on cities for a large-scale global city typology prediction (focusing on the transportation aspect of cities). Using data extracted from Wikipedia, we develop four typology classification models to predict congestion, auto-heavy, transit-heavy, and bike-friendly type cities (based on ground truth labels by Oke et al. (2019). We do so by algorithmically extracting lines from a city's Wikipedia page which semantically match a typology (via the SBERT NLP model) and use the typology-wise match scores to derive a low-dimensional keyline based feature vector (4-dimensional) representing the city. Using our proposed method, we demonstrate how a city typology classification based on limited samples (~300 cities) can be proliferated to an enormous scope spanning over 2000 cities across the world.

Our methodology permits integration of information from textual and numeric components (on Wikipedia) in the typology prediction models and provides sufficient flexibility for expansion of the model feature vector allowing incorporation of additional variables. Our study finds Wikipedia articles to be informative about transportation-based typology indicators. To the best of our knowledge, this is the first time the text-based information from Wikipedia articles is used as a data source for cities in this manner. This opens new opportunities for utilizing text-based data for transportation studies.

Our novel approach of using text-based information from Wikipedia for understanding citytypologies, and the outcomes presented in our study can assist a diverse group of stakeholders in transportation and urban planning fields. Additionally, we believe our method will reinforce existing studies utilizing crowd-sourced data leading to advances in strategic urban and transportation planning particularly in data-scarce regions of developing countries.

Future research should look at further collaboration with microtransit providers to understand the performance effects of city typologies and to focus more on empirically capturing good fitting forecast models using the proposed methodologies. Other emerging technologies should also be considered, especially where data are limited: e.g., planning electric vehicle fleets and charging infrastructure, pilots for autonomous vehicle fleets. A portfolio dashboard can be implemented to help a microtransit provider evaluate their portfolios and analyze alternative portfolio designs.

7. 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 7.1. Summary of research outputs

Output type	Description	Link/source
Paper	Rath, S., Liu, B., Yoon, G., Chow, J.Y.J., Microtransit deployment portfolio management using simulation-based data upscaling, submitted to 101st TRB Annual Meeting for presentation only.	Not available yet
Paper	Rath, S., Chow, J.Y.J., Worldwide city transport typology prediction with sentence-BERT based supervised learning via Wikipedia, submitted to 101st TRB Annual Meeting for presentation only.	Not available yet
Paper	R. Zimmerman, "Network-based Drivers of Technical and Social Innovations in Integrated Food, Water and Energy (FEW) Systems." Proceedings of the International Conference on Sustainable Development (ICSD) September 21-22, 2020, New York, NY: Columbia University Earth Institute and other collaborators.Posted November 2020.	https://ic-sd.org/wp-content/uploads/2020/11/Rae-Zimmerman.pdf
Paper	He, B. Y., Chow, J. Y. J., 2021. Entropy maximizing gravity model of passenger and mobility fleet origin-destination patterns with partially observed service data. *Transportation Research Record*, 2675(6), 235-253.	https://doi.org/10.1177/0361198121992074
Data	Calibrated simulation-based models for 6 cities	https://zenodo.org/record/5517983#.YUozC7hKg2w
Data	Estimated portfolio model and performance measures for two alternative portfolios	https://zenodo.org/record/5517983#.YUozC7hKg2w
Presentation	2021 C2SMART panel/webinar	https://www.youtube.com/watch?v=rA7T

		LsiXQ98
Presentation	2021 INFORMS Annual Meeting presentation	https://www.abstractsonline.com/pp8/?__hstc=194041586.9ad974a5999e3a9e202e99f21eba80a4.1598648681888.1630698234610.1630762393840.57&__hssc=194041586.1.1630762393840&__hsfp=2759698710&hsCtaTracking=76a3f7ff-51d5-4ec3-9afc-6681cc8dc243%7C1799fe6c-2007-47fc-9053-bd9abe03f130#!/10390/presentation/6213
Presentation	2022 TRANSED Microtransit Conference	Abstract submitted
Presentation	101st TRB Annual Meeting: Microtransit deployment portfolio management using simulation-based data upscaling	Paper submitted
Presentation	101st TRB Annual Meeting: Worldwide city transport typology prediction with sentence-BERT based supervised learning via Wikipedia	Paper submitted
Presentation	2021 ASCE Metropolitan Section Infrastructure Group June seminar: Flexible & Adaptable Infrastructure for a Post-Covid World	https://register.gotowebinar.com/register/3426197394343709198
Presentation	ASCE International Conference on Sustainable Infrastructure: Small Changes, Large Effects: Interconnected Infrastructure Networks in Food Supply Chain Disruptions and Multi-Modal Transportation Solutions	Abstract submitted and accepted for presentation
Presentation	Enlarging Infrastructure- based Public Services with Integrated Frameworks for Risk Reduction and Equity for Severe Weather, Climate Change and Pandemics	Presented at NYU Urban Initiative 2021 Urban Research Day, New York, NY on March 5, 2021

References

American Community Survey, 2019. Subject tables, United States Census Bureau. [Online] Available at: https://www.census.gov/acs/www/data/data-tables-and-tools/subject-tables/ [Accessed 29 July 2021].

Bliss, L., 2017. Bridj Is Dead, but Microtransit Isn't, *Bloomberg Citylab*.

Boylan, C., 2019. Saving money on Tesla model 3 charging with SmartcCharge NY — a hands-on review. CleanTechnica newsletter.

Calafiore, A. et al., 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems* 85, 101539.

Cantarella, G. & Cascetta, E., 1995. Dynamic processes and equilibrium in transportation networks: towards a unifying theory. *Transportation Science* 29(4), 305-329.

Caros, N. & Chow, J.Y.J., 2021. Day-to-day market evaluation of modular autonomous vehicle fleet operations with en-route transfers. *Transportmetrica B* 9(1), 109-133.

Cervero, R., 1998. *The transit metropolis: a global inquiry*. Island press.

Chatman, D. G., & Noland, R. B., 2011. Do public transport improvements increase agglomeration economies? A review of literature and an agenda for research, *Transport Reviews* 31(6), 725-742.

Chow, J.Y.J. & Djavadian, S., 2015. Activity-based market equilibrium for capacitated multimodal transport systems. *Transportation Research Part C* 59, 2-18.

Chow, J.Y.J., Golde, S. and Zimmerman, R. 2021. Understanding the Role of Microtransit in a Multi-modal Ecosystem, Webinar. New York, NY: NYU C2SMART. https://c2smart.engineering.nyu.edu/2021/03/24/understanding-the-role-of-microtransit-in-a-multi-modal-ecosystem/; https://www.youtube.com/watch?v=rA7TLsiXQ98

Chow, J.Y.J., Rath, S., Yoon, G., Scalise, P., Alanis Saenz, S., 2020. Spectrum of Public Transit Operations: From Fixed Route to Microtransit. FTA Report NY-2019-069-01-00. https://c2smart.engineering.nyu.edu/wp-content/uploads/2020/04/Chow-FTA-Report-NY-2019-069-01-00.pdf

Chow, J.Y.J., Regan, A., Ranaiefar, F. & Arkhipov, D., 2011. A network option portfolio management framework for adaptive transportation planning. *Transportation Research Part A* 45(8), 765-778.

Cich, G., Knapen, L., Maciejewski, M., Bellemans, T., & Janssens, D., 2017. Modeling demand responsive transport using SARL and MATSim. *Procedia Computer Science* 109, 1074-1079.

Cooper, R., Edgett, S. & Kleinschmidt, E., 1998. *Portfolio management for new products*. Addison Wesley Lgonman, Inc., Reading MA.

Cranshaw, J., Schwartz, R., Hong, J. & Sadeh, N., 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. *Proceedings of the 6th International AAAI Conference on Web and Social Media* 6(1), 58-65.

Creutzig, F., Baiocchi, G., Bierkandt, R., Pichler, P. P., & Seto, K. C., 2015. Global typology of urban energy use and potentials for an urbanization mitigation wedge. *Proceedings of the National Academy of Sciences* 112(20), 6283–6288.

Cronemberger, F., Gil-Garcia, J. R., Costa, F. X. & Pardo, T. A., 2018. Smart cities depictions in Wikipedia articles: reflections from a text analysis approach. *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, 560–567.

CTPP, 2016. Census Data for Transportation Planning Applications, AASHTO. [Online], Available at: https://ctpp.transportation.org/, [Accessed 29 July 2021].

Currie, G. & Fournier, N., 2020. Why most DRT/Micro-Transits fail–What the survivors tell us about progress. *Research in Transportation Economics* 83, 100895.

Daganzo, C. & Ouyang, Y., 2019. A general model of demand-responsive transportation services: From taxi to ridesharing to dial-a-ride. *Transportation Research Part B* 126, 213-224.

Derrible, S. & Kennedy, C., 2010. The complexity and robustness of metro networks. *PhysicaA: Statistical Mechanics and its Applications*, 389(17), 3678–3691.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1.

Djavadian, S. & Chow, J.Y.J., 2017a. Agent-based day-to-day adjustment process to evaluate dynamic flexible transport service policies. *Transportmetrica B* 5(3), 281-306.

Djavadian, S. & Chow, J.Y.J., 2017b. An agent-based day-to-day adjustment process for modeling 'Mobility as a Service' with a two-sided flexible transport market. *Transportation Research Part B* 104, 36-57.

EPA, 2021. Smart location database. [Online], Available at: https://www.epa.gov/smartgrowth/smart-location-mapping#SLD, [Accessed 30 July 2021].

Ferenchak, N. and Marshall, W.E., 2021. Bicycling facility inequalities and the causality dilemma with socioeconomic/sociodemographic change *Transportation Research Part D Transport and Environment* 97(3):102920.

Fielbaum, A., Jara-Diaz, S. & Gschwender, A., 2017. A parametric description of cities for the normative analysis of transport systems. *Networks and Spatial Economics*, 17(2), 343–365.

Guo, Z., 2011. Mind the map! The impact of transit maps on path choice in public transit. *Transportation Research Part A: Policy and Practice*, 45 (7), 625–639.

Haglund, N., Mladenović, M. N., Kujala, R., Weckström, C., & Saramäki, J., 2019. Where did Kutsuplus drive us? Ex post evaluation of on-demand micro-transit pilot in the Helsinki capital region. *Research in Transportation Business & Management*, 32, 100390.

Harris, C. D., 1943. A functional classification of cities in the United States. *Geographical Review* 33(1), 86-99.

Hasan, S., Zhan, X. & Ukkusuri, S. V., 2013. Understanding urban human activity and mobilit ypatterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pp. 1-8.

Horn, M., 2002. Fleet scheduling and dispatching for demand-responsive passenger services. *Transportation Research Part C: Emerging Technologies*, 10(1), 35-63.

Horowitz, J., 1984. The stability of stochastic equilibrium in a two-link transportation network. *Transportation Research Part B* 18(1), 13-28.

INRIX, 2019. Micromobility Potential in the U.S., UK and Germany.

Jones, K. S., 1974. Automatic indexing. *Journal of documentation* 30(4), 393-432.

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T., 2016. Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759.

Jung, J. & Chow, J.Y.J., 2019. Effects of charging infrastructure and non-electric taxi competition on electric taxi adoption incentives in New York City. *Transportation Research Record*, 2673(4), 262-274.

Kortum, K., 2018. TRANSPORTATION RESEARCH CIRCULAR E-C236 National Academies–TRB Forum on Preparing for Automated Vehicles and Shared Mobility. http://onlinepubs.trb.org/onlinepubs/circulars/ec236.pdf

Krok, A., 2016. Car2Stop: Car2Go shuts down services in San Diego, *CNET*.

Lenormand, M., Picornell, M., Cantú-Ros, O.G., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., San Miguel, M. and Ramasco, J.J., 2015. Comparing and modelling land use organization in cities. *Royal Society open science*, 2(12), 150449.

Lenormand, M. & Ramasco, J. J., 2016. Towards a better understanding of cities using mobility data. *Built Environment*, 42(3), 356–364.

Le, Q. & Mikolov, T., 2014. Distributed representations of sentences and documents. *International conference on machine learning*, 32, 1188-1196.

Litman, T., 2021a. Introduction to Multi-Modal Transportation Planning Principles and Practices, Victoria Transport Policy Institute. https://www.vtpi.org/multimodal_planning.pdf

Litman, T., 2021b. *New Mobilities: Smart Planning for Emerging Transportation Technologies, transportation.* Island Press.

Louail, T., Lenormand, M., Ros, O.G.C., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J. and Barthelemy, M., 2014. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(1), 1-12.

Louf, R. & Barthelemy, M., 2014. A typology of street patterns. *Journal of The Royal Society Interface*, 11(101), 20140924.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14), 281-297.

Mahmassani, H., 1990. Dynamic models of commuter behavior: Experimental investigation and application to the analysis of planned traffic disruptions. *Transportation Research Part A* 24(6), 465-484.

Mahmassani, H. & Chang, G., 1986. Experiments with departure time choice dynamics of urban commuters. *Transportation Research Part B* 20(4), 297-320.

Markov, I., Guglielmetti, R., Laumanns, M., Fernández-Antolín, A. and de Souza, R., 2021. Simulation-based design and analysis of on-demand mobility services. *Transportation Research Part A* 149, 170-205.

Marshall, A., 2019. Ford Axes Its Chariot Shuttles, Proves Mobility Is Hard., *Wired*.

Martí, P., Serrano-Estrada, L. & Nolasco-Cirugeda, A., 2019. Social media data: Challenges,opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, 74, 161–174.

Ma, T., Chow, J.Y.J., Klein, S. & Ma, Z., 2021. A user-operator assignment game with heterogeneous user groups for empirical evaluation of a microtransit service in Luxembourg. *Transportmetrica A* 17(4), 946-973.

Ma, T., Rasulkhani, S., Chow, J.Y.J. & Klein, S., 2019. A Dynamic Ridesharing Dispatch and Idle Vehicle Repositioning Strategy with Integrated Transit Transfers. *Transportation Research Part E* 128, 417–442.

Mihov, I. and Pangilinan, C., 2020. Towards a New Model of Public Transportation, *Uber*, https://d1nyezh1ys8wfo.cloudfront.net/static/PDFs/Transit+Horizons+vF.pdf

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Murphy, K. P., 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nalic, D., Mihalj, T., Bäumler, M., Lehmann, M., Eichberger, A. and Bernsteiner, 2020. Scenario Based Testing of Automated Driving Systems: A Literature Survey. Proc. FISITA Web Congr., 30.

National Academies of Sciences, Engineering, and Medicine, 2021. The Role of Transit, Shared Modes, and Public Policy in the New Mobility Landscape. Washington, DC: The National Academies Press. https://doi.org/10.17226/26053.

National Association of City Transportation Officials NACTO, 2019a. Shared Micromobility in the U.S. in 2018. https://.org/wp-content/uploads/2019/04/NACTO_Shared-Micromobility-in-2018_Web.pdf

National Association of City Transportation Officials NACTO, 2019b. Guidelines for Regulating Shared Micromobility https://nacto.org/wp-content/uploads/2019/09/NACTO_Shared_Micromobility_Guidelines_Web.pdf

Oke, J., Aboutaleb, Y.M., Akkinepally, A., Azevedo, C.L., Han, Y., Zegras, P.C., Ferreira, J. and Ben-Akiva, M.E., 2019. A novel global urban typology framework for sustainable mobility futures. *Environmental Research Letters*, 14(9), 095006.

OSM, 2021. Python OSMnx library (Open Street Map). [Online] Available at: https://github.com/gboeing/osmnx [Accessed 30 July 2021].

OTP, 2021. Open Trip Planner API. [Online] Available at: https://www.opentripplanner.org/, [Accessed 30 July 2021].

Pantelidis, T. P., Chow, J. Y. J. & Rasulkhani, S., 2020. A many-to-many assignment game and stable outcome algorithm to evaluate collaborative mobility-as-a-service platforms. *Transportation Research Part B: Methodological*, 140, 79-100.

Pennington, J., Socher, R. & Manning, C., 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, 1532–1543.

Pinto, H., Hyland, M., Mahmassani, H. & Verbas, I., 2020. Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Research Part C* 113, 2-20.

Priester, R., Kenworthy, J. & Wulfhorst, G., 2013. The diversity of megacities worldwide: Challenges for the future of mobility. *Megacity mobility culture*, Springer, 23–54.

Reid, R. 2021. How autonomous vehicles will change road designs. Civil Engineering magazine. https://www.asce.org/publications-and-news/civil-engineering-source/civil-engineering-magazine/article/2021/08/how-autonomous-vehicles-will-change-road-designs

Reimers, N. & Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Rocklage, E., Kraft, H., Karatas, A. & Seewig, J., 2017. Automated scenario generation for regression testing of autonomous vehicles. *IEEE 20th International conference on intelligent transportation systems*, 476-483.

Salton, G., 1968. *Automatic information organization and retrieval*. McGraw-Hill Computer Science Series, McGraw-Hill Book Co., New York.

Shaheen, S. & Chan, N., 2016. Mobility and the sharing economy: Potential to facilitate the first-and last-mile public transit connections. *Built Environment*, 42(4), pp. 573-588.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D. and Ermon, S., 2019. Predicting economic development using geolocated Wikipedia articles. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2698–2706.

Smith, M., Hazelton, M.L., Lo, H.K., Cantarella, G.E. and Watling, D.P., 2014. The long term behaviour of day-to-day traffic assignment models. Transportmetrica A 10(7), 647-660.

Smith, M. J., 1984. The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. *Transportation Science* 18(3), 245-252.

Smith, S., 2021. Via to Provide On-Demand Rides in Arlington, Texas.

Statista, 2019. Number of cities, towns and villages (incorporated places) in the United States in 2019, by population size. [Online] Available at: https://www.statista.com/statistics/241695/number-of-us-cities-towns-villages-by-population-size/, [Accessed 29 July 2021].

Thomson, J. M., 1978. *Great cities and their traffic*. Middlesex: Penguin Books.

Transdev, 2021. [Online] Available at: https://transdevna.com/services-and-modes/microtransit/ [Accessed 29 July 2021].

Transitfeeds, 2021. [Online] Available at: https://transitfeeds.com/ [Accessed 30 July 2021].

Tuncali, C., Fainekos, G., Ito, H. & Kapinski, J., 2018. Simulation-based adversarial test generation for autonomous vehicles with machine learning components. IEEE Intelligent Vehicles Symposium (IV), 1555-1562.

U.S. Department of Transportation, Bureau of Labor Statistics, 2019b Intermodal Passenger Connectivity Database (IPCD). Available at https://data-usdot.opendata.arcgis.com/ as of November 2019.

U.S. DOT BTS (2015) Passenger Travel Facts and Figures 2015, p. 12, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/PTFF_Complete.pdf

U.S. DOT BTS 2020 Transportation Statistics Annual Report 2020. Washington, DC. https://ntlrepository.blob.core.windows.net/lib/79000/79200/79277/TSAR_2020_Compressed_20210104.pdf

U.S. Department of Transportation, Bureau of Transportation Statistics, 2021. Intermodal Passenger Connectivity Database, available at www.bts.gov

U.S. Department of Transportation (DOT), Federal Highway Administration (FHWA), 2018. Summary of Travel Trends: 2017 National Household Travel Survey. https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf

U.S. Environmental Protection Agency, 2013. Our Built and Natural Environments: A Technical Review of the Interactions among Land Use, Transportation, and Environmental Quality, Second edition. Washington, D.C.: EPA http://www.epa.gov/dced/pdf/b-and-n/b-and-n-EPA-231K13001.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems,* NIPS, 6000–6010.

Via, 2021. [Online] Available at: https://ridewithvia.com/ [Accessed 29 July 2021].

Volinski, J., 2019. Microtransit or General Public Demand–Response Transit Services: State of the Practice. TCRP Synthesis of Transit Practice Project J-7, Volume Topic SB-30.

Wardman, M., 2004. Public transport values of time. *Transport policy*, 11(4), 363-377.

Watling, D. & Hazelton, M., 2003. The dynamics and equilibria of day-to-day assignment models. *Networks and Spatial Economics* 3(3), 349-370.

Wikimedia statistics, 2021. [Online] Available at: stats.wikimedia.org//#/all-projects [Accessed 30 July 2021].

Wikipedia API, 2014. Wikipedia API 0.5.4. [Online] Available at: https://pypi.org/project/Wikipedia-API/ [Accessed 30 July 2021].

Wikipedia, 2021. List of towns and cities with 100,000 or more inhabitants. [Online] Available at: en.wikipedia.org/wiki/List_of_towns_and_cities_with_100,000_or_more_inhabitants, 2021 [Accessed 30 July 2021].

Yan, X., Levine, J. & Zhao, X., 2019. Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. *Transportation Research Part C* 105, 683-696.

Yergin, D., 2021. The Major Problems Blocking America's Electric Car Future, *Politico.*

Yoon, G., Rath, S. & Chow, J., 2021. A simulation sandbox to compare fixed-route, flexible-route transit, and on-demand microtransit system designs. working paper.

Zhan, X., Ukkusuri, S. V. & F. Zhu, 2014. Inferring urban land use using large-scale socialmedia check-in data. *Networks and Spatial Economics,* 14(3), p. 647–667.

Zimmerman, R., 2019a. Human Behavioral Factors that Shape Urban Physical Infrastructure Services, in Proceedings from EDRA 50: Sustainable urban environments, edited by A. Beth, R. Wener, B. Yoon, R. A. Rae, and J. Morris. Brooklyn, NY: Environmental Design Research Association.

Zimmerman, R., 2019b. Challenges of Multimodal Networks Interconnecting Transportation Technologies: CAV and other Mode Innovations, University Transportation Research Center 2019 Transportation Technology Summit, New York, November 1, 2019.

Zimmerman, R., 2021a. Abrupt Transformations of Food Supply and Demand Networks and Interconnected Water and Energy in the COVID-19 Pandemic 2nd Food Energy Water Nexus conference. American Institute of Chemical Engineers (AIChE), the Institute for Sustainability (virtual).

Zimmerman, R., 2021b. Small Changes, Large Effects: Interconnected Infrastructure Networks in Food Supply Chain Disruptions and Multi-Modal Transportation Solutions, Presentation for the ASCE International Conference on Sustainable Infrastructure, December 2021 (virtual).

Zimmerman, R., Restrepo, C.E., Sellers, J., Amirapu, A., and Pearson, T. R., 2014. Promoting Transportation Flexibility in Extreme Events through Multi-Modal Connectivity, U.S. Department of Transportation Region II Urban Transportation Research Center, New York, NY: NYU-Wagner.

Zimmerman, R. and Sherman, M., 2011. To Leave An Area After Disaster: How Evacuees from the WTC Buildings Left the WTC Area Following the Attacks*, Risk Analysis*, 31(5), 787-804.

Zimmerman, R., Zhu, Q. and Dimitri, C., 2016. Promoting Resilience for Food, Energy and Water Interdependencies*, Journal of Environmental Studies and Sciences*, 6(1), 50-61.

Zimmerman, R., Zhu, Q. and Dimitri, C., 2018. A Network Framework for Dynamic Models of Urban Food, Energy and Water Systems (FEWS), *Journal of Environmental Progress & Sustainable Energy*, 37(1), 122-131.

Appendix A

The estimated feature coefficient values of the ridership and VMT forecast models are listed in Table A.1 and Table A.2 respectively.

Table A.1. Ridership forecast model estimated feature coefficient values

Feature	Estimated coefficient	Feature	Estimated coefficient	Feature	Estimated coefficient
Intercept	2.21E-01	mean income (dollars) x fix fare	3.07E-06	HH density x PP1	1.19E-01
mean income (dollars)	-2.90E-06	auto ownership per HH x street density	-1.69E-01	HH density x PP2	1.70E-01
auto ownership per HH	3.75E+00	auto ownership per HH x HH density	-4.05E-04	HH density x fix fare	-1.39E-01
street density	5.82E-02	auto ownership per HH x transit stop density	-2.04E-02	transit stop density x mean TRIPEQ	1.27E+02
HH density	8.83E-02	auto ownership per HH x employment density	4.17E-02	transit stop density x PP1	-2.76E+00
transit stop density	3.30E+01	auto ownership per HH x mean TRIPEQ	2.86E-01	transit stop density x PP2	9.87E+01
employment density	1.00E-01	auto ownership per HH x PP1	3.63E-01	transit stop density x fix fare	-2.04E+01
mean TRIPEQ	2.29E+00	auto ownership per HH x fix fare	-3.87E-01	employment density x mean TRIPEQ	6.09E-01
PP1	-1.14E+00	street density x HH density	-1.63E-02	employment density x PP1	-4.66E-02
PP2	2.34E-15	street density x transit stop density	-1.56E-01	employment density x PP2	4.41E-01
fix fare	-6.25E-02	street density x mean TRIPEQ	8.86E-02	employment density x fix fare	-7.92E-02
mean income (dollars) x street density	1.37E-07	street density x PP1	5.78E-02	mean TRIPEQ x PP1	-4.08E-01
mean income (dollars) x transit stop density	1.80E-04	street density x PP2	2.22E-03	mean TRIPEQ x PP2	1.05E-01
mean income (dollars) x employment density	3.88E-07	street density x fix fare	-2.70E-02	mean TRIPEQ x fix fare	-4.12E-02
mean income (dollars) x PP1	-1.50E-05	HH density x employment density	4.39E-03	PP1 x fix fare	-2.67E-02
mean income (dollars) x PP2	7.53E-07	HH density x mean TRIPEQ	6.54E-01		

Table A.2. VMT forecast model estimated feature coefficient values

Feature	Estimated coefficient	Feature	Estimated coefficient	Feature	Estimated coefficient
Intercept	-3.26E+00	mean income (dollars) x fix fare	1.78E-06	HH density x mean TRIPEQ	-1.10E+01
mean income (dollars)	2.86E-05	auto ownership per HH x street density	-7.74E-01	HH density x PP1	4.03E+00
auto ownership per HH	-1.62E+00	auto ownership per HH x HH density	-9.01E-01	HH density x PP2	9.38E-01
street density	2.71E-01	auto ownership per HH x transit stop density	-3.04E+02	HH density x fix fare	-5.65E-01
HH density	8.03E-01	auto ownership per HH x employment density	9.06E-01	transit stop density x employment density	-5.71E+01
transit stop density	-1.36E+01	auto ownership per HH x mean TRIPEQ	5.56E+01	transit stop density x mean TRIPEQ	1.29E+01
employment density	-6.39E-01	auto ownership per HH x PP1	4.48E+00	transit stop density x PP1	3.97E+02
mean TRIPEQ	-3.25E+01	auto ownership per HH x PP2	-5.53E-01	transit stop density x PP2	4.44E+02
PP1	-4.43E-01	auto ownership per HH x fix fare	-4.92E+00	transit stop density x fix fare	-6.47E+01
PP2	2.57E-15	street density x HH density	-7.76E-02	employment density x mean TRIPEQ	1.51E+00
fix fare	3.42E+00	street density x transit stop density	-5.68E+00	employment density x PP1	-7.87E-01
mean income (dollars) x auto ownership per HH	-2.15E-05	street density x employment density	8.80E-03	employment density x PP2	1.82E+00
mean income (dollars) x street density	5.17E-07	street density x mean TRIPEQ	1.67E+00	employment density x fix fare	2.50E-01
mean income (dollars) x HH density	9.61E-06	street density x PP1	2.23E-01	mean TRIPEQ x PP1	-6.83E+00
mean income (dollars) x transit stop density	7.90E-04	street density x PP2	4.61E-01	mean TRIPEQ x PP2	-2.44E+00
mean income (dollars) x employment density	-6.38E-06	street density x fix fare	-5.27E-02	mean TRIPEQ x fix fare	-5.29E-01
mean income (dollars) x mean TRIPEQ	-1.08E-04	HH density x transit stop density	1.16E+02	PP1 x fix fare	8.88E-01
mean income (dollars) x PP1	1.11E-05	HH density x employment density	1.94E-01	PP2 x fix fare	1.41E+00
mean income (dollars) x PP2	1.72E-05				