Sustainable and Healthy Communities through Integrating Mobility Simulations in the Urban Design Process

Center for Transportation, Environment, and Community Health
Final Report

by
Timur Dogan, Samitha Samaranayake

January 27, 2022

DISCLAIMER

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Sustainable and Healthy Communities through Integrating Mobility Simulations in the Urban Design Process		5. Report Date January 27, 2022 6. Performing Organization Code
7. Author(s) Timur Dogan (https://orcid.org/0000-0003-0749-8465) Samitha Samaranayake(https://orcid.org/0000-0002-5459-3898)		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Architecture, Cornell University, 129 Sibley Dome, Cornell University, Ithaca, NY 14853		10. Work Unit No. 11. Contract or Grant No. 69A3551747119
12. Sponsoring Agency Name and Address U.S. Department of Transportation 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report 10/01/2020– 09/30/2021 14. Sponsoring Agency Code US-DOT
15. Supplementary Notes		

16. Abstract

Rapid urbanization and new global construction estimated to be 250x NYC by 2050 is increasing traffic congestion, pollution, and related health threats. Thus, it is imperative that we develop new modeling capabilities that allow urban designers to quantify the performance of mobility solutions, sustainability, public health impacts, pedestrian thermal comfort and pollution exposure during the earliest stages of a design process. Embedded in a generative, performance-driven design process, such a tool can significantly facilitate the design of healthy and sustainable urban habitats that promote active mobility.

17. Key Words	18. Distribution Statement
Walkability, active mobility, urban design, computational design, modeling, software	Public Access: Conference Proceedings and Journal Paper. Free software downloadable from urbano.io

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

Form DOT F 1700.7 (8-69)

Abstract

Well-informed mobility-aware decision-making in spatial planning and urban design practice is imperative to reducing vehicle miles traveled (VMT) and advocating sustainable transportation modes. However, integrating transportation consideration into the early-stage design process remains challenging due to the lack of tools to inform designers about the mobility implications of their scenarios. This paper proposes a data-driven framework that utilizes open data sources including the travel survey data, census data, and POI data for modeling and simulating the spatiotemporal mode-distance choice distribution of urban home-based trips (i.e., trips originating from home). The framework has four main steps including (1) Population clustering, (2) Travel choice modeling, (3) Trip generation and simulation, and (4) Analysis. This framework, for the first time, allows urban designers and planners to efficiently quantify the amount of mode shift that can be elicited and the total VMT that can be reduced by different design scenarios. This methodology is generalizable and expandable due to its data-driven nature, which makes it a practically applicable tool to tackle varying design questions and be adapted to different geographical contexts. Through a proof-of-concept case study in California, we provide evidence that the dense and mixed-use urban development can significantly reduce VMT, advocate active transportation, and increase public transit ridership.

Keywords: mobility, simulation, urban planning, urban design

Introduction

By 2050, the global urban population is projected to grow by 2.5 billion, and roughly two-thirds of the world's population will live in urban areas (United Nations, Department of Economic and Social Affairs, and Population Division, 2018). This indicates a massive construction volume for new urban habitats and the continuous densification of current cities in the coming thirty years. Hence, urban mobility systems are at a crossroads where problems like congestion and pollution could be further aggravated with the rapid urbanization. The widely recognized objectives for future planning include reducing travel distances in cities and increasing accessibility to more sustainable urban transport solutions (UN-Habitat, 2013; Wegener *et al.*, 2017; Rupprecht Consult, 2019; Ceder, 2020). Many municipalities have started seeking integrative and holistic planning strategies to mitigate congestions, protect the environment, and enhance the quality of life. For example, California Transportation Plan (CTP) 2050 envisions reducing total vehicle miles traveled (VMT) by up to 27 percent and supporting a shift from 13 percent to 23 percent of all trips occurring by non-auto modes such as walking, biking, and public transit (Caltrans, 2021). Once achieved, it can significantly reduce the local transportation sector's greenhouse gas (GHG) emissions and enable vibrant, healthy communities. Moreover, based on new evidence from mobility research in the COVID-19 pandemic, shorter commute distances and more accessible active mobility choices can benefit urban resilience during the public health emergency response (Yang *et al.*, 2021).

Well-informed mobility-aware decision-making in spatial planning and urban design practice is imperative to reducing VMT and advocating sustainable transportation modes. Built environment factors, such as zoning regulations, placement of urban services, and density allocation, have fundamental impacts on human travel behavior (Stead and

Marshall, 2001; Grazi, van den Bergh and van Ommeren, 2008). Although these factors are found to have modest individual impacts, typically just a few percent of total travel, they are synergistic and therefore have a significant combined effect on transportation (Ewing and Cervero, 2010; Litman, 2021). Urban infrastructure and spatial patterns are much more costly to change once established than policies like parking pricing or gasoline taxing. This makes spatial planning and urban design approaches widely regarded as stable, long-term, and cost-effective ways to alleviate transportation congestion and emission (Yigitcanlar and Kamruzzaman, 2014; Ma *et al.*, 2018).

Despite the urgent need, integrating transportation consideration into the spatial planning and urban design process remains challenging due to the lack of tools to quantify the mobility implications of design scenarios. More specifically, there is currently no straightforward way to directly inform designers about the amount of mode shift that can be elicited and the total VMT that can be reduced by implementing their design ideas. State-of-the-art approaches in transportation and land use planning (Sadek *et al.*, 2011; Wegener, 2021), including activity-based model, modified four-step model, and integrated transport-land use models such as TRANUS, ILUTE, are not suitable for the early-stage design scenario testing because they try to model the equilibrium within the entire urban system and require detailed inputs such as road capacity, street hierarchy, building occupancy, speed limits, and travel costs. However, designers often do not have access to this detailed information, and collecting data or making assumptions for all of these inputs is challenging and unfeasible in the design phase. Also, calibrating and implementing these highly-specialized models is complex, demanding, and expensive. These limitations motivate research to develop simplified and targeted models for evaluating spatial planning and design strategies (Litman, 2021). In this regard, one type of more simplified and widely adopted framework in planning and design applications is indicator-based models. They use a group of engineered variables or composite scores to estimate, mostly with regressions, a certain dependent variable such as the percentage of one travel mode or the VMT (Moudon and Stewart, 2013; Lee, Jeong and Kim, 2016; Berhie and Haq, 2017). Common indicators include Space Syntax (Hillier *et al.*, 1976) and several D's (Cervero and Kockelman, 1997; Ewing and Cervero, 2010). However, there are debates around these conventional methods regarding their relatively low sensitivity to design and planning impacts, the influence of the analysis scale, and the lack of behavioral interpretability (Sadek *et al.*, 2011; Handy, 2018; Pafka, Dovey and Aschwanden, 2020).

Recent development in design-sensitive travel demand forecasting frameworks has two emergent trends. The first trend is accessibility-based, disaggregate-level simulation frameworks. They use address-level accessibility measures based on Points of Interest (POI) data to model trip-level choices and then aggregate them to reveal spatiotemporal patterns (Kou *et al.*, 2020; Sevtsuk, Basu and Chancey, 2021). The second trend is learning-based frameworks. They leverage machine learning techniques to simplify processes like data collection and travel behavior prediction (Aschwanden *et al.*, 2021). However, to our knowledge, there is no framework to date that can predict spatiotemporal patterns of multi-modal split and VMT simultaneously for any type of urban trip at fine spatial resolution. There are a few references in the choice modeling literature who studied the joint choice of travel mode and travel distance (Vega and Reynolds-Feighan, 2009; Ding *et al.*, 2014), but they were exclusively used for assessing policies like increasing car travel costs while minimal research efforts have been exerted to frame this joint choice question in spatial planning and urban design applications.

This paper proposes a data-driven framework that utilizes learning-based models to simulate the mode-distance choice distribution of urban home-based trips (i.e., trips originating from home). The output trips can be aggregated to derive accumulative modal split or VMT by high-resolution geographical units such as Census Block Group (CBG). This framework is sensitive to planning and design modifications such as urban densification, demographic changes, commercial development, or new public transportation construction. To showcase the workflow, we collect the training data and train a choice prediction model for California. We test the fidelity of the new predictive model on a held-out validation dataset and find that it has high accuracy in predicting the aggregate mode-distance distribution. We also apply the trained model in a proof-of-concept case study in the South Bay Region of Los Angeles County, which quantifies the mobility impacts of a neighborhood-scale urban renewal scenario by different development stages. We find our model adequately sensitive to capture the varying travel modal split and VMT produced by different stages. Results prove that a dense, mixed-use neighborhood development can entail considerably increased usage in modes of walking, biking, and public transit, as well as a significant reduction in VMT.

Overall, this research aims to equip planners and designers with the modeling and simulation capability that allows integrating mobility-related metrics into early-stage design consideration. The key innovation of this paper lies in (1) a simple yet effective way to parametrize urban accessibility features based on POI data and to use them in representing various types of planning and design modifications; (2) a learning-based approach to predict the joint choice of travel mode and distance level for individual trips so that the aggregate modal-split and VMT can be computed; (3) a data-driven framework that is generalizable and expandable since it is entirely driven by training data and do not require pre-defined expert-designed heuristics as in most traditional methods.

Methodology

Overview of the framework

The data-driven framework, as illustrated in Figure 1, has four main steps which can be divided into the modeling part and the simulation part. The modeling part aims to model travel behaviors by (1) Population clustering and (2) Travel choice modeling. These two steps need to be conducted only once for a region, such as a state like California, which can provide adequate trip data for training the models and whose population can be assumed to share similar travel profiles. The simulation part aims to produce the trips by (3) Trip generation and simulation, and compute mobility metrics through (4) Analysis. These two steps can be performed at a fine geographical level, such as CBG, for any area of interest within the modeled region.

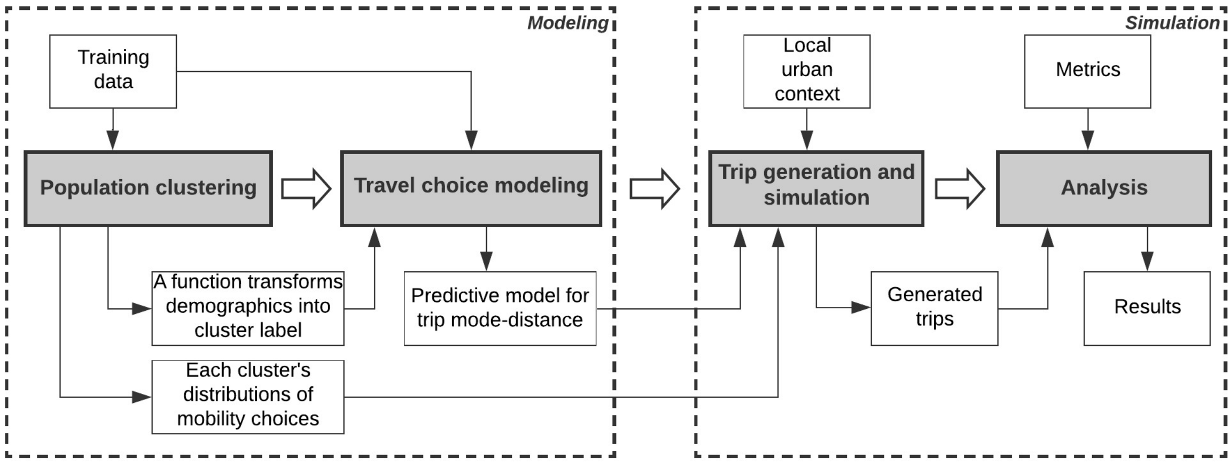

Figure 1. Data-driven modeling and simulation framework.

More specifically, (1) Population clustering step identifies population clusters of homogeneous mobility patterns. This step has two outputs that inform the rest of the framework including a function that transforms the demographic features into a specific cluster label, and each cluster's distributions of mobility choices. (2) Travel choice modeling step trains a learning-based classifier that predicts the individual trip's joint choice of travel mode and distance range. (3) Trip generation and simulation step takes inputs about the local urban context to produce trips in the simulated area. (4) Analysis step computes a series of metrics that help quantify the mobility performance.

Feature engineering and data preparation

Most of the required data is available publicly or can be computed using open data sources in the US. In this paper, we evaluate the proposed framework for California utilizing the following data sources:

- The 2017 National Household Travel Survey (NHTS) Add-On data for California (Caltrans, 2017). This dataset provides 94082 geocoded urban trip records of 23329 surveyed Californians. There are 12 other states in the US where such data is available (U.S. Bureau of Transportation Statistics, 2017);
- The 2018 American Community Survey (ACS) five-year estimates (US Census Bureau, 2018). This is a nationwide dataset that provides population information at various spatial levels.
- POI data from OpenStreetMap.org (OSM). This data provides the location of POIs and information about their service types as POI tags (OpenStreetMap, 2021).
- The 2017 Public Use Microdata Sample (PUMS) five-year estimates (US Census Bureau, 2017). This dataset provides disaggregate population samples with each individual's

demographic features. The geographical unit of sampling is called Public Use Microdata Area (PUMA) and there are around one thousand samples collected for each PUMA in California.

Table 1 describes all features used for the training process in terms of the feature type, discretization levels, and the data source. Demographic features ***(D)*** describe the traveler by age, occupation, and household characteristics. Time features ***(T)*** include season, weekday and weekend flags, and the time of day to describe differences in daily and seasonal mobility patterns. Activity features ***(A)*** explain the trip purpose by specifying the activity at origin and the activity at destination. Urban environment features ***(UE)*** describe the spatial context of the trip's origin location such as the population density and the accessibility to various urban services. In this paper, we use CBG as an example geographical unit to aggregate *UE* features. Mode-distance features ***(MD)*** describe the travel mode and the distance level of a trip. The trip count feature ***(TC)*** indicates the number of trips that the traveler makes during the time specified by ***T***. The final training dataset is a set of trip records from NHTS where each trip has all derived features as listed in Table 1.

Table 1. Description of features used for the training process.

Feature name	Type	Levels / Description	Source
Demographic *(D)*			
household vehicle	Categorical	not own vehicle; own vehicle	NHTS
age	Categorical	<= 17; 18 to 24; 25 to 64; >= 65	
household size	Categorical	1; 2; 3; 4; > 4	
household income	Categorical	< 50k; 50k to 75k; 75k to 100k; 100k to 125k; 125k to 150k; > 150k	
household workers	Categorical	0; 1; >1	
occupation	Categorical	sales / service; clerical / administrative / support; manufacturing / construction / maintenance / farming; professional / managerial / technical; student; other	
Time *(T)*			
season	Categorical	winter (Dec, Jan, Feb); spring (Mar, Apr, May); summer (Jun, Jul, Aug); fall (Sep, Oct, Nov)	NHTS
weekend	Categorical	weekend; weekday	
time of day	Categorical	early morning (12pm - 6am); morning (6am - 11am); noon (11am - 3pm); afternoon (3pm - 5pm); evening (5pm - 9pm); night (9pm - 12pm)	
Activity *(A)*			
activity at origin / destination	Categorical	home; work; school; shop / errands; social / recreational; meal; other	NHTS
Urban environment *(UE)*			
CBG school accessibility	Numerical	no. school within different distance levels (< 0.5 mile; 0.5 - 1 mile; 1 - 1.5 mile; 1.5 - 2 mile; 2 - 3 mile; 3 - 7 mile; > 7 mile) from the center point of the trip's origin CBG	OSM
CBG office accessibility	Numerical	no. office within different distance levels from the center point of the trip's origin CBG	
CBG shop/errands accessibility	Numerical	no. shop/errands within different distance levels from the center point of the trip's origin CBG	
CBG social/recreation accessibility	Numerical	no. social/recreation within different distance levels from the center point of the trip's origin CBG	

CBG meal accessibility	Numerical	no. meal within different distance levels from the center point of the trip's origin CBG	
CBG public transit accessibility	Numerical	no. public transit within different distance levels from the center point of the trip's origin CBG	
CBG population density	Numerical	total population/land area (km2) of the trip’s origin CBG	ACS
Mode-distance *(MD)*			
mode	Categorical	walking/biking; driving; public transit; other	NHTS
distance level	Categorical	< 0.5 mile; 0.5 - 1 mile; 1 - 1.5 mile; 1.5 - 2 mile; 2 - 3 mile; 3 - 7 mile; > 7 mile	
Trip count *(TC)*			
trip count	Categorical	0; 1; 2; 3; 4; 5 trips are made by the traveler during the time specified by ***T***.	NHTS

Table 2 shows how the POI tags in OSM are mapped to urban service types. The processes of accessing OSM data, mapping POI tags to service types, and deriving accessibility features are conducted in Grasshopper, Rhino3D (McNeel, R., & others, 2010) with the help of a mobility analysis toolkit called Urbano.io (Dogan *et al.*, 2020). The rest of the modeling and simulation processes are conducted in Python with the machine learning package scikit-learn (Pedregosa *et al.*, 2011).

Table 2. POI tags in OSM correspond to the urban service types.

Urban Service Type	**Corresponding POI Tags in OpenStreetMap**
school	amenity=school, university, college
office	office=*
shop/errands/services	æamenity=bank, pharmacy, post_box, atm, post_office, vending_machine, car_wash, marketplace, fuel, telephone, library, charging_station, bicycle_rental, veterinary, car_rental, driving_school, community_centre, ice_cream; shop=*
social/recreation	amenity=place_of_worship, bar, pub, townhall, social_facility, cinema, bbq, nightclub, arts_centre, shower, theatre; leisure=*
meal	amenity=restaurant, cafe, fast_food, food_court
public transit	public_transport=*

* all tags with the keyword.

Population clustering

Population clustering is a process that identifies groups of individuals with similar mobility patterns. Mobility patterns can be described as the tendencies to make various mobility choices such as mode-distance, activity, and trip count. Figure 2 illustrates the process where the population is clustered based on their different tendencies over time.

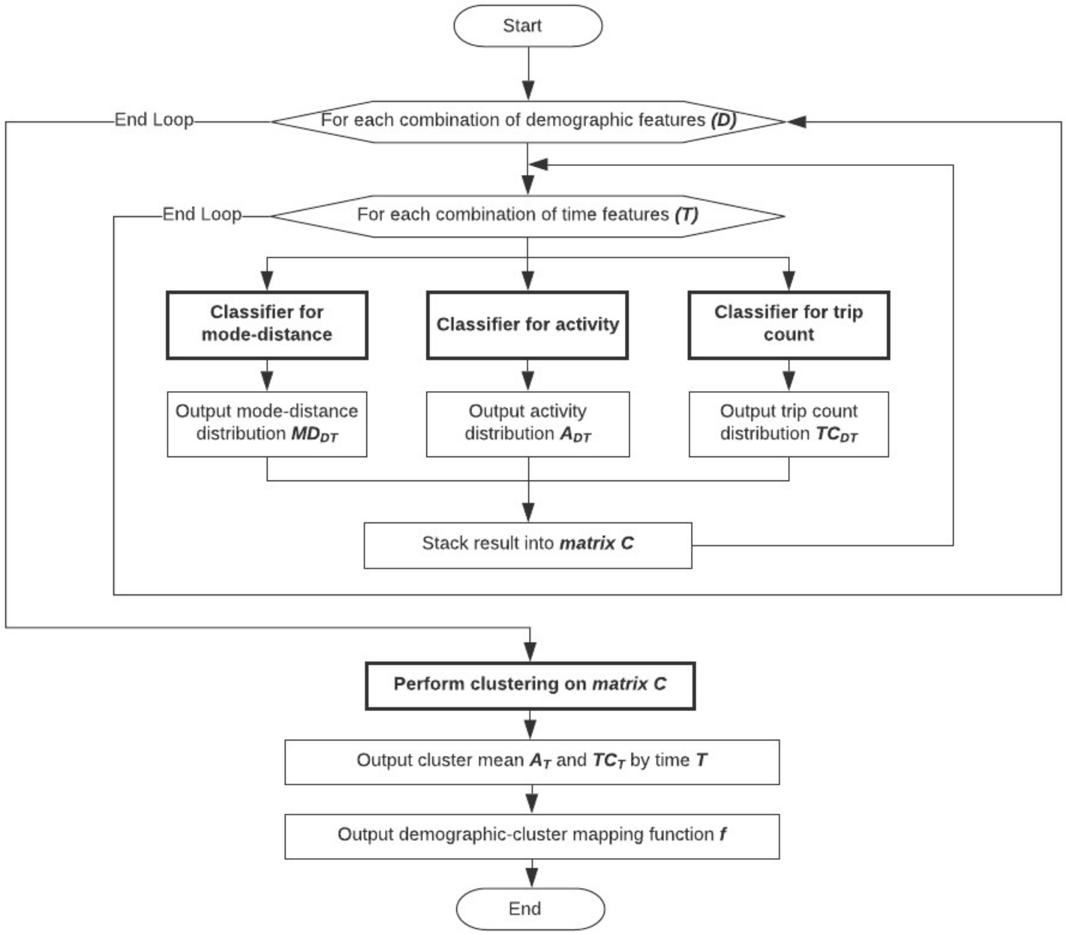

Figure 2. Flow chart of population clustering.

Three probability distributions are used to specify the tendencies to make mobility choices: mode-distance distribution ***MD*** (Equation 1), activity distribution ***A*** (Equation 2), and trip count distribution ***TC*** (Equation 3):

$$\begin{matrix} x_{m_1 d_1} & \cdots & x_{m_1 d_j} \\ & \ddots & \vdots \\ x_{m_i d_1} & \cdots & x_{m_i d_j} \end{matrix} \tag{1}$$

$$\begin{matrix} y_{o_1 d_1} & \cdots & y_{o_1 d_j} \\ & \ddots & \vdots \\ y_{o_i d_1} & \cdots & y_{o_i d_j} \end{matrix} \tag{2}$$

$$TC = (z_0 \cdots z_k) \tag{3}$$

where $x_{m_i d_j}$ denotes the probability of taking mode m_i and distance level d_j, $y_{o_i d_j}$ denotes the probability of having activity o_i at the origin and activity d_j at the destination, and z_k denotes the probability of making k trips. ***MD***, ***A***, and ***TC*** all have elements that sum to one. The modeled choice options for the mode, the distance level, the activity, and the trip count are listed in Table 1.

Three Random Forest classifiers are trained separately to predict ***MD***, ***A***, and ***TC*** for all possible combinations of demographic features ***D*** and time features ***T***. The predicted results are stacked into ***matrix C*** (Equation 4):

$$\begin{matrix} MD_{D_1T_1} & \cdots & MD_{D_1T_j}, A_{D_1T_1} & \cdots & A_{D_1T_j}, TC_{D_1T_1} & \cdots & TC_{D_1T_j} \\ & \ddots & \vdots \qquad \vdots & \ddots & \vdots \qquad \vdots & \ddots & \vdots \\ MD_{D_iT_1} & \cdots & MD_{D_iT_j}, A_{D_iT_1} & \cdots & A_{D_iT_j}, TC_{D_iT_1} & \cdots & TC_{D_iT_j} \end{matrix} \qquad (4)$$

where each row is a feature vector for one demographic group $\boldsymbol{D_i}$ across all time frames $\boldsymbol{T_j}$.

Then, the K-Means clustering algorithm is performed on ***matrix C*** where each row is matched to a cluster label. As a result, the cluster mean distribution $\boldsymbol{A_T}$ and $\boldsymbol{TC_T}$ by time ***T*** can be directly derived by averaging the corresponding elements in the clustered ***matrix C***. Also, by indexing the clustered ***matrix C***, a mapping function can be created as in Equation 5:

(5)

where any given demographic specified by features ***D*** can be mapped to a cluster label *c*.

Travel choice modeling

The predictive model for the trip mode-distance uses a Random Forest classifier that predicts the probability distribution of the mode-distance for a given trip. The output ***MD*** takes the same form as in Equation 1 where $x_{m_i d_j}$ denotes the probability that the predicted trip takes mode m_i and distance level d_j. Figure 3 shows the data flow of the proposed classifier. The time, activity, and urban environment features are directly input into the classifier while the demographic features go through the mapping function ***f*** and are translated into a population cluster label.

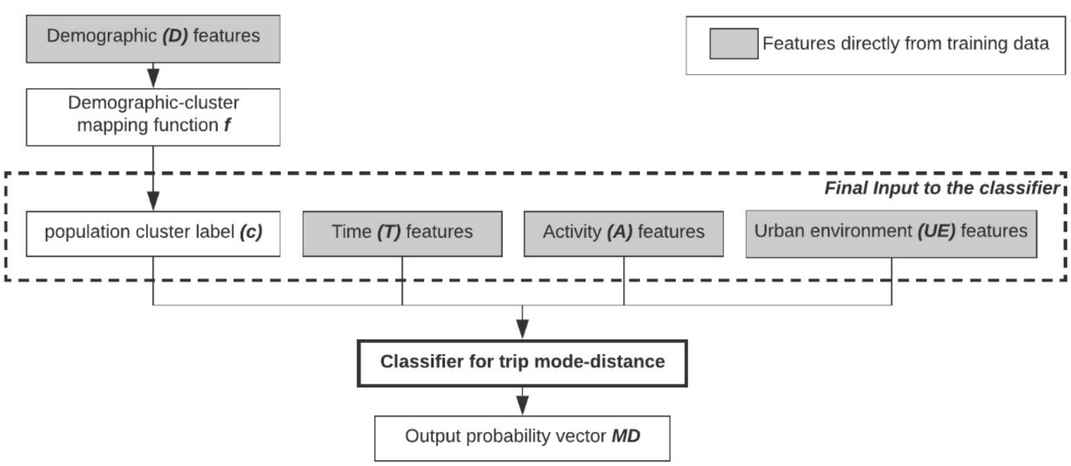

Figure 3. Input and output data of the trip mode-distance classifier.

Model assessment

To assess the clustering validity, we first examine how the derived clusters differ by their mobility patterns. This is achieved by comparing the mean activity distribution $\boldsymbol{A_T}$ and mean trip count distribution $\boldsymbol{TC_T}$, of all derived clusters. Besides, we also examine how the derived clusters differ by their demographics. The cluster demographic can be computed by mapping each sample from the PUMS data to a cluster label and deriving the cluster mean values of demographic features D. A successful clustering should produce clearly differentiated clusters that can recognize varying mobility patterns of different representative demographics.

To assess the effectiveness of the trip mode-distance classifier, we split the data where 70% is used for training and 30% is held out for validation. The validation approach is to compare the predicted $\boldsymbol{MD}$ distribution in the validation dataset, derived by averaging the predicted $\boldsymbol{MD}$ for all validation samples, against the true $\boldsymbol{MD}$ distribution in the validation dataset. Additionally, the feature importance of the Random Forest classifier is studied to provide insights about the most impactful features on the mode-distance choice of people.

Trip generation and simulation

Figure 4 illustrates the trip generation and simulation process which outputs a list of home-based trips by geographical units (e.g. CBG in this paper).

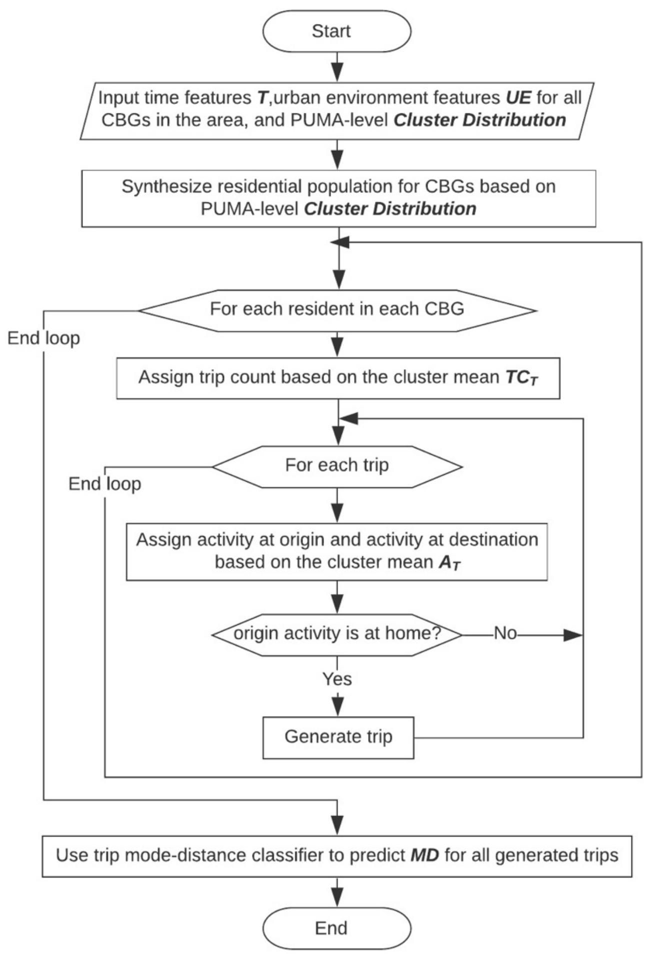

Figure 4. Flow chart of the simulation process.

The required simulation inputs include: (1) the user-defined time features ***T***. (2) the urban environment features ***UE*** for all CBGs in the area, which can be obtained in the same way as for the training data. (3) the newly defined PUMA-level ***Cluster Distribution*** as in Equation 6:

$$Cluster\ Distribution\ = (p_1\ , p_2\ \dots, p_c) \tag{6}$$

where p_c is the probability that a random resident in the PUMA belongs to cluster c. It can be derived by mapping each sample from the PUMS data to a cluster label and calculating the aggregate cluster distribution for the PUMA.

Given the simulation inputs, the first procedure is to synthesize the residential population for each CBG. It is a process of random sampling individuals according to the corresponding PUMA's ***Cluster Distribution*** with a sample size equivalent to the CBG's total population. All CBGs within a PUMA use the same ***Cluster Distribution*** for their population synthesis.

After the population synthesis, a sample enumeration process is conducted where the trips are iteratively generated by sampling the trip count for each resident and then sampling the activity for each trip. Note that these two sampling processes are purely based on the cluster mean distribution $\boldsymbol{TC_T}$ and $\boldsymbol{A_T}$. This is due to an assumption we made to simplify the framework that the choice of activities and trip count are only dependent on the traveler and the time, but independent of the urban environment.

It is also worth noting that only home-based trips are simulated because only these trips have a defined origin CBG. For non-home-based trips, the trip origin CBG is undefined and therefore the corresponding urban environment variables cannot be specified. One possible solution is to add a trip-sending process that assigns a specific destination for each trip and keeps track of the traveler's location across different time frames. However, this is beyond the scope of this paper.

Analysis

The predicted ***MD*** for simulated trips can be aggregated into statistics that describe the overall home-based travel behaviors. Assuming that the simulation generates K number of home-based trips for a CBG, P_m can be computed to show the aggregate percentage of trips choosing mode m as in Equation 6:

(6)

where x_{kmd_j} is the entry of $\boldsymbol{MD_k}$ for trip k that denotes the probability of this trip taking mode m and travel by distance level d_j.

To investigate the average travel distance per person by mode m, D_m can be computed as in Equation 7:

(7)

where $\boldsymbol{E}$ denotes the estimated value of distance for distance level j. It can be the median trip distance of level j, which is 0.25 miles, 0.75 miles, 1.25 miles, 1.75 miles, 2.45 miles, 4.45 miles, and 14.78 miles respectively for the distance level 1 to level 7 in our training dataset. When the mode m is set to automobile driving, D_m is equivalent to the average VMT per person.

Case study

The South Bay Region in Los Angeles County is used as a case study to demonstrate the simulation process and the analysis metrics. This is an area with the densest trip survey samples in California based on the spatial distribution of trip origins in the training data as shown in Figure 5a. Figure 5b shows the boundaries of CBGs and PUMAs covered in this area.

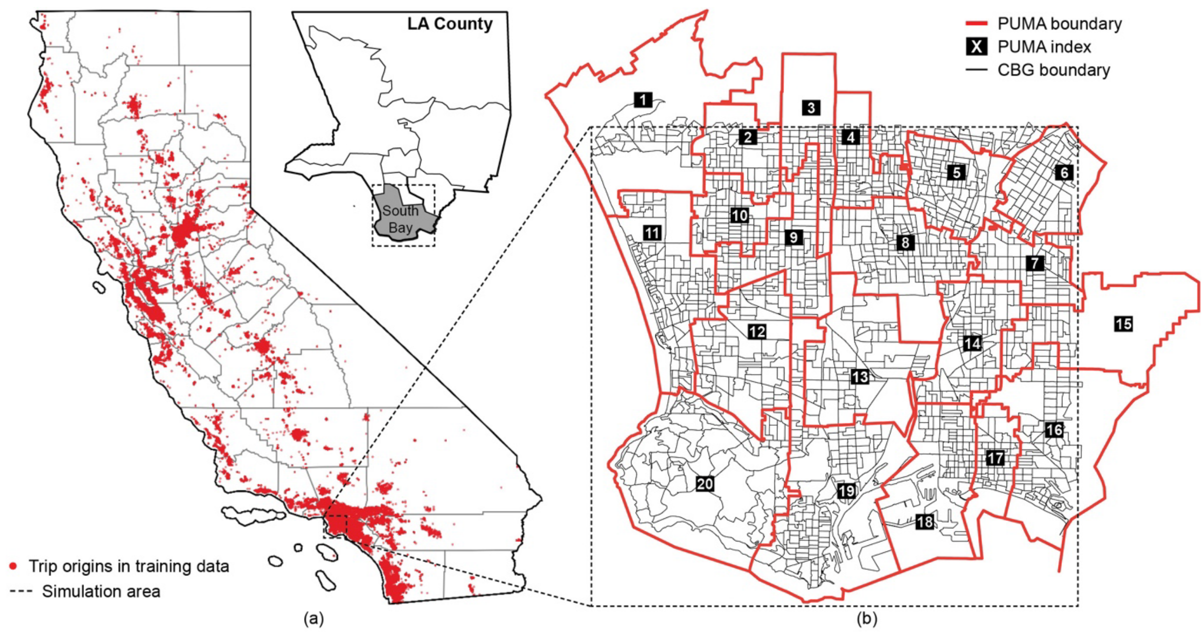

Figure 5. (a) California state map with trip origins in the training data. (b) Simulation area with PUMA and CBG boundaries.

The case study aims to investigate the mobility impact of a development scenario on a neighborhood covering 12 CBGs with approximately 17 square kilometers (Figure 6). The development is divided into three stages: (1) Population is densified by adding 1000 new residents in each CBG in the neighborhood (Figure 6b). It is assumed that the new population follows the same ***Cluster Distribution*** as the existing population, and there are no preference changes for all clusters. (2) Public transit accessibility is increased by introducing 250 new public transit services to the neighborhood and the surrounding area (Figure 6c). (3) Commercial developments are added. More specifically,10 shops, 10 social recreational services, and 10 meal service locations are added to the center location of each CBG in the neighborhood (Figure 6d).

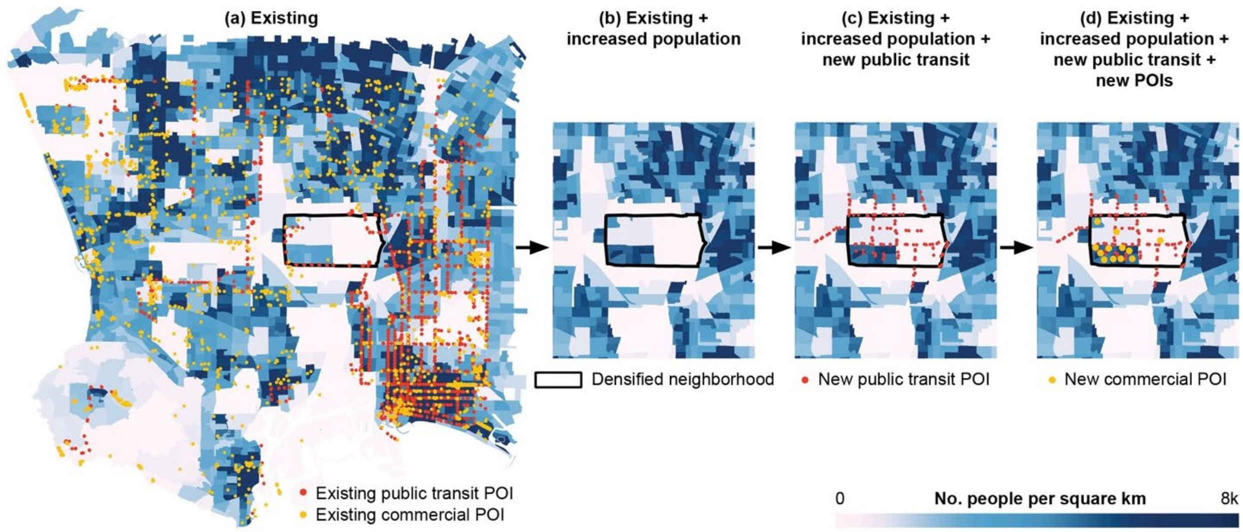

Figure 6. (a) Existing urban environment. (b) Development stage 1: the population is densified. (c) Development stage 2: public transit accessibility is increased. (d) Development stage 3: commercial developments are added.

By performing the simulation for a typical weekday (i.e. six time periods including early morning, morning, noon, afternoon, evening, and night) based on the existing environment and the different development stages, metrics like the daily average trip mode choice proportions ($\boldsymbol{P_{walking/biking}}$, $\boldsymbol{P_{public\ transit}}$, $\boldsymbol{P_{driving}}$) and the daily total VMT per person ($\boldsymbol{D_{driving}}$) can be derived and compared to study the mobility impacts of each development scenario.

Results

Assessment of population clustering result

We extract 12 population clusters from the clustering process (see the *Clustering* section of the *Supplementary Material*). Figure 7 shows the cluster mean trip count distributions $\boldsymbol{TC_T}$ and activity distributions $\boldsymbol{A_T}$ for three example time periods $\boldsymbol{T}$ (early morning, morning, and evening in winter weekdays). These distributions reveal the varying mobility patterns among specific clusters. For example, trip count distributions show that cluster #6, #7, #8, #12 are more inclined to travel during the early morning than other clusters. Activity distributions show that cluster #4, #5, #10, #11 have more tendencies to do school-related activities (e.g. from home to school or from school to home).

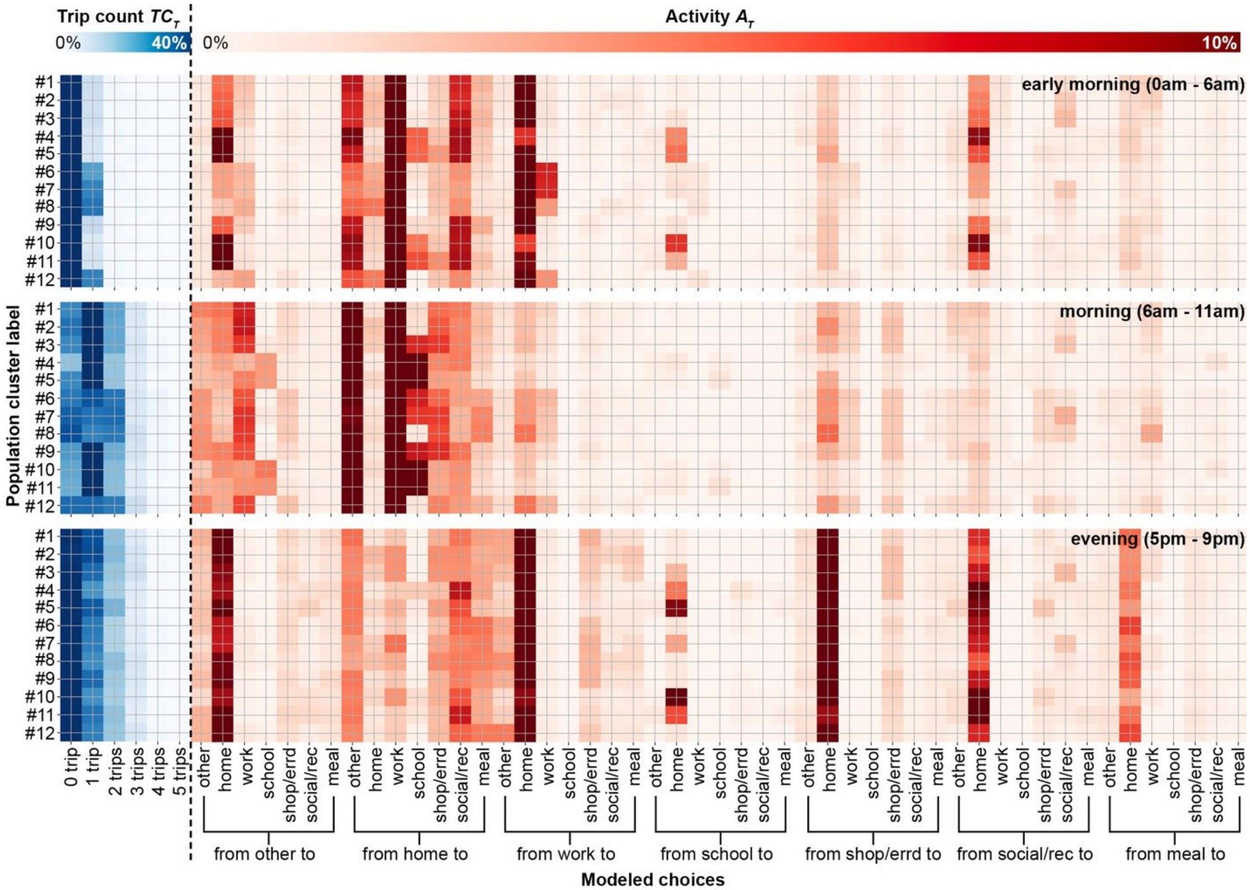

*Figure 7. Cluster mean trip count **TC_T** and activity **A_T** for three example time periods **T**: early morning, morning, and evening in winter weekdays. The darker color indicates the more probability that people from the cluster (labeled on the y axis) will make the corresponding mobility choice (labeled on the x-axis).*

By deriving and comparing the cluster mean demographic features as shown in Table 3, we can further relate the mobility patterns to specific demographics. For example, cluster #6, #7, #8, #12 consist predominantly of workers from the industry of manufacturing, construction, maintenance, and farming. Cluster #4, #5, #10, #11 are predominantly young people who are students. Further, the clusters are also clearly differentiated by household characteristics, including the size, vehicle ownership, workers count, and income. Based on the degree of dissimilarity among the derived clusters and the inherent consistency regarding their behavioral trends and demographics, it can be concluded that the proposed clustering approach effectively recognizes representative demographic groups with their distinct mobility patterns.

Table 3. Cluster mean demographic features.

Population Cluster label	Average age	Average household size	Average household vehicle count	Average household income	Average household workers count	Percentage of sales / service	Percentage of clerical / administrative / support	Percentage of manufacturing / construction / maintenance/ farming	Percentage of professional / managerial /	Percentage of student
#1	46.3	3.9	2.7	136213.6	2.0	0.3	0.2	0.0	0.2	0.0
#2	45.2	3.5	0.0	92388.3	1.7	0.3	0.2	0.0	0.1	0.0
#3	66.0	2.5	0.0	31164.3	0.0	0.0	0.0	0.0	0.0	0.0
#4	10.9	3.8	1.5	30902.3	0.0	0.0	0.0	0.0	0.0	1.0
#5	10.0	4.5	0.0	62963.1	1.5	0.0	0.0	0.0	0.0	1.0
#6	55.7	2.7	1.9	49938.0	0.0	0.0	0.0	1.0	0.0	0.0
#7	49.6	3.1	0.0	34642.2	0.0	0.0	0.0	1.0	0.0	0.0
#8	44.1	3.6	0.0	66156.0	1.8	0.0	0.0	1.0	0.0	0.0
#9	69.5	2.3	1.9	74370.1	0.0	0.0	0.0	0.0	0.0	0.0
#10	12.2	3.6	0.0	9831.6	0.0	0.0	0.0	0.0	0.0	1.0
#11	10.0	4.9	2.4	111502.6	1.8	0.0	0.0	0.0	0.0	1.0
#12	43.9	4.3	2.8	98096.1	2.2	0.0	0.0	1.0	0.0	0.0

Assessment of mode-distance classifier

After training the mode-distance classifier (see the *Classification* section of the *Supplementary Material*), we apply it to the validation dataset and compare the predicted and the actual distributions of trip mode-distance as shown in Figure 8. It demonstrates the aggregate prediction accuracy of the proposed choice prediction model.

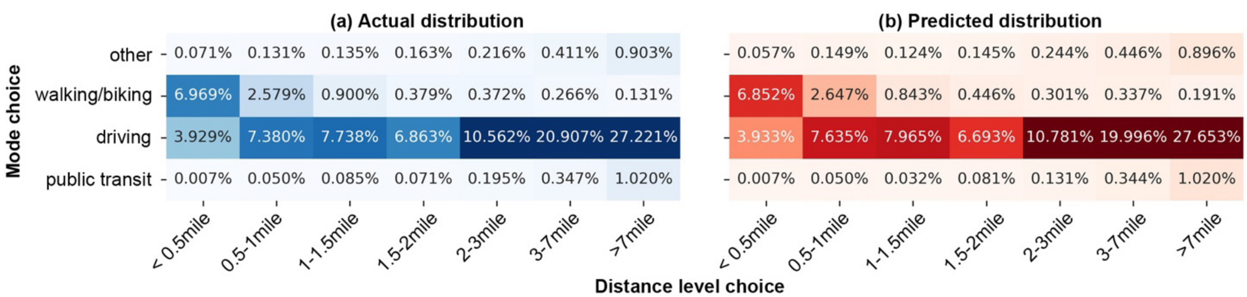

Figure 8. (a) Actual distribution of trip mode-distance in the validation dataset. (b) Predicted distribution.

To further investigate the most impactful factors for the choice of trip mode-distance, the feature importance ranking for the Random Forest classifier is plotted as in Figure 9. We find that all top-ranked features are related to the urban environment ***UE*** and activity ***A***. More specifically, among urban environment features, the CBG population density is the most significant feature. It is closely followed by features that describe the accessibility to the meal, social recreational, shop, and errand services within a half-mile range. Among activity features, the most determining activities are working, shopping, and running errands. Whether a trip starts from or goes to these activities influences the predicted mode-distance choice. These conclusions align well with numerous previous studies, which mention that population density has a statistically significant association with vehicle travel (Newman and Kenworthy, 2011), half-mile accessibility is closely associated with walking behavior (Handy, 2018), and various types of trips such as shopping, recreational, and commute trips show different travel patterns (Litman, 2021).

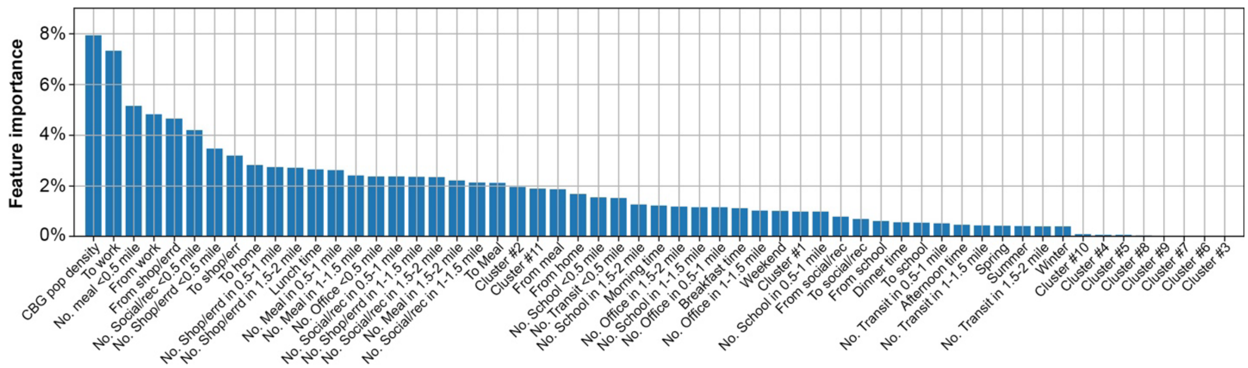

Figure 9. The feature importance ranking for the Random Forest trip mode-distance classifier.

Case study results

The ***Cluster Distributions*** of the 20 PUMAs in the simulation area are shown in Figure 10. It demonstrates varying demographics living in different areas. For example, PUMA 1, PUMA 12, and PUMA 20 have the highest proportion of cluster #1 population who have the highest average household income (Table 3). PUMA 4 has the highest cluster #11 population who are predominantly young students.

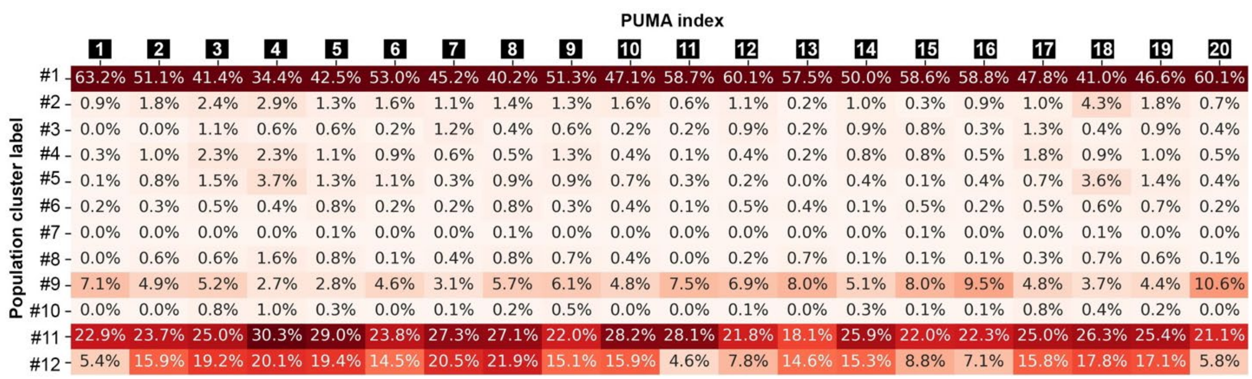

*Figure 10. PUMA-level **Cluster Distribution** in the simulation area.*

We first conduct the simulation process for a typical winter weekday. The derived CBG-level spatial distributions of daily average walking and biking trip proportion $\boldsymbol{P_{walking/biking}}$, driving trip proportion $\boldsymbol{P_{driving}}$, public transit trip proportion $\boldsymbol{P_{public\ transit}}$, and daily total VMT per person $\boldsymbol{D_{driving}}$ are respectively shown in Figure 11, Figure 12, Figure 13, and Figure 14. Note that there are no valid results for the CBGs without a residential population because there are no home-based trips generated for these CBGs.

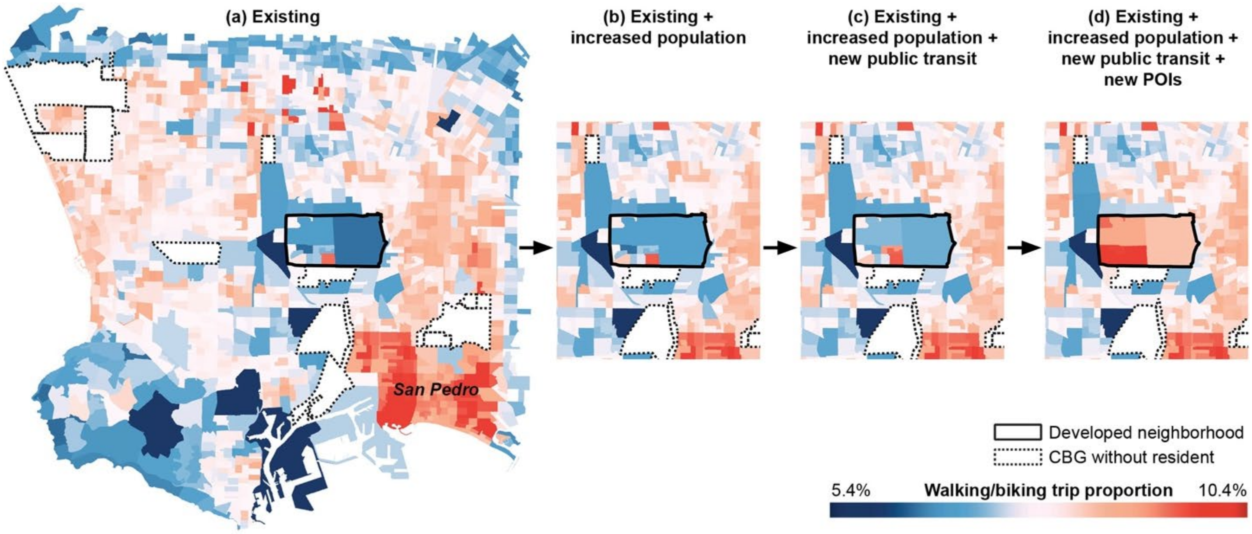

Figure 11. Simulated walking/biking trip proportion ***($P_{walking/biking}$)*** in the winter weekday.

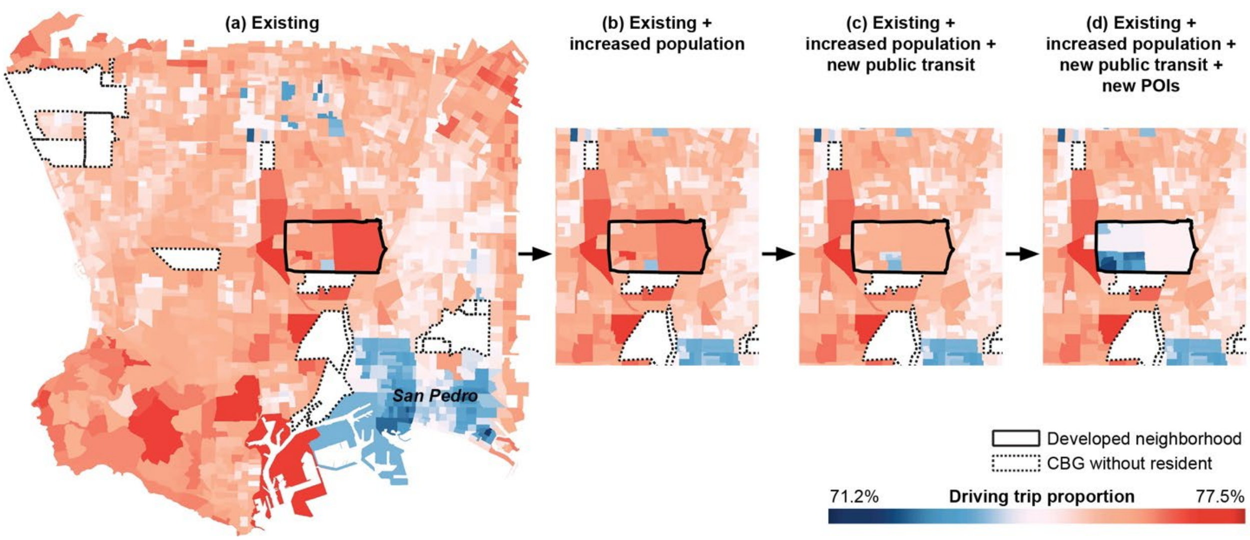

Figure 12. Simulated driving trip proportion ***($P_{driving}$)*** in the winter weekday.

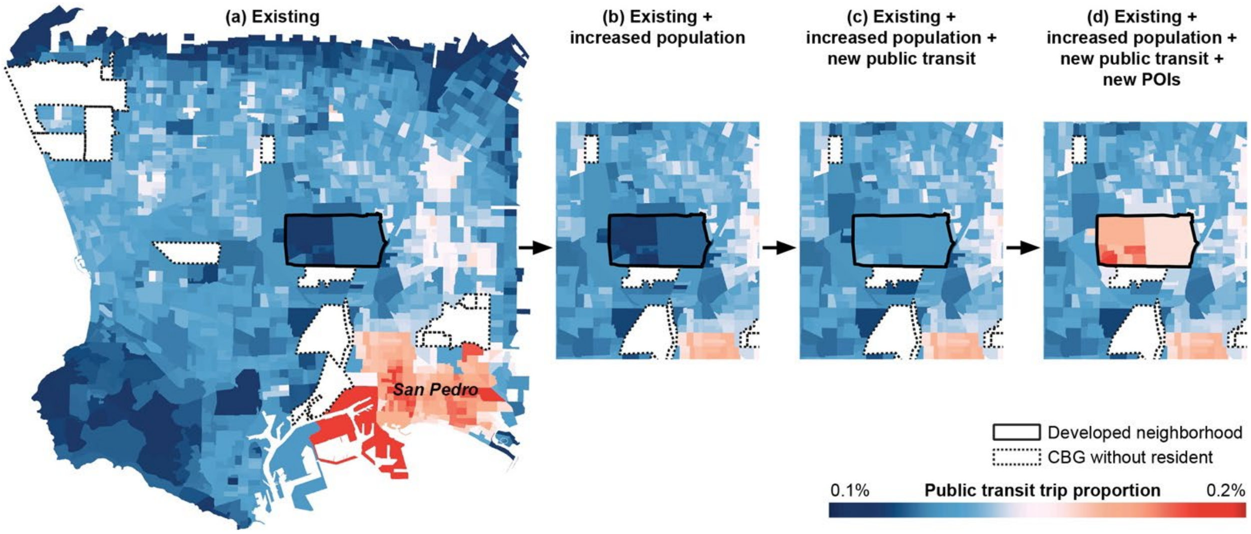

Figure 13. Simulated public transit trip proportion ***($P_{public\ transit}$)*** in the winter weekday.

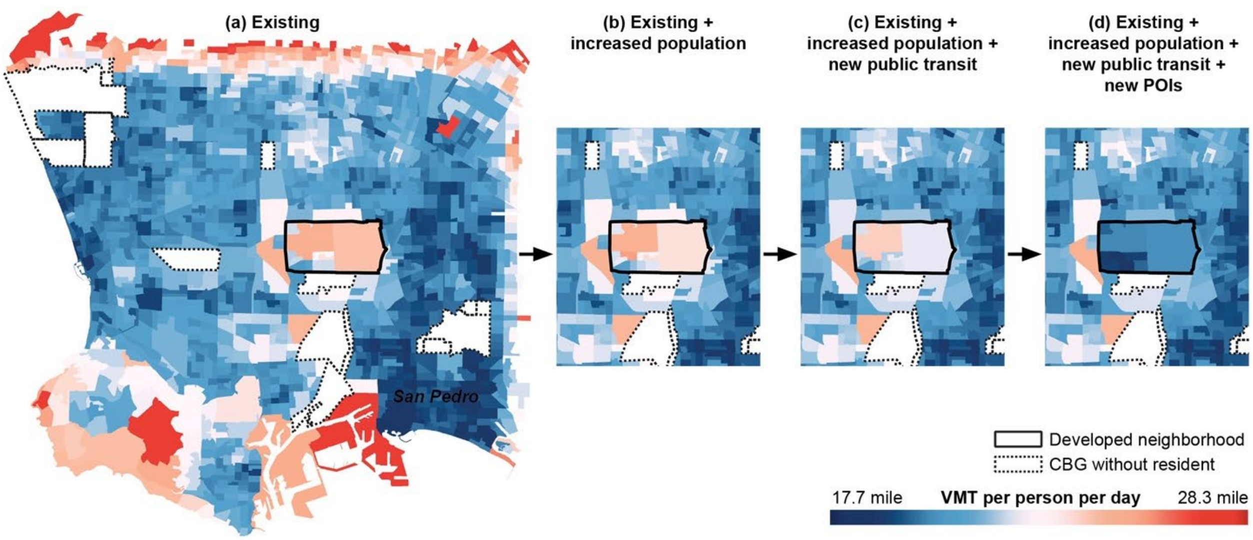

Figure 14. Simulated VMT per person per day ***($D_{driving}$)*** on the winter weekday.

Results show that the San Pedro area (south-east corner of the simulation area) produces the overall highest trip proportion by walking, biking, and public transit, as well as the least VMT per person. Comparatively, the neighborhood to be developed is one of the most automobile-dependent areas in the existing environment (see Figure 12a). By updating the urban environment in each development stage, the travel patterns in the developed neighborhood and its surrounding area are gradually changed and eventually reach a similar level as observed in the San Pedro area.

To further quantify the development impacts on local residents' home-based trips, Table 4 reports the average simulated metrics for the CBGs in the neighborhood. Two additional time frames - winter weekend and summer weekday - are also simulated so that temporal and seasonal differences can be observed. Results show that a large neighborhood-scale urban renewal program can significantly promote active transportation modes and reduce VMT. More specifically, after three development stages, the walking and biking trips are increased by 33 percent (from around 8.7% to around 11.6%), and the public transit trips are doubled (from around 0.6% to around 1.2%). The driving trips are decreased from around 88% to around 85%. The VMT is reduced by nearly 4 miles per person per day. Considering that the total population in the development neighborhood is around 22000, the total VMT reduction in this neighborhood is around 32 million per year. Regarding the temporal differences, results show that people generally travel less on weekends and in winters than weekdays and summers, which leads to slightly different magnitudes of travel impacts in different time frames. In addition, it is worth noting that in the first stage where only the population is densified and no new POIs are introduced, the estimated VMT per person is increasing since there are insufficient urban services available close by and the newly added residents still need to drive to distant destinations.

Table 4. Simulated metrics for the developed neighborhood.

Simulated time	**Metric**	**Existing**	**Existing+ increased population**	**Existing+ increased population+ new public transit**	**Existing+ increased population+ new public transit+ new POIs**
Winter weekday	$P_{walking/biking}$	8.72%	8.87%	9.44%	11.67%
	$P_{public\ transit}$	0.60%	0.60%	0.77%	1.32%
	$P_{driving}$	88.45%	88.31%	87.57%	85.30%
	VMT^{*}	22.81	22.94	22.33	19.24
Winter weekend	$P_{walking/biking}$	8.94%	9.06%	9.61%	11.91%
	$P_{public\ transit}$	0.42%	0.41%	0.58%	0.96%
	$P_{driving}$	88.59%	88.48%	87.77%	85.59%
	VMT^{*}	21.64	21.80	21.24	17.87
Summer weekday	$P_{walking/biking}$	8.64%	8.79%	9.44%	11.59%
	$P_{public\ transit}$	0.61%	0.61%	0.69%	1.24%
	$P_{driving}$	88.56%	88.41%	87.68%	85.46%
	VMT^{*}	23.03	23.20	22.60	19.36

* The unit of ***VMT*** is mile per person per day.

Discussion

Practical implications of case study results

The new neighborhood's reduced VMT and the changed travel modal split can be extrapolated into substantial economic, environmental, and societal gains. For example, the 32 million annual VMT reduction indicates approximately 117 million kilogram CO2 emission, 1 million gallons of gasoline consumption, and \$1.6 million pollution costs saved per year, assuming that CO2 emission per mile driven is 368.4 grams, gasoline consumption per mile driven is 0.041 gallon (Arkansas DEQ, 2008), and pollution costs per mile driven are 5 cents (Litman, 2010). The 33 percent increase in walking and biking trips indicates a reduction of individual pedestrian and biker injury risk by 16%, according to a correlation established by Jacobsen (2003) ($1 - 1.33^{-0.6} = 0.16$). Moreover, public health benefits can also be derived from the framework simulation results. Examples are the decrease of the obesity population (Frank, Andresen and Schmid, 2004) and medical costs associated with obesity (Edwards, 2008), which can be computed based on the projected daily miles traveled per person by walking and biking.

The case study further discovers that population densification alone is insufficient to reduce VMT because we find that the estimated VMT per person increases in the first development stage. Thus, a mixed-use development strategy is crucial to provide new residential developments with adequate services and accessible public transport. This finding provides evidence that supports popular theories of Walkable City (Speck, 2013) and Smart Growth (US EPA, 2006) which advocate compact and mixed-use urban development. The new simulation framework can be a practical tool to help planners and designers to make well-informed decisions regarding the optimal ways of densification.

The theoretical contribution of the framework

The primary theoretical contribution of this research is the development of a design-sensitive, multi-modal, and data-driven travel demand forecasting framework. To our knowledge, this framework is the first of its kind and has unique advantages that allow it to be implemented in planning and urban design applications.

Firstly, the framework provides a comprehensive yet straightforward approach to parametrize various types of design modifications in the built environment. The urban environment features ***UE*** can be used to represent spatial changes such as new constructions and rezoning. The spatial demography represented by the ***Cluster Distribution*** can be used to address anticipated population change. The case study provides an example of modeling densified population and new constructions of commercial and public transportation services. However, it is possible to investigate other design questions with this framework, such as (1) Comparing different service allocation scenarios and answering questions like which combination of the service types serves the neighborhood best, how many of them are needed, and where they should be placed; (2) Predicting the influence of increasing certain types of demographics in the local population, such as senior population when planning senior housings, young population when planning school districts or new dormitories, or workers of a certain industry when planning new industrial areas; (3) Analyzing a specific time frame for time-sensitive decision-making such as designing and planning for temporary urban events; (4) Analyzing a specific activity type for targeted design analysis such as home-school travel environment. Overall, the framework allows a significant degree of freedom for planners and urban designers to spatially, demographically, and temporally contextualize the mobility analysis for their projects.

Secondly, this framework uses a joint choice prediction model for trip mode-distance that allows predicting the multi-modal split and their distance-related metrics such as VMT. These metrics are normally challenging to compute and require dedicated data collection and expert modeling in traditional transportation models. We propose a simpler framework applicable for the preliminary scenario testing in planning and urban design practices, and its fidelity to planning and design-related questions has been demonstrated in the case study.

Further, the spatial resolution of the framework is theoretically unlimited and only constrained by data availability. The latitudes and longitudes of the geocoded trips in the survey data that we used for California are rounded to two decimal places, which indicates a precision level of around one kilometer. Thus, we engineered the features, calibrated the models, and conducted the simulation at the level of CBG, which has a similar spatial resolution. The modeling and simulation precision could be improved significantly with higher-resolution trip datasets accompanied with the ***UE*** features aggregated at a finer spatial resolution such as block or building level.

Lastly, the framework is generalizable and expandable due to its data-driven nature. The proposed framework can be deployed to any other location worldwide provided that the data is available. In contrast, expert-designed rules in traditional transportation models usually need greater efforts to be adjusted and calibrated to incorporate a changed geographical context. Also, if specific urban features are of interest to certain design questions but cannot be captured by the ***UE*** data used in this paper (e.g. availability of bike lanes, quality of street design, or microclimates in urban spaces), the data can be added to the training dataset as new feature columns, and the predictive model can be retrained to make it sensitive to these new design aspects. However, one potential hindering factor is the availability and quality of built environment data which may impair the reliability and sensitivity of the data-driven simulation. For example, the OSM dataset used in this paper may not contain all POIs available. Some design-related features, such as landscapes, microclimates, or infrastructures, can be challenging to derive because of the lack of high-quality, high-resolution data sources. Future studies based on this framework need to explore a more comprehensive set of ***UE*** features, test the sensitivity of the prediction, and provide practical guidance regarding variable selection.

The current limitation of the framework proposed in this paper is that only home-based trips are simulated because only home-based trips have a defined origin CBG and, therefore, can be matched to the corresponding ***UE***. However, the methodology and the case study results remain to be valuable and informative because home-based activities are the major type of daily trips. According to our NHTS trip data, home-based trips account for more than 33 percent of all trips while no other activity-based trips account for more than 16 percent. This makes the home-based travel behavior a significant indicator of the local mobility environment of the neighborhood. To enable holistic forecasting of both home-based and non-home-based trips, a trip-sending module and population location tracking over a series of timesteps are required. This is beyond the scope of this paper and is subject to future work.

Conclusion

This study provides a unique perspective in design-sensitive travel demand forecasting by developing a data-driven mode-distance choice prediction framework for home-based trips. This framework utilizes open data sources including the travel survey data, census data, and POI data for modeling and simulating the spatiotemporal travel behavior of urban residents. The framework is also generalizable and expandable due to its data-driven nature, which makes it a practically applicable tool to tackle varying design questions and be adapted to different geographical contexts. Through a proof-of-concept case study, we provide evidence that dense and mixed-use urban development can significantly reduce VMT, advocate active transportation, and increase public transit ridership.

Despite that the associations between the built environment and human travel behaviors have been extensively discussed in previous studies, this research, for the first time, allows urban designers and planners to efficiently test the changes in the aggregate modal-split and the reduction in VMT induced by different scenarios during the early-stage design process. We believe that our research can facilitate the co-design of mobility solutions and urban forms, which will empower the design and planning community in addressing the sustainability and livability issues in future urban mobility systems.

References

Arkansas DEQ (2008) *Cars and Air Pollution*. Available at: https://www.adeq.state.ar.us/air/planning/ozone/cars.aspx (Accessed: 28 March 2020).

Aschwanden, G.D. *et al.* (2021) 'Learning to walk: Modeling transportation mode choice distribution through neural networks', *Environment and Planning B: Urban Analytics and City Science*, 48(1), pp. 186–199. doi:10.1177/2399808319862571.

Berhie, G. and Haq, S. (2017) 'Land Use and Transport Mode choices: Space Syntax Analysis of American Cities', *Enquiry: A Journal for Architectural Research*, 14, pp. 1–22. doi:10.17831/enq:arcc.v14i1.429.

Caltrans (2017) *2017 National Household Travel Survey California Add-on*. Available at: https://nhts.dot.ca.gov/.

Caltrans (2021) *California Transportation Plan 2050*, p. 7. Available at: https://dot.ca.gov/programs/legislative-affairs/reports.

Ceder, A. (Avi) (2020) 'Urban mobility and public transport: future perspectives and review', *International Journal of Urban Sciences*, 0(0), pp. 1–25. doi:10.1080/12265934.2020.1799846.

Cervero, R. and Kockelman, K. (1997) 'Travel demand and the 3Ds: Density, diversity, and design', *Transportation Research Part D: Transport and Environment*, 2(3), pp. 199–219. doi:10.1016/S1361-9209(97)00009-6.

Ding, C. *et al.* (2014) 'Modeling the Joint Choice Decisions on Urban Shopping Destination and Travel-to-Shop Mode: A Comparative Study of Different Structures', *Discrete Dynamics in Nature and Society*, 2014, p. e492307. doi:10.1155/2014/492307.

Dogan, T. *et al.* (2020) 'Urbano: A Tool to Promote Active Mobility Modeling and Amenity Analysis in Urban Design', *Technology|Architecture + Design*, 4(1), pp. 92–105. doi:10.1080/24751448.2020.1705716.

Edwards, R.D. (2008) 'Public transit, obesity, and medical costs: Assessing the magnitudes', *Preventive Medicine*, 46(1), pp. 14–21. doi:10.1016/j.ypmed.2007.10.004.

Ewing, R. and Cervero, R. (2010) 'Travel and the Built Environment: A Meta-Analysis', *Journal of the American Planning Association*, 76(3), pp. 265–294. doi:10.1080/01944361003766766.

Frank, L.D., Andresen, M.A. and Schmid, T.L. (2004) ‘Obesity relationships with community design, physical activity, and time spent in cars’, *American Journal of Preventive Medicine*, 27(2), pp. 87–96. doi:10.1016/j.amepre.2004.04.011.

Grazi, F., van den Bergh, J.C.J.M. and van Ommeren, J.N. (2008) ‘An Empirical Analysis of Urban Form, Transport, and Global Warming’, *The Energy Journal*, 29(4), pp. 97–122.

Handy, S. (2018) ‘Enough with the “D’s” Already — Let’s Get Back to “A”’, p. 3.

Hillier, B. *et al.* (1976) ‘Space Syntax’, *Environment and Planning B: Planning and Design*, 3(2), pp. 147–185. doi:10.1068/b030147.

Jacobsen, P.L. (2003) ‘Safety in numbers: more walkers and bicyclists, safer walking and bicycling’, *Injury Prevention*, 9(3), pp. 205–209. doi:10.1136/ip.9.3.205.

Kou, Z. *et al.* (2020) ‘Quantifying greenhouse gas emissions reduction from bike share systems: a model considering real-world trips and transportation mode choice patterns’, *Resources, Conservation and Recycling*, 153, p. 104534. doi:10.1016/j.resconrec.2019.104534.

Lee, G., Jeong, Y. and Kim, S. (2016) ‘Impact of Individual Traits, Urban Form, and Urban Character on Selecting Cars as Transportation Mode using the Hierarchical Generalized Linear Model’, *Journal of Asian Architecture and Building Engineering*, 15(2), pp. 223–230. doi:10.3130/jaabe.15.223.

Litman, T. (2010) ‘Evaluating public transit benefits and costs: Best Practices Guidebook.’, *World Transit Research* [Preprint]. Available at: https://www.worldtransitresearch.info/research/2707.

Litman, T. (2021) ‘How Land Use Factors Affect Travel Behavior’, p. 91.

Ma, X. *et al.* (2018) ‘Joint Analysis of the Commuting Departure Time and Travel Mode Choice: Role of the Built Environment’, *Journal of Advanced Transportation*, 2018, p. e4540832. doi:10.1155/2018/4540832.

McNeel, R., & others (2010) *Rhinoceros 3D*. Seattle, WA: Robert McNeel & Associates. Available at: https://www.rhino3d.com/.

Moudon, A.V. and Stewart, O. (2013) ‘Tools for Estimating VMT Reductions from Built Environment Changes’, p. 36.

Newman, P. and Kenworthy, J. (2011) ‘“Peak Car Use”: Understanding the Demise of Automobile Dependence’, *World Transport Policy & Practice*, 17(2), pp. 31–42.

OpenStreetMap (2021) *OpenStreetMap Taginfo*. Available at: https://taginfo.openstreetmap.org/ (Accessed: 12 November 2021).

Pafka, E., Dovey, K. and Aschwanden, G. d. p. a. (2020) ‘Limits of space syntax for urban design: Axiality, scale and sinuosity’, *Environment and Planning B: Urban Analytics and City Science*, 47(3), pp. 508–522. doi:10.1177/2399808318786512.

Pedregosa, F. *et al.* (2011) ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12(85), pp. 2825–2830.

Rupprecht Consult (2019) *Guidelines for Developing and Implementing a Sustainable Urban Mobility Plan, Second Edition*. Available at: https://www.eltis.org/mobility-plans/sump-guidelines (Accessed: 24 June 2021).

Sadek, A.W. *et al.* (2011) ‘Reducing Vehicle Miles Traveled Through Smart Land-Use Design’. Available at: https://trid.trb.org/view/1136731 (Accessed: 10 November 2021).

Sevtsuk, A., Basu, R. and Chancey, B. (2021) ‘We shape our buildings, but do they then shape us? A longitudinal analysis of pedestrian flows and development activity in Melbourne’, *PloS One*, 16(9), p. e0257534. doi:10.1371/journal.pone.0257534.

Speck, J. (2013) *Walkable city: how downtown can save America, one step at a time*. First paperback edition. New York: North Point Press, a division of Farrar, Straus and Giroux.

Stead, D. and Marshall, S. (2001) ‘The Relationships between Urban Form and Travel Patterns. An International Review and Evaluation’, *European Journal of Transport and Infrastructure Research*, 1(2). doi:10.18757/ejtir.2001.1.2.3497.

UN-Habitat (ed.) (2013) *Planning and design for sustainable urban mobility: global report on human settlements 2013*. Abingdon, Oxon: Routledge.

United Nations, Department of Economic and Social Affairs, and Population Division (2018) *World urbanization prospects: the 2018 revision*.

U.S. Bureau of Transportation Statistics (2017) *2017 NHTS Add-on Partners*. Available at: https://www.bts.gov/table-3-2017-nhts-add-partners (Accessed: 28 October 2021).

US Census Bureau (2017) *American Community Survey (ACS), 5-Year Public Use Microdata Sample (PUMS)*, *The United States Census Bureau*. Available at: https://www.census.gov/programs-surveys/acs/microdata/access.html (Accessed: 9 August 2021).

US Census Bureau (2018) *2018 ACS 5-year Estimates*, *https://data.census.gov/ cedsci*.

US EPA (2006) *Smart Growth and Transportation*. Available at: https://www.epa.gov/smartgrowth/smart-growth-and-transportation (Accessed: 19 March 2020).

Vega, A. and Reynolds-Feighan, A. (2009) ‘A methodological framework for the study of residential location and travel-to-work mode choice under central and suburban employment destination patterns’, *Transportation Research Part A: Policy and Practice*, 43(4), pp. 401–419. doi:10.1016/j.tra.2008.11.011.

Wegener, M. (2021) ‘Land-Use Transport Interaction Models’, in Fischer, M.M. and Nijkamp, P. (eds) *Handbook of Regional Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 229–246. doi:10.1007/978-3-662-60723-7_41.

Wegener, S. *et al.* (2017) *Active Mobility – the New Health Trend in Smart Cities, or even More?*

Yang, Y. *et al.* (2021) ‘Urban design attributes and resilience: COVID-19 evidence from New York City’, *Buildings and Cities*, 2(1), pp. 618–636. doi:10.5334/bc.130.

Yigitcanlar, T. and Kamruzzaman, M. (2014) ‘Investigating the interplay between transport, land use and the environment: a review of the literature’, *International Journal of Environmental Science and Technology*, 11(8), pp. 2121–2132. doi:10.1007/s13762-014-0691-z.