Commuting Flow Prediction using OpenStreetMap Data

Kuldip Singh Atwal[1*], Taylor Anderson[1], Dieter Pfoser[1] and Andreas Züfle[2]

[1]George Mason University, Geography and Geoinformation Science, Fairfax, VA 22030, United States.
[2]Emory University, Department of Computer Science, Atlanta, GA 30322, United States.
[*]corresponding:katwal@gmu.edu.

Abstract

Accurately predicting commuting flows is crucial for sustainable urban planning and preventing disease spread due to human mobility. While recent advancements have produced effective models for predicting these recurrent flows, the existing methods rely on datasets exclusive to a few study areas, limiting the transferability to other locations. This research broadens the applicability of state-of-the-art commuting flow prediction models by employing features from freely accessible and globally available OpenStreetMap data. We show that the prediction accuracy of several state-of-the-art models using open data is comparable to location-specific and proprietary data. Our experiments indicate that consistent with theoretical and analytical models, building types, distance, and population are the determining characteristics for mobility related to commuting. Furthermore, our experiments show that predicted flows closely match ground truth flows. It helps establish the practical relevance of flow prediction models for real-world applications such as urban planning and epidemiology.

Introduction

Understanding how individuals routinely move from one place to another is as challenging as it is significant [1, 2]. Commuting flow prediction estimates the number of people moving between regions in a geographic area based on descriptive features, such as population [3], distance to other locations [4], and land use type [5]. Commuting flow prediction is helpful in many applications, such as understanding migration patterns [6, 7], urban planning [8, 9], and epidemiology [10, 11]. Considering that commuting flows vary little from day to day [12, 13], the goal is typically to predict a set of static flows where each flow represents the average number of daily commuters between origin-destination pairs, i.e., home and work locations [14, 15]. Therefore, similar to other approaches [9, 16], we define the term flow prediction as the task of predicting repetitive static flows rather than forecasting flows along a series of points in time using historical data, which is a time series problem.

Analytical flow prediction approaches include spatial interaction models such as the gravity model [17] and its extensions, including the radiation model [18–20], the intervening opportunities model [21, 22], and the competing migrants model [23]. Each model proposes different characteristics to predict accurate flows. For example, the gravity model assumes that the flow between locations is a function of two main characteristics: (i) the population at both locations and (ii) the distance between them. In another example, the intervening opportunities model replaces distance with the number of opportunities at the destination location that satisfy the trip objective [24]. Thus, when predicting commuter flows, the "opportunity" in question might be the number of commercial businesses.

More recently, machine learning models for commuting flow prediction far outperform the traditional mathematical approaches when comparing the predicted flows with ground truth [16, 25–28]. These models leverage machine learning approaches that can more flexibly incorporate different features of the origin-destination and can capture complex and non-linear relationships in the data [29–31]. Many studies use spatiotemporal characteristics to address the flow prediction problem using neural networks [32–35], which can also be combined with ordinary differential equations [36]. A current state-of-the-art model, the Geo-contextual Multitask Embedding Learner (GMEL) [9] learns commuting flows based on origin-destination features and their spatial contexts. GMEL uses 65 features derived from the 2015 NYC Primary Land Use Tax Lot Output (PLUTO)[37] dataset. In another example, the ConvGCN-RF model [38] uses convolutional neural network, graph convolutional network, and a random forest regressor to predict the commuting flow based on origin-destination features related to land use, as well as the residential and working population for homogeneous spatial units in the region of Beijing, China. Spadon et al. [39] derive 22 urban features from datasets provided by the Brazilian Institute of Geography and Statistics (IBGE) to predict intercity commuting in Brazil.

Despite the ability of such models to accurately predict flows, these high-performing models use a large number of input features derived from location-specific data sets that are not available outside of the study area. It makes the use of the model in other data-poor study regions challenging. In addition, given the variety of different input features used across models, it is difficult to compare models independent of the used data.

Our goal in this research is to address the limitations that restrict the applicability of current commuting flow prediction models to arbitrary study areas. More precisely, we assess the effectiveness of these models by employing a minimal set of input features obtained from a globally accessible dataset called OpenStreetMap (OSM) [40]. Moreover, since numerous models are assessed using high-level metrics, such as Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Common Part of Commuters (CPC), which provide limited insight into the model's ability to replicate authentic patterns intrinsic to commuting flows, we investigate the degree to which these models prove valuable in predicting significant mobility flows at different scales. The extensive analysis of flows explains some of the underlying phenomena driving commuting mobility. Motivated by features used in previous theoretical work, including the gravity model and intervening opportunities model, we consider three characteristics to address the flow prediction problem: building types, distance, and population. Specifically, we extract nine input features from open data based on these characteristics that potentially drive commuters' mobility, as follows:

- The number (count), density, and area of residential and non-residential buildings, respectively (six features),
- Region population and population density (two features), and
- Distance between census tracts (one feature)

The feature generation leverages existing work on using a machine learning approach to classify OSM building types [41] beyond the information available in OSM. Additionally, we use Open Source Routing Machine (OSRM), an OSM-based routing API [42], to generate trip duration between all pairs of regions used to represent distance. Using these features, we first provide a fair comparison of different models for predicting commuter flows. Our first case study focuses on New York City (NYC), USA, at the census tract granularity, where we compare two state-of-the-art models, including GMEL [9] and Deep Gravity [27], and eXtreme Gradient Boosting (XGBoost) and random forests (RF) as out-of-the-box models commonly used for commuting prediction [25, 26, 39]. The 2015 Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) data [43] is used to evaluate the effectiveness of our approach. We compare model performance using OSM-derived features with region-specific features unavailable outside the study area. Finally, we demonstrate the inherent flexibility of using OSM-derived features by predicting commuting flows for Fairfax County, USA. Results from both case studies validate the intuitive understanding that the destination flows, commuters going to workplaces, are concentrated in a few places.

Table 1: Notations used in the study

Notation	Meaning
$A = \{a_1, ..., a_n\}$	The study region
a_i	A subregion of the study region
n	The number of subregions
T_{ij}	The ground truth commuter flow from Region a_i to Region a_j
$\widehat{T}_{ij}$	The estimated commuter flow from Region a_i to Region a_j
d_{ij}	Spatial distance between two subregions
$O_i = \sum_j T_{ij}$	The total outflow of region a_i (to any other region)
$I_i = \sum_j T_{ji}$	The total inflow of region a_i (to any other region)
$\widehat{O}_i = \sum_j \widehat{T}_{ij}$	The estimated outflow of region a_i (to any other region)
$\widehat{I}_i = \sum_j \widehat{T}_{ji}$	The estimated inflow of region a_i (to any other region)

Results

Results show that we can get accurate flow predictions between census tracts using features derived from open data, and population, building type, and distance are the significant characteristics driving commuting mobility. The evidence from experiments at multiple scales suggests our approach produces meaningful mobility patterns while providing notable insights into the commuting flows. Before presenting our findings, we briefly define the commuting flow prediction problem.

Problem Definition

The commuting flow prediction problem can be defined as follows. Table 1 summarizes the used notations.

Definition 1 (Commuting Flow Prediction)**.** Let A denote a study region partitioned into n smaller regions $(a_1, ..., a_n)$, such as census tracts in the United States. For each region a_i, let f_i denote a corresponding set of features, and for each pair of regions a_i, a_j, let d_{ij} denote a distance measure between regions. Given these features and distance, the task is to predict the commuting flow T_{ij} for each pair of regions $a_i, a_j \in A$.

Benchmark Results

Using OSM data and the same set of derived features for New York City (NYC), Table 2 provides the commuting flow prediction accuracy for state-of-the-art models GMEL [9] and Deep Gravity [27], and out-of-the-box models XGBoost [44] and RF [45]. To evaluate model performance, we use the RMSE [46], the Coefficient of Determination R^2 [47], and the Common Part of Commuters metric [48].

The RMSE is defined as follows:

$$RMSE(A) = \sqrt{\frac{\sum_{a_{ij}}(\widehat{T}_{ij} - T_{ij})^2}{n}}$$

where A is the NYC study region, $\widehat{T}_{ij}$ is the predicted commuting flow (c.f. Definition 1), T_{ij} is the ground truth flow obtained for NYC using LODES data, and n is the number of census tracts of NYC.

RMSE values are notoriously difficult to interpret. For example, it is not clear to what degree a prediction with an RMSE of 2.279 is accurate. As such, we also provide the Coefficient of Determination R^2 and Common Part of Commuters (CPC) to provide an additional evaluation of model accuracy. Although the R^2 is well known and measures the fraction of variance explained by the model, the Common Part of Commuters (CPC) is less known. Thus, we define CPC, as follows:

$$CPC(A) = \frac{2\sum_{a_{ij}} min(\widehat{T}_{ij}, T_{ij})}{\sum_{a_{ij}} \widehat{T}_{ij} + \sum_{a_{ij}} T_{ij}}$$

CPC is 0 when predicted and ground truth flows do not overlap and 1 when both are identical [49].

Based on the results presented in Table 2, GMEL has the lowest RMSE and highest CPC and R^2 in comparison to XGBoost, Deep Gravity, and RF. Note that the two state-of-the-art models, GMEL and Deep Gravity, are originally implemented to predict commuting flow using a different set of input features, making them difficult to compare. Therefore, in order to evaluate the performance of the models independent of the data, the models are implemented using the same set of input features derived from OSM. The experiment shows that GMEL is the best-performing model compared to other models using the same features.

Table 2: Evaluation of different flow prediction models using OSM data

Model	RMSE	CPC	R^2
GMEL	2.279	0.495	0.535
XGBoost	3.125	0.261	0.111
Deep Gravity	3.144	0.325	0.078
RF	3.228	0.218	0.051

Comparative Analysis

Given our results showing that GMEL is the best-performing model, we next compare the performance of the originally proposed GMEL model, which leverages the PLUTO dataset [37] available only for New York City, with the performance of GMEL using globally available OSM data. To distinguish between the two, we call the original model GMEL-PLUTO and our approach GMEL-OSM throughout the rest of the paper. In other words, GMEL-PLUTO uses region-specific PLUTO data for flow prediction, while GMEL-OSM uses features derived from OSM data.

Table 3: Comparison of OSM and PLUTO data using GMEL model for NYC

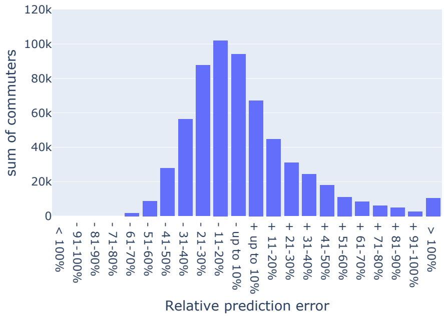
Features	RMSE	CPC	R^2
GMEL-OSM	2.279	0.495	0.535
GMEL-PLUTO	2.084	0.536	0.611

Table 3 shows that a comparable level of prediction accuracy can be achieved overall when using features derived from globally accessible and freely available OSM data. The R^2 value indicates that the three characteristics account for an 53.5% variation in commuting flows. Additionally, GMEL-OSM utilizes a smaller set of features to achieve accuracy close to GMEL-PLUTO with 65 features.

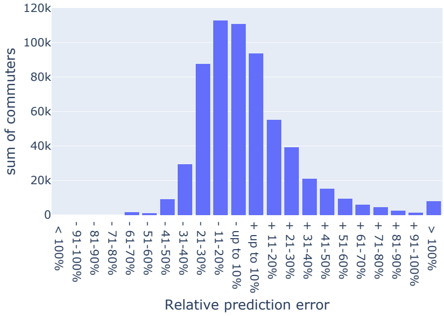
To better understand the ability of the models to capture meaningful mobility patterns beyond aggregate metrics, we also evaluate the predicted sum of outgoing commuters from an origin location a_i denoted as $\hat{O}_i = \sum_j \hat{T}_{ij}$, which we call *outflows*, and the predicted sum of incoming commuters to a destination location a_i denoted as $\hat{I}_i = \sum_j \hat{T}_{ji}$, which we call *inflows*. The $\hat{O}_i$ and $\hat{I}_i$ for each region a_i stemming from the GMEL-OSM and GMEL-PLUTO predictions are then compared to the ground truth values $O_i = \sum_j T_{ij}$ and $I_i = \sum_j T_{ji}$ derived from LODES data for NYC.

Figure 1 shows the distribution of relative prediction errors for the outflows $\frac{O_i - \hat{O}_i}{O_i}$ and the inflows $\frac{I_i - \hat{I}_i}{I_i}$ for GMEL-OSM (Figures 1a and 1c) and for GMEL-PLUTO (Figure 1b and 1d). We observe that GMEL-OSM is comparable with GMEL-PLUTO to predict outflows, but performs somewhat weaker for inflows. It is likely due to the nature of commuting flows, with inflows being limited to a small group of destination census tracts (cf. discussion in the Data Section). Even so, the results show the practicality of predicted flows compared to ground truth data. Out of those census tracts where flow is over-predicted by more than 100%, many have a commuting flow count of 10 individuals or fewer. It indicates that our approach is capable of predicting real-world commuting mobility at the tract level, where the flow count is generally more than 10.

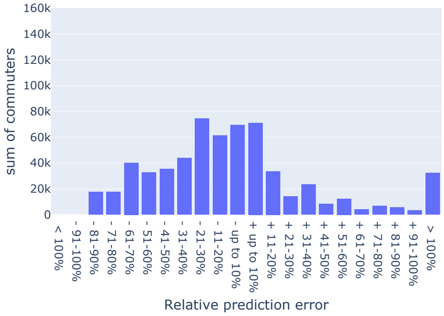
To assess the accuracy of the predicted inflows and outflows for census tracts, Figure 2 shows scatter plots comparing the ground truth flows against the predicted flows using GMEL-OSM (Figures 2a and 2c) and GMEL-PLUTO (Figures 2b and 2d). Both models tend to overestimate inflows that are smaller in the real world and underestimate large inflows, as indicated by the points that fall above and below the identity line. Likewise, both models also tend to overestimate smaller outflows. Again, while both models produce similar



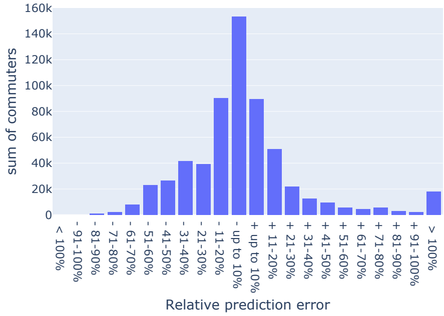
(a) Percentage of under or overestimation of NYC commuters' outflows using GMEL-OSM.



(b) Percentage of under or overestimation of NYC commuters' outflows using GMEL-PLUTO.



(c) Percentage of under or overestimation of NYC commuters' inflows using GMEL-OSM.



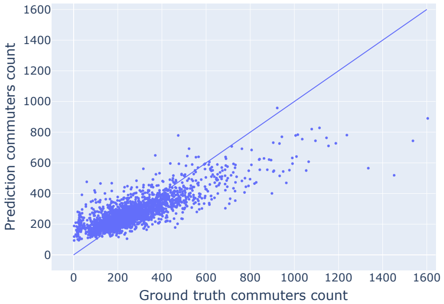
(d) Percentage of under or overestimation of NYC commuters' inflows using GMEL-PLUTO.

Fig. 1: Comparison of GMEL-OSM and GMEL-PLUTO commuters under or overestimation in NYC flows.

results for outflows, GMEL-PLUTO (65 custom feature model) seems to perform better when predicting the inflows, essentially confirming the results of Figure 1 at a more granular level.

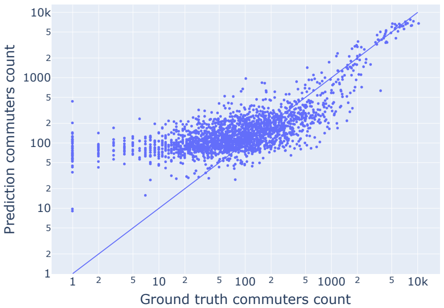
We note that the maximum number of commuters going to a census tract is much higher than coming from a home location, which is consistent in both prediction models and the ground truth. It indicates that the inflows are much denser to specific census tracts or workplaces. We investigate and explain this phenomenon in our Data Section.

We can also map the differences between predicted and ground truth outflows as presented in Figure 3 and inflows presented in Figure 4. Positive relative prediction errors indicate over-prediction and are depicted in shades of blue colors. In contrast, negative percentages indicate under-prediction and are shown in shades of red. Green shows a prediction largely matching the ground truth flows. Note that the large tracts in the south of the study area are mostly

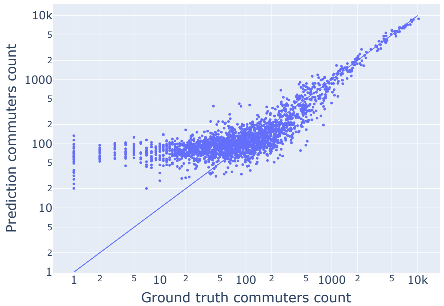


(a) Comparison of NYC commuters' outflows using GMEL-OSM with ground truth.

(b) Comparison of NYC commuters' outflows using GMEL-PLUTO with ground truth.



(c) Comparison of NYC commuters' inflows using GMEL-OSM with ground truth (log-log scale).



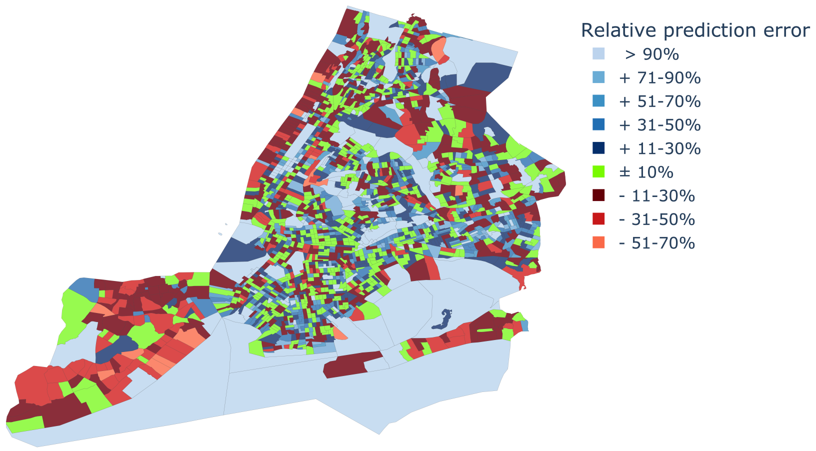
(d) Comparison of NYC commuters' inflows using GMEL-PLUTO with ground truth (log-log scale).

Fig. 2: Comparison of GMEL-OSM and GMEL-PLUTO commuters with ground truth in NYC flows.

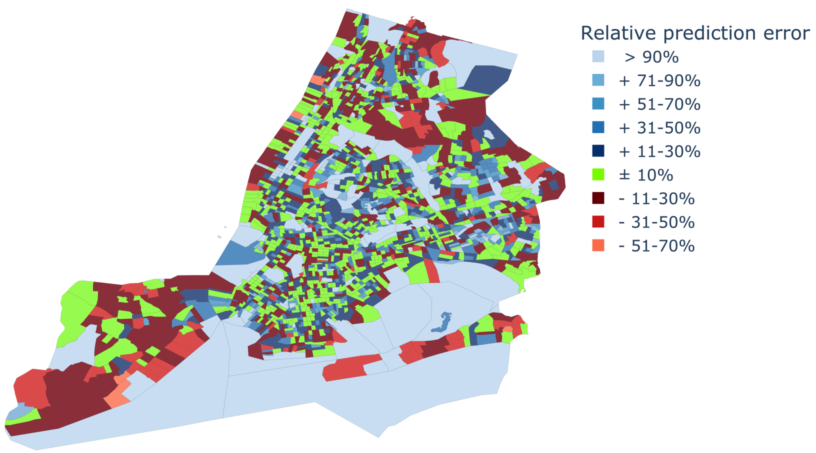
comprised of water, thus having small in and outflows. As a result, minor flow prediction errors for these census tracts provide high relative percentage errors and as such are shown as large light blue areas.

Upon comparing Figures 3 and 4, we can see that GMEL-OSM and GMEL-PLUTO flow predictions are very similar in terms of the relative prediction error. Both approaches have less success in predicting destination flows. It is once again likely due to the large number of features used in GMEL-PLUTO that are likely better at capturing the inflows to destination census tracts. We discuss steps that we may take to address this in future work in the Discussion Section.

To better understand the utility of predicted commuter flows, we also performed experiments focusing on a single origin (destination) tract to understand how well models can capture the distribution of destination (origin) tracts to (from) this tract. For this purpose, we select the census tract having the median outflow (GeoID: 36047037300, denoted as the *Origin Median*)



(a) Relative errors of NYC outflows using GMEL-OSM.

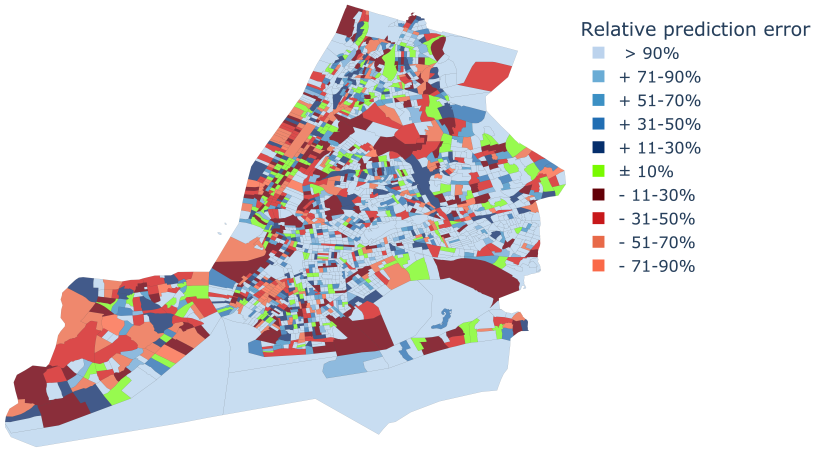


(b) Relative errors of NYC outflows using GMEL-PLUTO.

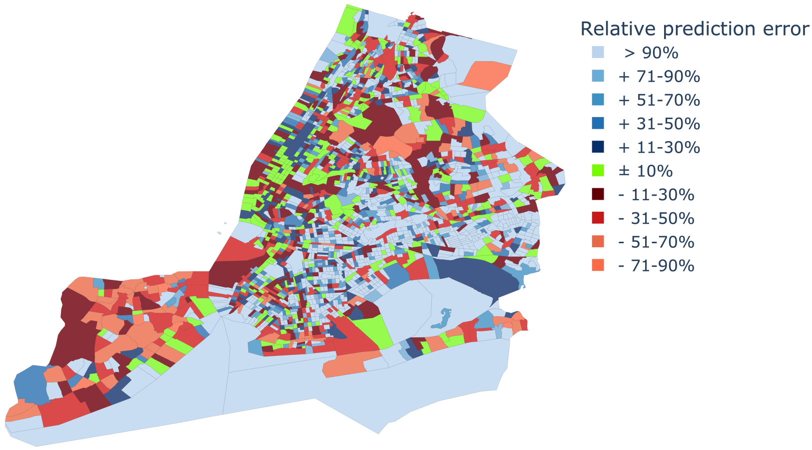
Fig. 3: Comparison of GMEL-OSM and GMEL-PLUTO in NYC outflows. Plotly version 5.13.0 was used to generate the maps.

and the census tract having the median inflow (GeoID 36005024800, denoted as the *Destination Median*). We use these two census tracts to evaluate (i) the distribution of outflows from the Origin Median to understand how well the models can understand where people commute to (from one specific census tract) and (ii) the distribution of inflows from the Destination Median to understand how well our models can capture the distribution of where people commute from (to one specific census tract).

Table 4 shows the results of these experiments. Out of all 448 census tracts in the NYC study region included in the test set, 354 tracts have a zero commuting flow from the Origin Median. The remaining 94 census tracts having



(a) Relative errors of NYC inflows using GMEL-OSM.



(b) Relative errors of NYC inflows using GMEL-PLUTO.

Fig. 4: Comparison of GMEL-OSM and GMEL-PLUTO in NYC inflows. Plotly version 5.13.0 was used to generate the maps.

non-zero commuting flows capture a total of 244 commuters. Using GMEL-OSM, we have 332 predicted zero commuting flows and 116 predicted non-zero commuting flow. Out of the predicted 116 predicted non-zero flows, 48 match with the 94 ground truth non-zero flows. Out of the 332 predicted zero flows, 286 match with the 354 ground truth flows. It yields an overall 74.5% accuracy in predicting whether any census tract has a non-zero flow from the Origin Median. Note that we round predictions to the nearest integer for this experiment, such as that a predicted zero flow is equivalent to a predicted flow of less than 0.5 individuals. We observe that for GMEL-PLUTO, the accuracy

Table 4: Single origin and destination census tract predictions

Census Tract	Approach	Zero Flows Count (Matching)	Non-Nero Flows Count (Matching)	Sum of Commuters
Origin Median	Ground Truth	354 (354)	94 (94)	244
	GMEL-OSM	332 (286)	116 (48)	212
	GMEL-PLUTO	345 (304)	103 (53)	201
Destination Median	Ground Truth	411 (411)	46 (46)	81
	GMEL-OSM	418 (393)	39 (21)	43
	GMEL-PLUTO	427 (398)	30 (17)	32

is higher at 79.6%, indicating that the model can better predict destination flows by leveraging PLUTO data.

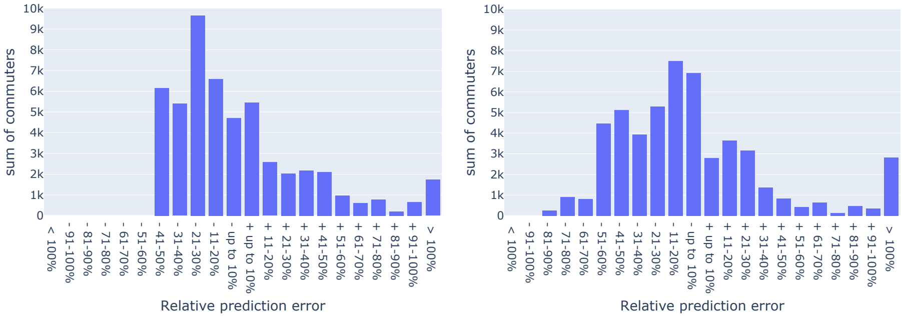
Similarly, by considering only the Destination Median as a single destination, GMEL-OSM and GMEL-PLUTO matched 90.5% and 90.8%, respectively, out of 457 origin tracts in the test set. We observe that the destination median has a relatively small number of only 81 incoming commuters in the ground truth. It is explained by the long-tail distribution of inflows, which we further investigate and explain in the Data Section.

Overall, we observe that while GMEL-OSM and GMEL-PLUTO provide very accurate flow predictions when aggregated to census tracts, the prediction of individual origin-destination flows remains challenging. The reason is that the vast majority of origin-destination flows are zero and among the non-zero flows, most flows are less than five individuals. Despite these small numbers, which correspond to rare events of individual origin-destination commutes, both GMEL-OSM and GMEL-PLUTO give good results.

Based on the results presented so far, we can conclude that there are marginal gains in performance by using a large number of region-specific features using GMEL-PLUTO, and we can achieve similar results with a small set of features derived from open data that is globally available. To examine whether GMEL-OSM is usable in other regions, we trained and tested the model for Fairfax County in Virginia and compared the predicted flows with the LODES data as ground truth. Note that we cannot compare GMEL-OSM with GMEL-PLUTO because the latter approach uses NYC-specific data, which is publicly unavailable for Fairfax.

Histograms in Figure 5 show the relative percentage errors of outflows and inflows at the tract level compared to the ground truth. Figure 6 demonstrates the trend of flow prediction for outflows and inflows, respectively. We observe that the model performance in Fairfax, VA is comparable, if not better than the NYC case study using GMEL-PLUTO. Based on the histograms, it appears that the commuting inflows for Fairfax are easier to predict and less extreme than in NYC.

Additionally, we trained GMEL-OSM using NYC data and tested the pre-trained model to predict the commuting flows for Fairfax to determine whether the model is useful in locations where training commuting flow data (obtained for the U.S. from LODES data) is not available. Table 5 shows that the model



(a) Percentage of under or overestimation of Fairfax commuters' outflows using GMEL-OSM.

(b) Percentage of under or overestimation of Fairfax commuters' inflows using GMEL-OSM.

Fig. 5: Commuters under or overestimation using GMEL-OSM for Fairfax.

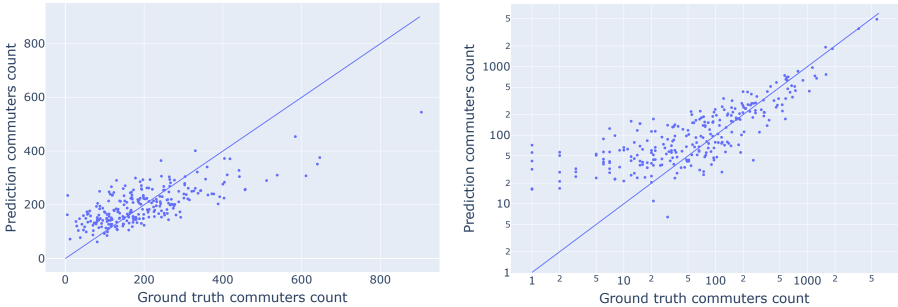
trained in NYC and transferred to Fairfax provides acceptable results by explaining 62.1% of the variation in the commuting flows of Fairfax, compared to 70.2% using the model that was trained using Fairfax LODES data.

Table 5: Comparison of GMEL-OSM in Fairfax using transfer learning

Training data	RMSE	CPC	R^2
Fairfax	6.476	0.643	0.702
NYC	7.427	0.572	0.621

Discussion

Results for the two study areas show that commuting flows can be accurately predicted using features derived from OSM data, which is globally available and freely accessible. Comparative results reveal that GMEL-OSM achieves accuracy close to region-specific GMEL-PLUTO, which outperforms other state-of-the-art models but cannot be used outside NYC due to a lack of input data for other regions. The learning framework of GMEL-OSM relies on geographic contextual information [50] for predicting commuting flows between origin-destination pairs of subregions. Our findings suggest that the OSM data captures the contextual information very well for the origin and destination locations, providing a rich and effective source of input features for GMEL-OSM. Besides aggregated results, the in-depth analysis demonstrates the usefulness of the predicted flows for urban planning [51], disease transmission [52, 53], and other applications [54, 55]. We find that inflows are concentrated in a few destinations while outflows are more evenly distributed, validating the intuition that people commute to a few workplaces and reside in dispersed locations. Our analysis shows that GMEL-OSM effectively captures



(a) Comparison of Fairfax commuters' outflows using GMEL-OSM with ground truth.

(b) Comparison of Fairfax commuters' inflows using GMEL-OSM with ground truth (log-log scale).

Fig. 6: Comparison of GMEL-OSM commuters prediction with ground truth for Fairfax flows.

this divergent phenomenon, matching the trend of outflows and inflows in the ground truth. Additionally, we also illustrate that the number of residential and non-residential buildings in census tracts plays a crucial role in predicting commuters' mobility. Our results indicate that building types, distance, and population are the essential characteristics driving commuting mobility.

While the population can be estimated at a fine-grained scale using OSM data [56, 57], for simplicity, we utilized the U.S. Census data as a proxy for this. In future work, we plan to extend our proposed approach for generating population features, alleviating the need for census data. To investigate the explainability of the input features, we might explore a unified mechanism for interpreting predictions such as SHapley Additive exPlanations (SHAP) [58]. It would help us understand which features are useful for better commuting flow predictions, potentially leading to more suitable feature selection for improving the performance of our approach. Where we found relatively weaker prediction accuracy for the destination flows, there is an opportunity to examine what features might improve this aspect of the predictions. Prior work shows the effectiveness of points of interest (PoIs) [59] and land use [60, 61] for predicting flows. Therefore, we would explore types of PoIs and land use as other characteristics driving mobility. Finally, our transfer learning results for Fairfax County show promise for future work in which we would plan to apply our approach to regions where LODES or equivalent commuting data is not publicly unavailable, potentially outside the U.S.

Methods

Models

We aim to predict commuting flows from three characteristics operationalized using globally available and openly accessible data. Therefore, we examine four models including GMEL, Deep Gravity, XGBoost, and random forest (RF), comparing their performance using the same set of features derived from OSM. GMEL employs graph representation learning by using the graph attention network (GAT) framework for capturing the geographic contextual information from the nearby regions for commuting flow predictions. Given the potentially unique characteristics of the regions, it uses two GATs separately for origin and destination locations. As described in the proposed model [9], we used one hidden layer and an embedding size of 128 as hyperparameters for GMEL-OSM. Deep Gravity utilizes deep neural networks to generate mobility flows using features retrieved from OSM and census data [27]. The main features include land use, points of interest, road networks, and the population of the study region. XGBoost is a regression tree gradient boosting model, a highly scalable learning system capable of efficiently handling sparse data and supporting multicore parallel computing for quick model exploration [44]. XGBoost has been shown to outperform traditional mathematical gravity and radiation models for commuting flow prediction using U.S. Census data [25]. Random forests are the ensemble of individual tree predictions averaged for regression problems and the prediction with maximum votes selected for classification problems [45]. Compared to the gravity model and artificial neural networks, the accuracy for the random forest is higher for predicting commuting flows in NYC in previous work [26]. As described in Results Section, we evaluate the comparative performance of these models for our approach using the parameters and configurations prescribed in the proposed studies.

Data

We use real-world commuting flows obtained from the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) 2015 dataset [43, 62] as ground truth for training and testing the models. LODES data captures the raw number of commuters between two regions at the census block level, and we aggregated it at the census tract level.

Across the 2,168 NYC census tracts, there are $2168^2 = 4,700,224$ pairwise flows, of which 905,837 are non-zero with a total of 3,031,641 commuters. Similarly, across the 263 Fairfax County census tracts, there are a possible 69,169 flows out of which 34,366 are non-zero flows, capturing 259,792 commuters. Unlike prior work [9, 12, 26], we include flows that are zero in the ground truth LODES data. While LODES data does not explicitly include zero flows in their data, the omitted flows between a pair of census tracts are implicitly assumed to be zero values, which are missing from the evaluation of prior work [9, 12, 26]. However, omitting such flows creates biased models that learn that any pair of origin-destination census tracts must always have at least a flow count of one commuter. Our experiments include all pairs of census tracts, including zero flows, eliminating the bias. In other words, we add zero flows to training and test sets of all evaluated models to allow a fair evaluation. We note that due to this difference, the quantitative results we report in the aggregated metrics in the Results Section (such as Table 2) are generally lower than reported in prior work, as our results include cases of flows where models predict a non-zero flow instead of a zero flow count in the ground truth. For training and testing, we split the flows into a 60% training set, a 20% validation set, and a 20% test set.

Table 6 presents the descriptive statistics for the NYC and Fairfax County LODES outflows O_i and inflows I_i aggregated at the tract level. We notice a much higher standard deviation of the inflow of commuters in both study regions. The maximum count of commuters for the inflows also highlights the significant difference in variance. Furthermore, the 3rd quantile values in both cases show the skewness in the distribution of commuters. These results demonstrate the concentrated nature of inflows in comparison to outflows, where the majority of commuters move to a small set of destination census tracts. Therefore, as our results suggest, it is much harder to predict the commuters' count for inflows.

Table 6: Descriptive statistics of ground truth data Data

Study Area	Flow Type	Mean	Standard Deviation	Min	25%	Median	75%	Max
NYC	Outflows	280	176	4	168	244	350	1604
	Inflows	280	817	1	34	81	190	10243
Fairfax	Outflows	197	120	5	111	173	255	904
	Inflows	197	482	1	21	67	180	5702

OSM is an open-source collaborative project that provides free access to geographic data collected by volunteers at the global level [40]. The OSM data is structured as a set of elements such as nodes, ways, and relations that represent points of interest, polylines or polygons, and more complex shapes consisting of relationships between simple elements. Tags of key and value pairs can describe all the elements. For instance, a polygon can be tagged with the key as building and value as a residential, describing a residential building. This way, OSM data provides extensive coverage of points, buildings, roads, parking lots, and many other types of geographic information via editable maps. The OSM data we used for this work consists of 1,090,752 NYC and 204,671 Fairfax building footprints.

Features

The features used in the models for predicting the flows are derived from OSM and the 2010 U.S. Census data [63]. Previous work shows that building types are missing from a vast majority of OSM data, and the spatial and non-spatial features of the data can be used to categorize buildings into residential or non-residential types [41]. We use this classification method to label the OSM buildings data and derive six input features for our study. In the first step of data preparation, we classify buildings for NYC and Fairfax. And in the second step, we calculate the count, area, and density of two building types for each census tract, resulting in six features.

We use population and the population density for each tract as two more input features. Although population estimates can be derived from OSM features in the same way [56, 57], we use census data as a proxy for this approach. Finally, we obtain the trip duration between the centroids of census tracts using Open Source Routing Machine (OSRM) [42] and use it as the edge feature for the geo-adjacency network of GMEL-OSM. OSRM also relies on the maps from the OSM road network for calculating the shortest paths between O-D pairs.

Data availability

Data are available from OSF at https://osf.io/sxzar/

Code availability

The code is available in a GitHub repository at https://github.com/heykuldip/commuting_flows_prediction

References

[1] Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. nature **453**(7196), 779–782 (2008)

[2] Alessandretti, L., Aslak, U., Lehmann, S.: The scales of human mobility. Nature **587**(7834), 402–407 (2020)

[3] Rong, C., Li, T., Feng, J., Li, Y.: Inferring origin-destination flows from population distribution. IEEE Transactions on Knowledge and Data Engineering **35**(1), 603–613 (2021)

[4] Levinson, D.M.: Accessibility and the journey to work. Journal of transport geography **6**(1), 11–21 (1998)

[5] Layman, C.C., Horner, M.W.: Comparing methods for measuring excess commuting and jobs-housing balance: empirical analysis of land use changes. Transportation Research Record **2174**(1), 110–117 (2010)

[6] Luca, M., Barlacchi, G., Lepri, B., Pappalardo, L.: A survey on deep learning for human mobility. ACM Computing Surveys (CSUR) **55**(1), 1–44 (2021)

[7] Jiang, R., Cai, Z., Wang, Z., Yang, C., Fan, Z., Chen, Q., Tsubouchi, K., Song, X., Shibasaki, R.: Deepcrowd: A deep model for large-scale citywide crowd density and flow prediction. IEEE Transactions on Knowledge and Data Engineering **35**(1), 276–290 (2021)

[8] Rodrigue, J.-P.: The Geography of Transport Systems. Routledge, UK (2020)

[9] Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., Silva, C.: Learning geo-contextual embeddings for commuting flow prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 808–816 (2020)

[10] Eubank, S., Guclu, H., Anil Kumar, V., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature **429**(6988), 180–184 (2004)

[11] Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the national academy of sciences **106**(51), 21484–21489 (2009)

[12] Yang, Y., Herrera, C., Eagle, N., González, M.C.: Limits of predictability in commuting flows in the absence of data for calibration. Scientific reports **4**(1), 1–9 (2014)

[13] Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. Journal of The Royal Society Interface **10**(84), 20130246 (2013)

[14] Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C.: Exploring universal patterns in human home-work commuting from mobile phone data. PloS one **9**(6), 96180 (2014)

[15] Masucci, A.P., Serras, J., Johansson, A., Batty, M.: Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. Physical Review E **88**(2), 022812 (2013)

[16] Yin, G., Huang, Z., Bao, Y., Wang, H., Li, L., Ma, X., Zhang, Y.: Convgcn-rf: A hybrid learning model for commuting flow prediction considering geographical semantics and neighborhood effects. GeoInformatica, 1–21 (2022)

[17] Zipf, G.K.: The p 1 p 2/d hypothesis: on the intercity movement of persons. American sociological review **11**(6), 677–686 (1946)

[18] Alonso, W.: The system of intermetropolitan population flows. Institute of Urban and Regional Development, University of California (1971)

[19] Simini, F., González, M.C., Maritan, A., Barabási, A.-L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012)

[20] Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M.C., Toroczkai, Z.: Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nature communications **5**(1), 5347 (2014)

[21] Stouffer, S.A.: Intervening opportunities: a theory relating mobility and distance. American sociological review **5**(6), 845–867 (1940)

[22] Kotsubo, M., Nakaya, T.: Kernel-based formulation of intervening opportunities for spatial interaction modelling. Scientific reports **11**(1), 950 (2021)

[23] Stouffer, S.A.: Intervening opportunities and competing migrants. Journal of regional science **2**(1), 1–26 (1960)

[24] Schneider, M.: Gravity models and trip distribution theory. Papers in Regional Science **5**(1), 51–56 (1959)

[25] Morton, A., Piburn, J., Nagle, N.: Need a boost? a comparison of traditional commuting models with the xgboost model for predicting commuting flows (short paper). In: 10th International Conference on Geographic Information Science (GIScience 2018) (2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik

[26] Pourebrahim, N., Sultana, S., Niakanlahiji, A., Thill, J.-C.: Trip distribution modeling with twitter data. Computers, Environment and Urban Systems **77**, 101354 (2019)

[27] Simini, F., Barlacchi, G., Luca, M., Pappalardo, L.: A deep gravity model for mobility flows generation. Nature communications **12**(1), 6576 (2021)

[28] Yao, X., Gao, Y., Zhu, D., Manley, E., Wang, J., Liu, Y.: Spatial origin-destination flow imputation using graph convolutional networks. IEEE Transactions on Intelligent Transportation Systems **22**(12), 7474–7484 (2020)

[29] Wang, Y., Yin, H., Chen, H., Wo, T., Xu, J., Zheng, K.: Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1227–1235 (2019)

[30] Koca, D., Schmöcker, J.D., Fukuda, K.: Origin-destination matrix estimation by deep learning using maps with new york case study. In: 2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 1–6 (2021). IEEE

[31] Rong, C., Feng, J., Ding, J.: Goddag: Generating origin-destination flow

for new cities via domain adversarial training. IEEE Transactions on Knowledge and Data Engineering (2023)

[32] Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)

[33] Liang, Y., Ouyang, K., Sun, J., Wang, Y., Zhang, J., Zheng, Y., Rosenblum, D., Zimmermann, R.: Fine-grained urban flow prediction. In: Proceedings of the Web Conference 2021, pp. 1833–1845 (2021)

[34] Zeng, H., Peng, Z., Huang, X., Yang, Y., Hu, R.: Deep spatio-temporal neural network based on interactive attention for traffic flow prediction. Applied Intelligence, 1–12 (2022)

[35] Robinson, C., Dilkina, B.: A machine learning approach to modeling human migration. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, pp. 1–8 (2018)

[36] Zhou, F., Li, L., Zhang, K., Trajcevski, G.: Urban flow prediction with spatial–temporal neural odes. Transportation Research Part C: Emerging Technologies **124**, 102912 (2021)

[37] NYC Department of City Planning: PLUTO and MapPLUTO. Accessed: 4/6/2023. https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[38] Yin, G., Huang, Z., Bao, Y., Wang, H., Li, L., Ma, X., Zhang, Y.: Convgcn-rf: A hybrid learning model for commuting flow prediction considering geographical semantics and neighborhood effects. GeoInformatica **27**(2), 137–157 (2023)

[39] Spadon, G., Carvalho, A.C.d., Rodrigues-Jr, J.F., Alves, L.G.: Reconstructing commuters network using machine learning and urban indicators. Scientific reports **9**(1), 11801 (2019)

[40] OpenStreetMap contributors: OpenStreetMap. Accessed: 4/6/2023 (2023). https://www.openstreetmap.org

[41] Atwal, K.S., Anderson, T., Pfoser, D., Züfle, A.: Predicting building types using openstreetmap. Scientific Reports **12**(1), 19976 (2022)

[42] Luxen, D., Vetter, C.: Real-time routing with openstreetmap data. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 513–516 (2011)

[43] U.S. Census Bureau: LODES data 2015 (https://lehd.ces.census.gov/

data/). Accessed: 4/6/2023

[44] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[45] Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

[46] Hancock, G.R., Freeman, M.J.: Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. Educational and Psychological Measurement **61**(5), 741–758 (2001)

[47] Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. PeerJ Computer Science **7**, 623 (2021)

[48] Lenormand, M., Huet, S., Gargiulo, F., Deffuant, G.: A universal model of commuting networks (2012)

[49] Lenormand, M., Bassolas, A., Ramasco, J.J.: Systematic comparison of trip distribution laws and models. Journal of Transport Geography **51**, 158–169 (2016)

[50] Feng, J., Li, Y., Lin, Z., Rong, C., Sun, F., Guo, D., Jin, D.: Context-aware spatial-temporal neural network for citywide crowd flow prediction via modeling long-range spatial dependency. ACM Transactions on Knowledge Discovery from Data (TKDD) **16**(3), 1–21 (2021)

[51] Zeng, J., Zhang, G., Rong, C., Ding, J., Yuan, J., Li, Y.: Causal learning empowered od prediction for urban planning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 2455–2464 (2022)

[52] Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A., Grenfell, B.T.: Synchrony, waves, and spatial hierarchies in the spread of influenza. science **312**(5772), 447–451 (2006)

[53] Ferguson, N.M., Cummings, D.A., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S.: Strategies for mitigating an influenza pandemic. Nature **442**(7101), 448–452 (2006)

[54] Li, M.-H., Chen, B.-Y., Li, C.-T.: A hybird method with gravity model and nearest-neighbor search for trip destination prediction in new metropolitan areas. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 6553–6560 (2022). IEEE

[55] Delventhal, M.J., Kwon, E., Parkhomenko, A.: Jue insight: How do cities change when we work from home? Journal of Urban Economics **127**, 103331 (2022)

[56] Bast, H., Storandt, S., Weidner, S.: Fine-grained population estimation. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1–10 (2015)

[57] Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., Zipf, A.: Fine-resolution population mapping using openstreetmap points-of-interest. International Journal of Geographical Information Science **28**(9), 1940–1963 (2014)

[58] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[59] Cai, M., Pang, Y., Sekimoto, Y.: Spatial attention based grid representation learning for predicting origin–destination flow. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 485–494 (2022). IEEE

[60] Lee, M., Holme, P.: Relating land use and human intra-city mobility. PloS one **10**(10), 0140152 (2015)

[61] Horner, M.W.: Spatial dimensions of urban commuting: a review of major issues and their implications for future geographic research. The Professional Geographer **56**(2), 160–173 (2004)

[62] Credit, K., Arnao, Z.: A method to derive small area estimates of linked commuting trips by mode from open source lodes and acs data. Environment and Planning B: Urban Analytics and City Science **50**(3), 709–722 (2023)

[63] U.S. Census Bureau: Population data 2010 (https://data.census.gov/table). Accessed: 4/6/2023

Acknowledgments

This work is supported by National Science Foundation Grant No. 2109647 titled "Data-Driven Modeling to Improve Understanding of Human Behavior, Mobility, and Disease Spread".

This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

Author contributions statement

K.S.A, T.A, A.Z, and D.P. designed the study. K.S.A, T.A, A.Z, and D.P. performed the analyses. K.S.A, T.A, A.Z, and D.P. conceived the experiments, K.S.A conducted the experiments. K.S.A, T.A, A.Z, and D.P. wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.S.A.