Computational Urban Science

ORIGINAL PAPER

Open Access

Commuting flow prediction using OpenStreetMap data

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

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 utility of state-of-the-art commuting flow prediction models with globally available OpenStreetMap data while achieving prediction accuracy comparable to location-specific and proprietary data. We show that the types of buildings, residential and non-residential, are a strong indicator for predicting commuting flows. Consistent with theoretical and analytical models, our experiments indicate that building types, distance, and population are the determining characteristics for mobility related to commuting. Our experiments show that predicted flows closely match ground truth flows. Our work enables accurate flow prediction using building types to support applications such as urban planning and epidemiology.

Keywords Commuting flows, OSM, Flow prediction, Graph attention networks

1 Introduction

Understanding how individuals commute between places is as challenging as it is significant (Gonzalez et al. 2008; Alessandretti et al. 2020). Commuting flow prediction estimates the number of people moving between regions in a geographic area based on descriptive features, such as population (Rong et al. 2021), distance to other locations (Levinson 1998), and land use type (Layman and Horner 2010). Commuting flow prediction is helpful in many applications, such as understanding migration patterns (Jiang et al. 2021), urban planning (Rodrigue 2020), and epidemiology (Balcan et al. 2009). Considering that commuting flows vary little from workday to workday (Yang et al. 2014), 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 (Masucci et al. 2013). Therefore, similar to other approaches (Liu et al. 2020; Yin et al. 2023), 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 (Zipf 1946) and its extensions, including the radiation model (Alonso 1971; Ren et al. 2014), the intervening opportunities model (Stouffer 1940; Kotsubo and Nakaya 2021), and the competing migrants model (Stouffer 1960). 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 (Schneider 1959). Thus,

*Correspondence:
Kuldip Singh Atwal
katwal@gmu.edu
[1] Geography and Geoinformation Science, George Mason University, 4400 University Drive, Fairfax, VA 22030, US
[2] Department of Computer Science, Emory University, 201 Dowman Drive, Atlanta, GA 30322, US

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 (Morton et al. 2018; Yao et al. 2020; Yin et al. 2023). 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 (Koca et al. 2021; Rong et al. 2023). Many studies use spatiotemporal characteristics to address the flow prediction problem using neural networks (Zhang et al. 2017; Liang et al. 2021; Robinson and Dilkina 2018), which can also be combined with ordinary differential equations (Zhou et al. 2021). A state-of-the-art model, the Geo-contextual Multitask Embedding Learner (GMEL) (Liu et al. 2020) predicts 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) (NYC 2015) dataset, which is only available for NYC. In another example, the ConvGCN-RF model (Yin et al. 2023) uses a 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. (2019) 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. This makes the use of the models in other data-poor study regions challenging. Another model, Deep Gravity (Simini et al. 2021) overcomes this limitation by leveraging features obtained from a globally crowdsourced and available dataset called OpenStreetMap (OSM) (2024). As described by the model's name, the selected features are inspired by the classic gravity model, considering population, distance, and other OSM features to predict flows.

Given the variety of input features used across the various models, it is difficult to compare and select the best-performing model for flow prediction. Therefore, in this study, we start with a benchmark of the two state-of-the-art models GMEL (Liu et al. 2020) and Deep Gravity (Simini et al. 2021), and two out-of-the-box models including eXtreme Gradient Boosting (XGBoost) and random forests (RF) (Morton et al. 2018; Spadon et al. 2019) against the same set of features derived from OSM. Our case study focuses on New York City (NYC), USA, at the census tract granularity. We first evaluate the flow prediction models against the 39 OSM features proposed by Deep Gravity, comparing the predictions against the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) data (Census Bureau 2015) as a ground truth. Moreover, since models are typically 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 lack of semantic information in OSM data limits the features that can be used for prediction (Vargas-Munoz et al. 2020; Liu and Long 2016). While Deep Gravity improves transferability to other study areas by using features that are directly available from OSM, there are opportunities to use feature engineering to improve the accuracy and applicability of flow prediction models. As such, we propose to improve flow predictions by additionally incorporating building type as an indicator, a key factor identified in the intervening opportunities model. Specifically, we extract nine input features from open data, 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)

It is hard to obtain building types from authoritative or open data sources, including OSM (Fonte et al. 2018). Therefore, our feature generation approach leverages existing work using a machine learning approach to classify building footprints by building type (residential/non-residential) using only OSM data (Atwal et al. 2022). Additionally, as used in GMEL, we employ Open Source Routing Machine (OSRM), an OSM-based routing API (Luxen and Vetter 2011), to generate trip duration between all pairs of regions representing distance.

Based on the benchmark results, we find that GMEL is consistently the best model for predicting commuting flows. Additionally, we show that the majority of models improve flow prediction accuracy when using building-type features derived from OSM versus the OSM features leveraged by Deep Gravity. Finally, we show that GMEL coupled with OSM building-type features produces comparable prediction accuracy to the original GMEL model trained on location-specific data.

Therefore, we couple the GMEL model with building-type features derived from OSM to predict commuting flows for Fairfax County, USA, for which the original features leveraged by GMEL are not available. Results from both case studies show that we can get accurate flow predictions between census tracts using features derived from open data without relying on location-specific features. Figure 1 shows the map of the study areas.

We note that OSM lacks complete building footprint data (Zhou et al. 2022), and coverage is uneven in different regions globally (Herfort et al. 2023). Therefore, our approach is applicable in urbanized areas where building footprint data can be obtained. As explained in Sect. 2.3 on the data, both case studies have more than 75% building coverage compared to authoritative sources. Additionally, the footprint data with building types is publicly available in the United States (F. de Arruda et al. 2024).

2 Methods

Before presenting our findings, we briefly define the commuting flow prediction problem.

2.1 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 subregions $(a_1, ..., a_n)$, such as census tracts in the United States. For each subregion a_i, let f_i denote a corresponding set of features, and for each pair of subregions 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 subregions $a_i, a_j \in A$.

2.2 Models

We aim to predict commuting flows from three characteristics operationalized using publicly available data such as OSM. 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

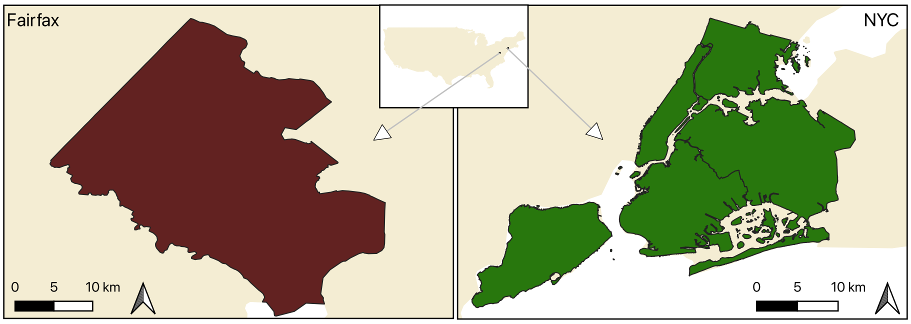


Fig. 1 The map of NYC and Fairfax study areas

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 subregion a_i to subregion a_j
$\widehat{T}_{ij}$	The estimated commuter flow from subregion a_i to subregion a_j
d_{ij}	Spatial distance between two subregions
$O_i = \sum_j T_{ij}$	The total outflow of subregion a_i (to any other subregion)
$I_i = \sum_j T_{ji}$	The total inflow of subregion a_i (from any other subregion)
$\widehat{O}_i = \sum_j \widehat{T}_{ij}$	The estimated outflow of subregion a_i (to any other subregion)
$\widehat{I}_i = \sum_j \widehat{T}_{ji}$	The estimated inflow of subregion a_i (from any other subregion)

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 (Liu et al. 2020), we used one hidden layer and an embedding size of 128 as hyperparameters for GMEL. Deep Gravity utilizes deep neural networks to generate mobility flows using features retrieved from OSM and census data (Simini et al. 2021). The main features include road network, points of interest, land use, 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 (Chen and Guestrin 2016). XGBoost has been shown to outperform traditional mathematical gravity and radiation models for commuting flow prediction using U.S. Census data (Morton et al. 2018). Random forests are the ensemble of individual tree predictions averaged for regression problems and the prediction with maximum votes selected for classification problems (Breiman 2001). 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 (Pourebrahim et al. 2019). As described in Sect. 3 on results, we evaluate the comparative performance of these models for our approach using the parameters and configurations prescribed in the proposed studies.

GMEL uses 65 features as urban indicators of NYC to predict commuting flows. The features, such as the number of buildings in each built year interval, the number of tax lots, and floor area ratio statistics, are region-specific indicators available only for NYC. Similarly, Deep Gravity employs 39 features obtained from OSM, such as road network, count of education and food points of interests, etc. As explained in Sect. 2.4 on features, our approach utilizes building types as the main indicator to derive nine features for predicting commuting flows.

To evaluate model performance, we use the root mean square error (RMSE) (Hancock and Freeman 2001), the Coefficient of Determination (Chicco et al. 2021) R^2, and the Common Part of Commuters (CPC) metric (Lenormand et al. 2012).

The RMSE is defined as follows:

$$RMSE(A) = \sqrt{\frac{\sum_{a_i,a_j}(\widehat{T}_{ij} - T_{ij})^2}{n^2}} \tag{1}$$

where $A = \{a_1, ..., a_n\}$ is the study region with subregions, T_{ij} is the ground truth flow between subregions a_i and a_j (obtained for NYC using LODES data) as defined in Definition 1, $\widehat{T_{ij}}$ is the predicted commuting flow, and n is the number of subregions (census tracts for NYC).

RMSE values are notoriously difficult to interpret. For example, it is not clear to what degree a prediction with an RMSE of 2.393 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:

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

The CPC ranges from 0 (no overlap between prediction and ground truth) to 1 (identical prediction and ground truth).

2.3 Data

We use real-world commuting flows obtained from the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) 2015 dataset (Census Bureau 2015; Credit and Arnao 2023) 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 $2168^2 = 4,700,224$ pair-wise 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 (Liu et al. 2020; Yang et al. 2014; Pourebrahim et al. 2019), 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 (Liu et al. 2020; Yang et al. 2014; Pourebrahim et al. 2019). 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 5) 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 2 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.

OSM is an open-source collaborative project that provides free access to geographic data collected by volunteers at the global level (OSM 2024). 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. However, the lack of semantic information is a challenge (Liu and Long 2016). The OSM data we used for this work consists of 1,090,752 NYC and 204,671 Fairfax building footprints, which cover 99% and 76% of buildings respectively for both study areas compared to authoritative sources (N.Y.C 2024; Fairfax County 2024). Figure 2 shows the workflow of our approach.

2.4 Features

The features used in the models for predicting the flows are derived from OSM and the 2010 U.S. Census data (Census Bureau 2010). 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 (Atwal et al. 2022). 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 (Bast et al. 2015; Bakillah et al. 2014), 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) (Luxen and Vetter 2011) and use it as the edge feature for the geo-adjacency network of GMEL. OSRM also relies on the maps

Table 2 Descriptive statistics of ground truth 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

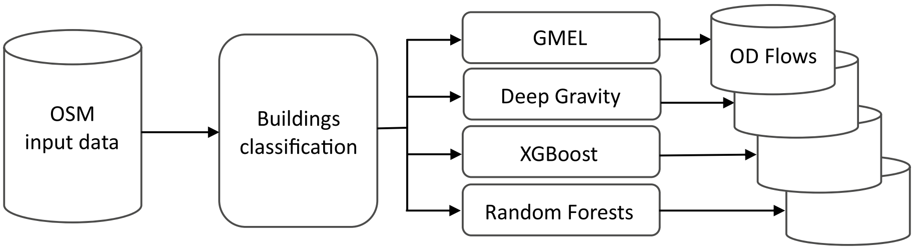


Fig. 2 The workflow of solving commuting flow prediction problem using OSM data with various machine learning models

from the OSM road network for calculating the shortest paths between O-D pairs.

3 Results

The evidence from experiments at multiple scales suggests our approach produces meaningful mobility patterns while providing notable insights into the commuting flows.

3.1 Benchmark results

Table 3 provides the NYC commuting flow prediction accuracy for the state-of-the-art models GMEL and Deep Gravity and out-of-the-box models XGBoost and RF using the OSM data for the same features employed in Deep Gravity. Table 4 provides the NYC commuting flow accuracy for the models using the new set of features based on building type.

Based on the results presented in Table 3, 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 benchmarked using the same set of input features used in Deep Gravity. The experiment shows that GMEL is the best-performing model compared to other methods using the same features.

In the next experiment, we replace Deep Gravity features with our proposed features based on building types. In essence, we enrich the OSM data using the building classification method proposed in (Atwal et al. 2022) and utilize the count, density, and area of residential and non-residential buildings as features for predicting commuting flows. Besides these features, we use population and distance features, which are also used in Deep Gravity. Compared to the prediction accuracy in Table 3 using Deep Gravity features, results in Table 4 show that with all else equal, all models outperform when using building types features except RF, which performs comparatively for both data sets. Therefore, we conclude that the type of buildings in an area is a major driving factor for commuting, and including such features can help predict the corresponding flows more accurately.

Table 3 Evaluation of different models using Deep Gravity features for predicting NYC commuting flows

Model	RMSE	CPC	R^2
GMEL	2.393	0.491	0.486
Deep Gravity	3.326	0.305	0.039
XGBoost	3.172	0.245	0.063
RF	3.204	0.225	0.049

Table 4 Evaluation of different models using building types features for predicting NYC commuting flows

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

Benchmark results show that GMEL is the best model with a different set of features. The graph attention mechanism of GMEL makes it capable of utilizing the geo-contextual information from the input features. Without the attention mechanism, Deep Gravity and other simpler models lack the spatial correlation between the outflows and inflows of commuting. Additionally, the residential and non-residential building types capture the correlation between home and work locations. Therefore, the nine features derived from building types consistently outperform the 39 features used in Deep Gravity.

3.2 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 (NYC 2015) available only for New York City, with the performance of GMEL using OSM data with buildings classification enrichment. 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 based on building types.

Table 5 shows that a comparable level of prediction accuracy can be achieved overall when using features derived from publically 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.

Table 5 Comparison of OSM and PLUTO data using GMEL model for NYC

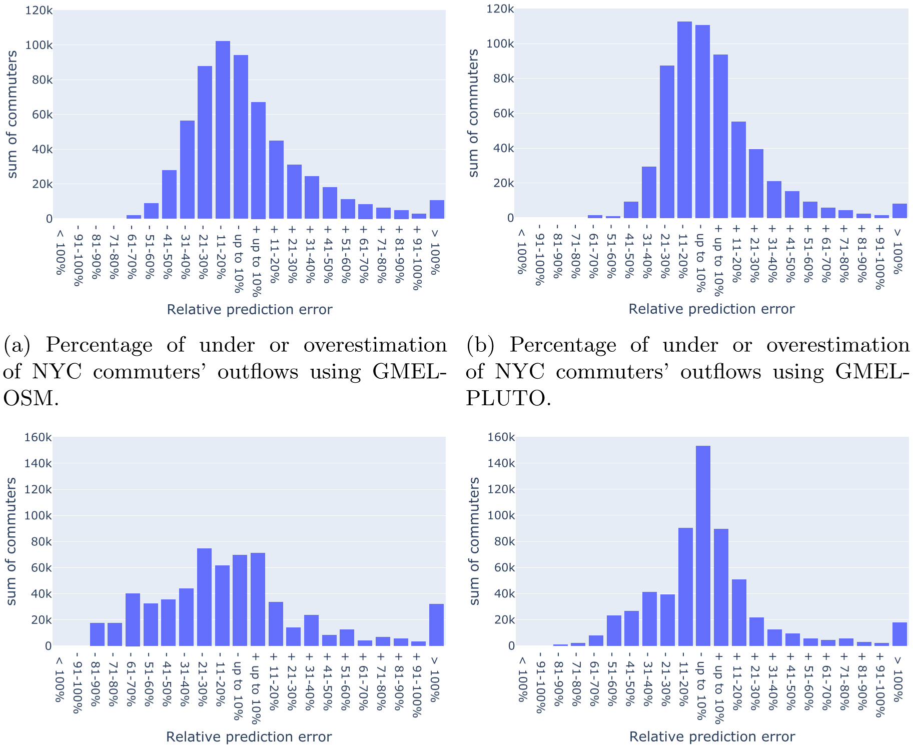
Features	RMSE	CPC	R^2
GMEL-OSM	2.279	0.495	0.535
GMEL-PLUTO	2.084	0.536	0.611

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 3 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 (Fig. 3a and c) and for GMEL-PLUTO (Fig. 3b and d). 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 Sect. 2.3 on the data). 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, Fig. 4 shows scatter plots comparing the ground truth flows against the predicted flows using GMEL-OSM (Fig. 4a and c) and GMEL-PLUTO (Fig. 4b and d). 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 results for outflows, GMEL-PLUTO (65 custom feature model) seems to perform better when predicting the inflows, essentially confirming the results of Fig. 3 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 Sect. 2.3 on the data.



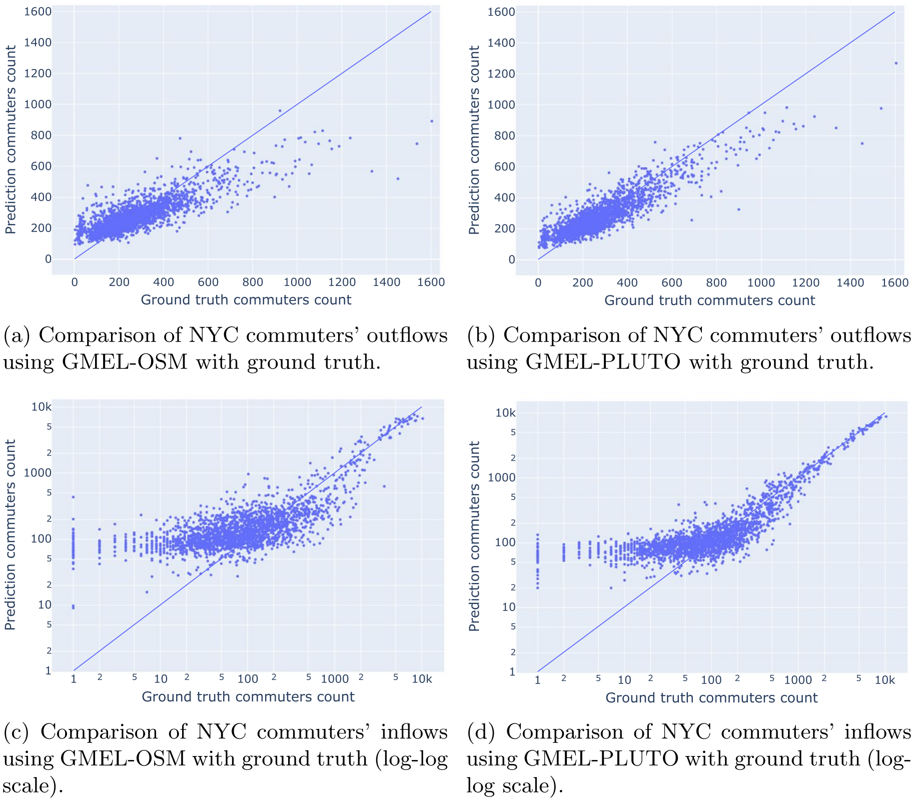
(a) Percentage of under or overestimation of NYC commuters' outflows using GMEL-OSM.

(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. 3 Comparison of GMEL-OSM and GMEL-PLUTO commuters under or overestimation in NYC flows



(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).

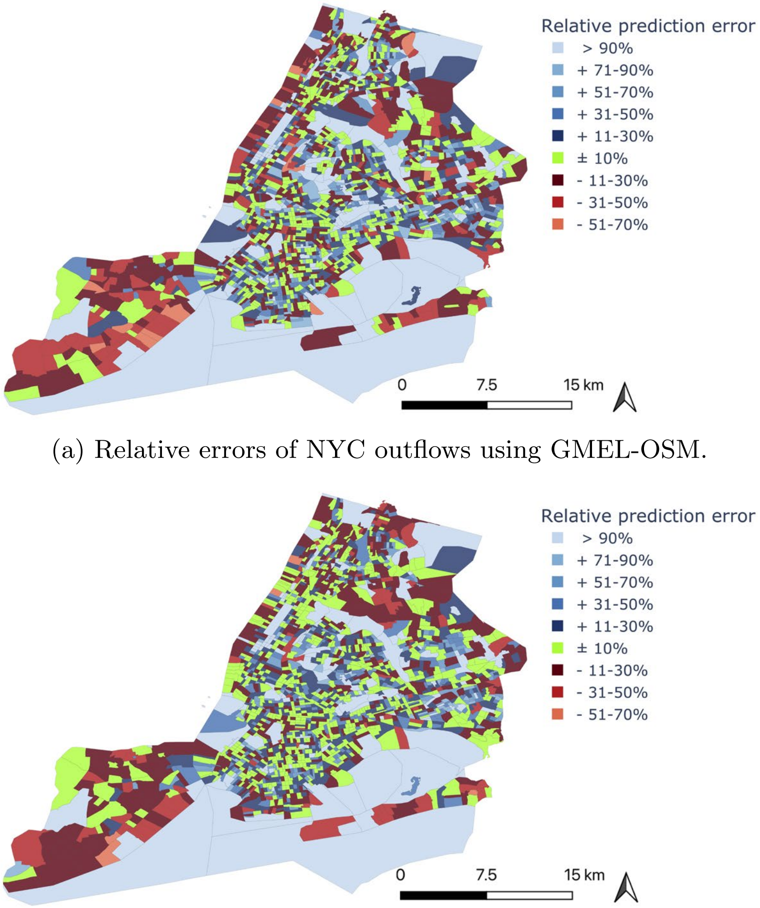
Fig. 4 Comparison of GMEL-OSM and GMEL-PLUTO commuters with ground truth in NYC flows

We can also map the differences between predicted and ground truth outflows as presented in Fig. 5 and inflows presented in Fig. 6. 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 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 Figs. 5 and 6, 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*) 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 6 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 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) Relative errors of NYC outflows using GMEL-OSM.

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

Fig. 5 Comparison of GMEL-OSM and GMEL-PLUTO in NYC outflows. Plotly version 5.13.0 was used to create the maps

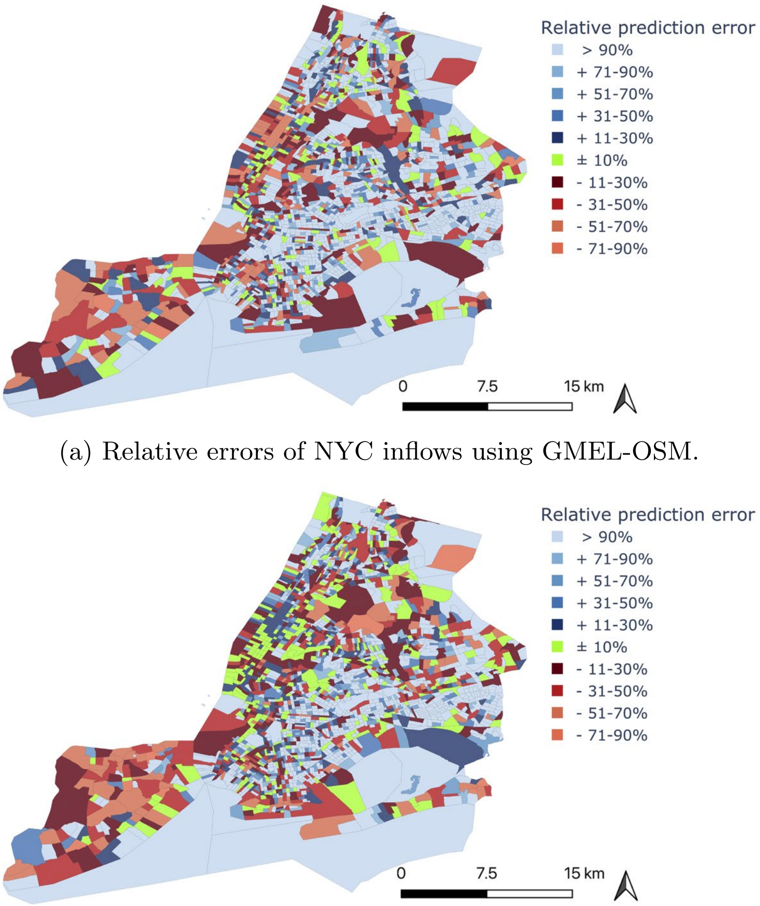
a non-zero flow from the Origin Median. Note that we round predictions to the nearest integer for this experiment, such as that a predicted zero flow is equivalent to a predicted flow of less than 0.5 individuals. We observe that for GMEL-PLUTO, the accuracy is higher at 79.6%, indicating that the model 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 publicly available OSM data. 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 Fig. 7 show the relative percentage errors of outflows and inflows at the tract level compared to the ground truth. Figure 8 demonstrates the



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

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

Fig. 6 Comparison of GMEL-OSM and GMEL-PLUTO in NYC inflows. Plotly version 5.13.0 was used to create the maps

Table 6 Single origin and destination census tract predictions

Census tract	**Approach**	**Zero flows Count**	**Non-zero flows Count**	**Sum of Commuters**
		(Matching)	**(Matching)**	
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

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 7 shows that the model trained in NYC and

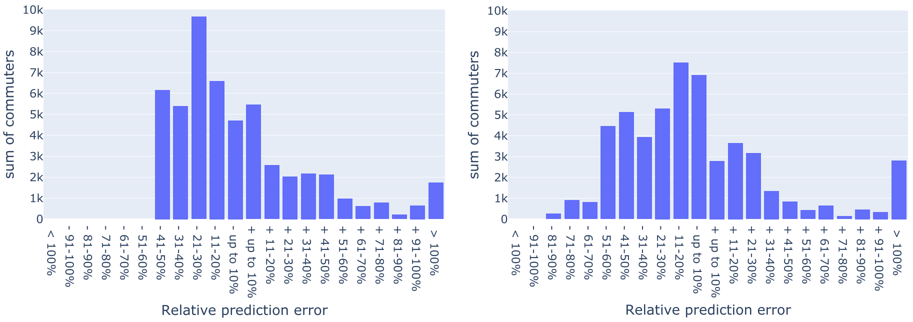
Table 7 Comparison of GMEL-OSM in Fairfax using transfer learning

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

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.

4 Discussion

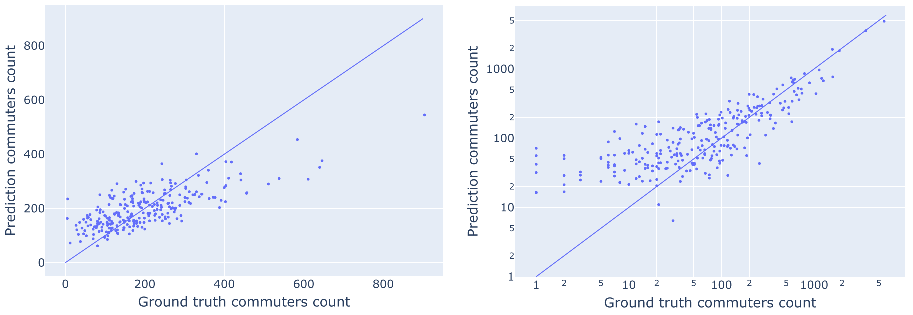
Results for the two study areas show that commuting flows can be accurately predicted using features derived from publicly available OSM data, which is regularly updated by volunteers. We show that the enrichment of OSM data with building types significantly improves the prediction accuracy of commuting flows. By utilizing a building classification method, our results outperform Deep Gravity features also obtained from OSM. Therefore, we illustrate that residential and non-residential building types in census tracts are crucial in predicting commuters' mobility. 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 (Feng et al. 2021) 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 (Zeng et al. 2022), disease transmission (Ferguson et al. 2006), and other applications (Li et al. 2022; Delventhal et al. 2022). However, since our approach relies on types of buildings, it requires building footprints data to derive the input features for the



(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. 7 Commuters under or overestimation using GMEL-OSM for Fairfax



(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. 8 Comparison of GMEL-OSM commuters prediction with ground truth for Fairfax flows

machine learning models, making it applicable to urbanized regions only. Nonetheless, our work demonstrates the significance of building types in predicting commuting mobility. In summary, 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 this divergent phenomenon, matching the trend of outflows and inflows in the ground truth. Additionally, 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 (Bast et al. 2015), 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) (Lundberg and Lee 2017). It would help us understand which other 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) (Cai et al. 2022) and land use (Lee and Holme 2015; Horner 2004) for predicting flows. Therefore, we would explore types of PoIs and land use as other characteristics driving mobility. Particularly, since the descriptive statistics of our study areas show the concentration of destination flows, we would like to further investigate this phenomenon. And, to improve the prediction accuracy of incoming flows, we will consider additional features to capture this behavior, such as commercial or workplace PoIs, economic characteristics, socio-economic indicators, etc. Another step in this direction is to use fine-grained building types such as industrial, educational, commercial, religious, etc. 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. Specifically, we would like to investigate the scale of transferability of our approach and come up with localized models according to urbanization characteristics and data availability in other regions.

Acknowledgements
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, 2018631, and 2109647).

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

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

Authors' contributions
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.

Funding
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".

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

Declarations

Competing interests
The authors declare no competing interests.

Received: 1 July 2024 Revised: 12 December 2024 Accepted: 31 December 2024
Published online: 20 January 2025

References

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

Alonso, W. (1971). The system of intermetropolitan population flows. Institute of Urban and Regional Development.

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

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

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

Bast, H., Storandt, S., & Weidner, S. (2015). Fine-grained population estimation. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 1–10). New York: Association for Computing Machinery.

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

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

Census Bureau, U. (2010). Population data. https://data.census.gov/table. Accessed 3 June 2024.

Census Bureau, U. (2015). Lodes data. https://lehd.ces.census.gov/data/. Accessed 3 June 2024.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York: Association for Computing Machinery.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science, 7*, e623.

Credit, K., & Arnao, Z. (2023). 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.

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

F de Arruda, H., Reia, S.M., Ruan, S., Atwal, K.S., Kavak, H., Anderson, T., & Pfoser, D. (2024). An openstreetmap derived building classification dataset for the united states. *Scientific Data, 11*(1), 1210.

Fairfax County, G. (2024). Fairfax county open geospatial data. https://www.fairfaxcounty.gov/maps/open-geospatial-data. Accessed 3 June 2024.

Feng, J., Li, Y., Lin, Z., Rong, C., Sun, F., Guo, D., & Jin, D. (2021). 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.

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

Fonte, C., Minghini, M., Antoniou, V., Patriarca, J., & See, L. (2018). Classification of building function using available sources of vgi. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, 209–215.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.

Hancock, G. R., & Freeman, M. J. (2001). 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.

Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap. *Nature Communications, 14*(1), 3985.

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

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

Koca, D., Schmöcker, J. D., & Fukuda, K. (2021). 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). IEEE.

Kotsubo, M., & Nakaya, T. (2021). Kernel-based formulation of intervening opportunities for spatial interaction modelling. *Scientific Reports, 11*(1), 950.

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

Lee, M., & Holme, P. (2015). Relating land use and human intra-city mobility. *PloS one, 10*(10), e0140152.

Lenormand, M., Huet, S., Gargiulo, F., & Deffuant, G. (2012). A universal model of commuting networks. *PLoS ONE, 7*(10), 1–7.

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

Li, M.-H., Chen, B.-Y., & Li, C.-T. (2022). 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). IEEE.

Liang, Y., Ouyang, K., Sun, J., Wang, Y., Zhang, J., Zheng, Y., Rosenblum, D., & Zimmermann, R. (2021). Fine-grained urban flow prediction. In *Proceedings of the Web Conference 2021* (pp. 1833–1845). New York: Association for Computing Machinery.

Liu, X., & Long, Y. (2016). Automated identification and characterization of parcels with openstreetmap and points of interest. *Environment and Planning B: Planning and Design, 43*(2), 341–360.

Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., & Silva, C. (2020). Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 34, pp. 808–816). Washington: AAAI Press.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 4768–4777). Curran Associates Inc.

Luxen, D. & Vetter, C. (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 513–516). New York: Association for Computing Machinery.

Masucci, A. P., Serras, J., Johansson, A., & Batty, M. (2013). Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 88(2), 022812.

Morton, A., Piburn, J., & Nagle, N. (2018). 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)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

NYC, D. o. C. P. (2015). Pluto and mappluto. https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page. Accessed 3 June 2024.

N.Y.C, O. (2024). Nyc opendata. https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh. Accessed 3 June 2024.

OSM, c. (2024). Openstreetmap. https://www.openstreetmap.org. Accessed 3 June 2024.

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

Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C., & Toroczkai, Z. (2014). Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature communications, 5*(1), 1–9.

Robinson, C. & Dilkina, B. (2018). A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 1–8). New York: Association for Computing Machinery.

Rodrigue, J.-P. (2020). *The geography of transport systems*. Routledge.

Rong, C., Feng, J., & Ding, J. (2023). Goddag: Generating origin-destination flow for new cities via domain adversarial training. *IEEE Transactions on Knowledge and Data Engineering, 35*(10), 10048–10057.

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

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

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

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

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

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

Vargas-Munoz, J. E., Srivastava, S., Tuia, D., & Falcao, A. X. (2020). Openstreetmap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine, 9*(1), 184–199.

Yang, Y., Herrera, C., Eagle, N., & González, M. C. (2014). Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports, 4*(1), 5662.

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

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

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

Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 31). Palo Alto: AAAI Press.

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

Zhou, Q., Zhang, Y., Chang, K., & Brovelli, M. A. (2022). Assessing OSM building completeness for almost 13,000 cities globally. *International Journal of Digital Earth, 15*(1), 2400–2421.

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.