Computers, Environment and Urban Systems 77 (2019) 101354

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Trip distribution modeling with Twitter data

Nastaran Pourebrahim[a,*], Selima Sultana[a], Amirreza Niakanlahiji[b], Jean-Claude Thill[c]

[a] University of North Carolina at Greensboro, Department of Geography, Environment, and Sustainability, 1400 Spring Garden Street, Greensboro, NC 27412, United States of America
[b] University of North Carolina at Charlotte, Department of Software and Information Systems, 9201 University City Blvd, Charlotte, NC 28223, United States of America
[c] University of North Carolina at Charlotte, Department of Geography and Earth Sciences, 9201 University City Blvd, Charlotte, NC 28223, United States of America

ARTICLE INFO

Keywords:
Machine learning
Artificial neural networks
Random forests
Travel demand modeling
Social media
Volunteered geographic information
Twitter

ABSTRACT

Integrating both traditional and social media data, this study compares the performance of gravity, neural network, and random forest models of commuting trip distribution in New York City. Trip distribution modeling has primarily employed traditional data sources and classical methods such as the gravity. However, with the emergence of social media during the past decade, the potential for integrating traditional and social media data while utilizing new techniques has been identified. Our findings indicate that the random forest model outperforms the traditional gravity and neural network models. Population, distance, number of Twitter users, and employment were identified as the four most influential predictors of trip distibution by the random forest model. While Twitter flows did not enhance the models' performance, the importance of the number of Twitter users at work destinations implies the potential for using social media data in travel demand modeling to improve the predictive power and accuracy.

1. Introduction

Worldwide increases in traffic congestion and air pollution in urban areas presents a need to better understand mobility patterns of urban populations and their travel demands (Shirzadi Babakan, Alimohammadi, & Taleai, 2015; Yang, 2013). A number of studies have examined either individual or collective mobility patterns at different spatial scales (i.e., Beiró, Panisson, Tizzoni, & Cattuto, 2016; González, Hidalgo, & Barabási, 2008; Hawelka et al., 2014). Mobility information of individuals can be aggregated to study the frequency of travel between different regions, as represented by origin-destination (OD) matrices (Barbosa et al., 2018). An OD matrix provides population flow patterns (trip distribution) studied for diverse purposes such as traffic forecasting, resource allocation, prediction of migration flows, and epidemic spreading (Beiró et al., 2016; Pourebrahim, Sultana, Thill, & Mohanty, 2018). Therefore, improving flow estimations has become critical across various domains of application (Barbosa et al., 2018).

Various trip distribution models have been developed over past decades to estimate population flows with greater accuracy (de Dios Ortuzar & Willumsen, 2011; Roy & Thill, 2003; Simini, González, Maritan, & Barabási, 2012; Wilson, 1970; Wilson, 1998; Zipf, 1946). Most of these models are heavily dependent on conventional data such as population censuses or travel diary surveys. The emergence of social media and location-based services in recent years has introduced new opportunities to the field of transportation (Yang, Jin, Cheng, Zhang, & Ran, 2015). Geospatial big data such as taxi trajectories, mobile phone records, and social media messages have attracted scholars to observe, understand, and visualize (i.e., Karduni et al., 2017) human activities in cities at fine spatio-temporal scales (Liu et al., 2015). These data significantly improve the visualization of human mobility patterns, yet there is a need to better understand and contextualize them in different steps of the travel demand modeling framework (Anda, Erath, & Fourie, 2017).

Traditionally, the gravity model and its derivatives have been used as the most reliable approach to predict trip distribution at fine spatial scales, such as commuting flows within cities (Lenormand, Bassolas, & Ramasco, 2016). The potential for developing hybrid approaches that integrate the vast volume of social media data with the gravity model has recently been noted (Beiró et al., 2016). While traditional models have used statistical methods rooted in sound mathematical foundations, they have been unable to account for nonlinearities and other irregularities in data (Golshani, Shabanpour, Mahmoudifard, Derrible, & Mohammadian, 2018). To alleviate these and other issues, Machine Learning (ML) techniques have been applied in different urban and transportation domains (i.e., Ghasri, Hossein Rashidi, & Waller, 2017; Karimi, Sultana, Shirzadi Babakan, & Suthaharan, 2019). A significant

* Corresponding author.
E-mail addresses: n_poureb@uncg.edu (N. Pourebrahim), S_sultan@uncg.edu (S. Sultana), aniakanl@uncc.edu (A. Niakanlahiji), Jean-Claude.Thill@uncc.edu (J.-C. Thill).

https://doi.org/10.1016/j.compenvurbsys.2019.101354
Received 24 February 2019; Received in revised form 15 June 2019
Available online 02 July 2019

body of literature exists at this time where various ML techniques have been evaluated on their ability to model travel demand, such as artificial neural networks (ANNs) (i.e., Ding, Wang, Wang, & Baumann, 2013; Mozolin, Thill, & Lynn Usery, 2000; Pourebrahim et al., 2018; Tillema, van Zuilekom, & van Maarseveen, 2006) and tree-based ensemble methods (i.e., Ghasri et al., 2017; Rasouli & Timmermans, 2014). While random forests (RFs) (Breiman, 2001) have been identified among the most advanced and most efficient ensemble methods for data classification and regression (Ghasri et al., 2017), they have so far been used in only a few studies of travel demand. Ghasri et al. (2017) and Rasouli and Timmermans (2014) have reported promising results with RF modeling of trip generation and modal split, yet their suitability and usefulness in trip distribution analysis remains to be thoroughly assessed.

Given the current state of research, our objective is to compare the performance of gravity, neural network, and random forest models of commuting trip distribution while combining both traditional and social media data. We also evaluate how information on personal mobility derived from social media affects commuting trip distribution by identifying the importance of different variables. To the best of our knowledge, this paper is one of the first to use machine learning approaches in trip distribution forecasting with social media data. The main contributions of this paper are threefold: (1) revealing the potential of social media data in trip distribution modeling at census tract level; (2) using machine learning techniques to predict trip distribution at census tract level; and (3) comparing the performance of gravity, neural network and random forest models to identify the best model for predicting trip distribution at census tract level. The paper is organized as follows. The review of related work is provided in section 2, followed by a presentation of the study area, data sources and methodology in section 3. Results are presented in section 4, with a discussion and concluding remarks in sections 5 and 6.

2. Related work

2.1. Travel demand modeling

The relationship between personal mobility flows and a range of personal and environmental factors has been studied to determine future travel demands within cities (Barbosa et al., 2018). Travel demand modeling has long been dominated by the four-step model with its steps being trip generation, trip distribution, modal split, and traffic assignment (McNally, 2007). The objective of this model is to estimate the traffic in the transportation networks. The model first identifies the amount of travel produced by each traffic zone in the area (trip generation), which is then distributed among other zones (trip distribution). Trip distribution accounts for network effects on personal mobilities and responds to the connectivity qualities of travel opportunities across the urban space. A broad family of models of spatial interaction, including the ubiquitous gravity model, has been developed to analyze and forecast urban trip distribution (de Dios Ortuzar & Willumsen, 2011; Roy & Thill, 2003; Simini et al., 2012; Wilson, 1970; Wilson, 1998; Zipf, 1946).

In his pioneering work, Zipf (1946) suggests that the number of persons moving between two communities is proportional to the product of their populations and is inversely related to the distance between them. This basic model has been extended over time by adding factors other than population. These factors that determine trip production rate at origins and trip attraction volumes at destinations have been examined in past studies. Population, housing type, household income, and car ownership have been identified as important factors determining volumes of trip production in the origin zone for commuting trips (Berger, 2012). Employment, density and land use, and points of interest have been identified as attraction factors of the destination zones (Berger, 2012; Yang, Herrera, Eagle, & González, 2015).

2.2. Social media and human mobility

Cities with large populations are early adopters of new information and communication technologies (Barbosa et al., 2018). As such, they have more mobile phone users, generating a huge volume of data through various social networking applications. Social media data, particularly Twitter, have attracted many scholars to explore human mobility at various spatial scales. For example, Hawelka et al. (2014) employed Twitter data to examine mobility patterns of international travelers. Several measures were explored including mobility rate, radius of gyration, diversity of destinations and the balance of inflows and outflows. Similarly, Kurkcu, Ozbay, & Morgul (2016) studied radius of gyration and user displacement for Twitter users in New York City. The research showed that mobility patterns of Twitter users follow the Lévy Flight and that the mobility flows estimated from Twitter posts are similar to ground-truth home-to-work trips at the county level. In their analysis of human mobility and activity patterns, Hasan, Zhan, and Ukkusuri (2013) studied the distribution of different activity categories using Twitter data, as well as the temporal, spatial and frequency distributions of places visited in New York, Chicago and Los Angeles. Liu, Zhao, Khan, Cameron, and Jurdak (2015) used Twitter data to investigate population distribution and human mobility flows in Australia at national, state, and metropolitan scales. Their study found that population distribution can be estimated from Twitter data at coarse spatial granularity.

While mobility patterns identified from Twitter data have been commonly reported through simple measures such as those mentioned above, a few studies have used more advanced modeling perspectives. Liu, Liu, et al. (2015) reported a high correlation between the mobility flows extracted from Twitter posts and the flows estimated by a gravity model. More recently, McNeill, Bright, and Hale (2017) determined that estimated Twitter commuting flows outperform the radiation model, especially for short trips with higher volume of commuters. Kim, Park, and Lee (2018) compared different possible predictors in a gravity model of inner-city traffic. They used the resident population, the number of employees, and the number of tweets as proxies of mass values and found the number of tweets to be the most powerful predictor. Employing Foursquare user check-in data, Yang, Jin, et al. (2015) combined clustering, regression, and gravity models to estimate an OD matrix of non-commuting trips in the Chicago urban area. The study concluded that the estimated OD matrix is similar to the ground-truth OD flow matrix. Beiró et al. (2016) integrated Flickr data with a standard gravity model under a stacked regression procedure. The results showed that the hybrid gravity model outperforms the traditional gravity model. The research validated the performance of the model using two ground-truth datasets of air travel and daily commuting in US counties. Utilizing Twitter data, Pourebrahim et al. (2018) compared gravity and neural network models to predict the commuter trip distribution in New York City. While these models had low predictive power, the findings indicated that adding Twitter data enhanced the performance of both models. The promising results achieved in these studies show the potential for enhancement in traditional modeling with social media data. However, more research is needed to fully grasp the scope of the predictive value of social media in estimating mobility patterns at fine spatial granularity. Applying new techniques and integrating traditional and new datasets are the first steps towards this goal.

2.3. Artificial neural networks and decision trees

Artificial neural networks (ANNs) started to be used in transportation research, including travel demand modeling, as an alternative to more conventional generalized linear models (GLM) and other econometric techniques from the beginning of the 1990s. Because ANNs have an inherent and demonstrated capability to capture nonlinearities and to be robust to alternative distributional properties of the data, they are

found attractive for policy and planning analysis (Golshani et al., 2018). The usefulness of ANNs in trip distribution analysis and forecasting has been a focus of research, but reported results remain mixed. Black (1995) compared a gravity model and ANNs for a three-region flow problem and a nine-region commodity flow problem. Better performance of ANNs was reported for flow prediction in comparison to the gravity model. Similarly, Celik (2004) showed that ANNs outperform a regression model in predicting short-term inter-regional commodity flows. However, Tillema et al. (2006) reported that for a fifteen-region trip distribution in Rotterdam Rijnmond, ANNs perform better than gravity models only when data are limited. In another study, Mozolin et al. (2000) compared the performance of ANNs to the gravity models for commuter trip distribution among counties of the Atlanta Metropolitan Area, as well as among census tracts. The study suggested that although ANNs may fit the data better, their predictive accuracy is poor due to uncontrollable over-fitting and lack of generalization power, which cast doubt on the longitudinal transferability of ANN forecasts. Similar results were observed by Pourebrahim et al. (2018) who investigated commuting trip distribution between the census tracts of New York City.

Empirical results show the uncertainty about the potential contribution of ANNs in the context of trip distribution. In addition to problems of over-fitting and lack of generalization (Mozolin et al., 2000), the perception of ANNs as a "black box" makes it difficult to understand and utilize the results for planning purposes (Tillema et al., 2006). Decision trees are another class of techniques that have been used to model discrete decisions in travel demand. Thill and Wheeler (2000a, 2000b) demonstrated the merit of decision tree induction learning of spatial choice behavior in Minneapolis-St. Paul. Pitombo, de Souza, and Lindner (2017) compared these models to traditional gravity models of trip distribution in Bahia, Brazil, and concluded they exhibit better accuracy when prediction of destination choices is assessed by aggregate metrics such as trip length distribution and goodness-of-fit measures. More recent tree-based ensemble methods have been widely used in various machine learning problems due to their simplicities and understandability. Decision tree ensembles have shown promising results in studying travel time prediction (Zhang & Haghani, 2015), trip generation (Rashidi & Mohammadian, 2011), and mode choice modeling (Rasouli & Timmermans, 2014). As a specific class of decision tree ensembles, random forests (RFs) have been applied in predicting traffic flow (Leshem & Ritov, 2007) and travel time (Hamner, 2010), but only a few studies have applied the technique in travel demand modeling. Rasouli and Timmermans (2014) and Sekhar, Minal, and Madhu (2014) identified that RF outperforms other methods in mode choice modeling. Similarly, Ghasri et al. (2017) reported that RF shows high accuracy in estimating the total number of trips and trip attributes in a tour of trips at a disaggregate individual level. Following this research paradigm; we explore the performance of RF compared to the gravity and ANN models in trip distribution modeling.

3. Methodology

We have selected New York City (NYC) as our study area (Fig. 1) due to the large volume of readily available Twitter data. We focused on commuting trips because they are temporally stable and account for the largest share of total flows in a population (Yang, Jin, et al., 2015). The census tracts are used as the geographic units for modeling commuting flows in NYC (Fig. 1).

3.1. Data collection and processing

The 2015 LEHD Origin-Destination Employment Statistics (LODES) for NYC were obtained as the mobility variable of interest (U.S. Census Bureau, 2015). The dataset reports the home and employment locations of workers, along with other characteristics such as age, earnings, industry distributions, and local workforce indicators. For each employee there is a home and a work census tract representing one commuting flow. The data were aggregated to census tract-to-tract commuting flows. The internal commuter flows, where home and work census tracts of commuters are the same, were excluded. In total, there were 903,685 origin-destination (OD) dyads and 2,580,596 home-work flows between census tracts in NYC that provide the dependent variable of this study. The number of OD dyads was further reduced to 878,132 after removing missing values on the input variables.

The key independent variables were selected based on previous commuting research (i.e., Sultana & Weber, 2007; Sultana & Weber, 2014). These variables include residential population, employment, household median income, household median size, and household median number of vehicles. These data were obtained for the year 2015 from SimplyAnalytics (2015), a socio-economic data provider. Two other input variables, points of interest (POI) and sprawl index for each census tract, were collected from the NYC OpenData (2017) and National Cancer Institute (2014), respectively. The POIs are amenities (i.e., shops, tourist attractions, etc.) that different city agencies consider to be a common place or place/point of Interest. The sprawl index is a measure of compactness of a census tract based on various socio-economic, land use, and street network data (Sultana, Pourebrahim, & Kim, 2018; Ewing & Hamidi, 2014). We calculated the network distance between the centroids of origin and destination census tracts as another independent variable in the dataset. The North America Detailed Streets (ArcGIS, 2018) dataset was used to create the network data and all the procedures and calculations were performed in ArcGIS Network Analyst environment.

Twitter is among the most popular social media platforms that has been successfully used in the past studies (i.e., Pourebrahim et al., 2018; Yang, 2013). Approximately 700 million tweets are posted on Twitter per day by 126 million daily active users (Internet Live Stats, 2019), making it an optimal data source for collection of information including texts, pictures, and geolocation (Pourebrahim, Sultana, Edwards, Gochanour, & Mohanty, 2019). The large volume of readily available Twitter posts in NYC allowed us to conduct this research. Geolocated tweets posted within NYC from June 2015 to May 2016 were obtained from the SOPHI data lake maintained by the Data Science Initiative (DSI) at the University of North Carolina at Charlotte. The dataset includes 1% sample of all tweets generated on Twitter during the timeframe. The tweets with precise location (longitude and latitude) were used, which resulted in 2,052,599 usable tweets. We then calculated the number of tweets and unique individual users in origin and destination census tracts. We kept the unique individual users (Twitter population) in the final analysis due to the high correlation ($r = 0.93$, $p < .05$) between the two variables. In addition, multicollinearity was tested by computing the variation inflation factor (VIF) of the independent variables. A VIF of < 2 for all the independent variables confirmed the absence of multicollinearity.

We extracted Twitter flows as another input variable by converting the Twitter dataset into an origin-destination matrix where origin represents home census tract and destination represents work census tract. Three filters were applied to the dataset: 1) Only users who posted geolocated tweets in two or more different census tracts were included; 2) The average length of stays for NYC international and domestic tourists is 9 and 1.9 days, respectively (Josephs, 2017). Therefore, Twitter users with a time interval of < 10 days between their first and last tweets were considered tourists and removed from the dataset; and 3) Oxford Internet Institute identifies Twitter users with an average of > 50 or 100–250 posts per day as bots or cyborgs, respectively (Nimmo, 2017; Nimmo, 2019). We used daily posting rate of 50 as the cutoff to remove suspicious bot activities. We also manually reviewed the "source" field ("generator" in our dataset) in the collected tweets following the work of Tsou, Zhang, and Jung (2017) to identify the bots or cyborgs with commercial purposes. For example, if a tweet was created on an iPhone device, the source field will be "Twitter for iPhone". However, an advertisement tweet could have "TweetMyJOBS"

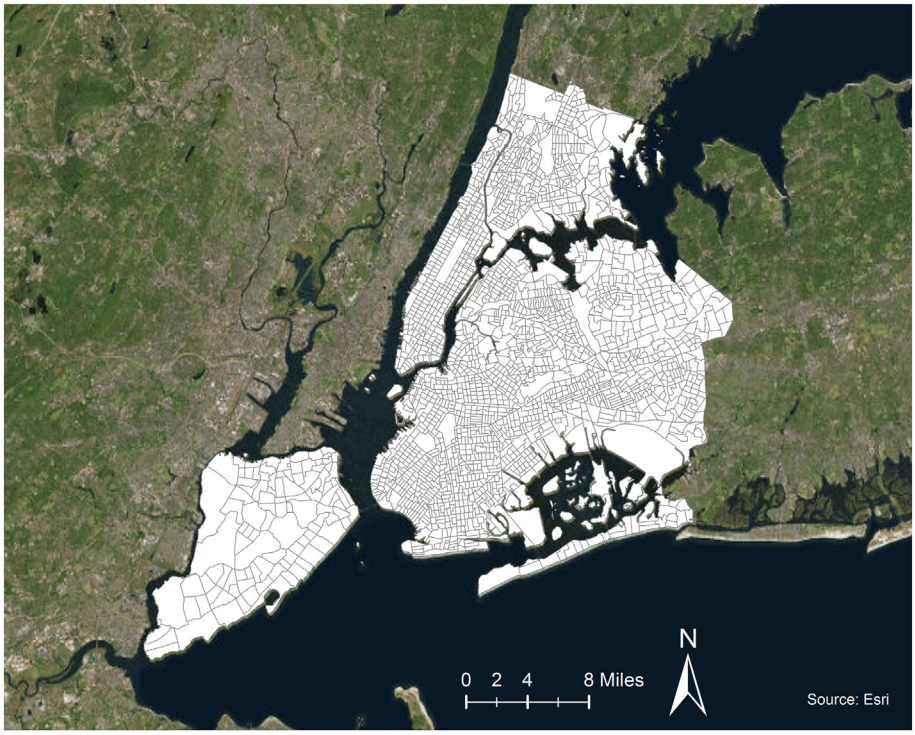

Fig. 1. Study area (NYC census tracts).

or "dlvr.it" in its source field (Tsou et al., 2017). These noises were classified as advertisement, traffic, news, weather, and job. The total of 117,190 tweets were removed based on 15 identified source of noises that include dlvrt.it, pinprick on iOS, Squarespace, COS App, kickalert, Beer Menus, dine here, Dance Deets, 511NY-Tweets, TTN NYC traffic, Cities, eLobbyist, iembot, TweetMyJOBS, and SafeTweet by TweetMyJOBS.

The 1 year timeframe of Twitter dataset allowed us to extract sufficient geolocated tweets of individuals in order to identify their home and work locations. The home and work census tracts for each user were identified and the direction of commuter flow was considered from home to work census tract. The home-work flows then were summed for all individuals having similar home-work locations. We identified the home and work census tracts of the Twitter users based on the common method of frequency counts with temporal (day-night) filtering (i.e., Jiang, Li, & Ye, 2018; McNeill et al., 2017). The most visited census tract at night (00:00–7:00) was considered the home location. If the frequency of visits to different census tracts were similar, the centroid of all the census tracts (represented as points) with the highest similar number of tweets was calculated to identify the home census tract. The most visited census tract during the day (8:00–17:00) excluding weekends and national holidays (Thanksgiving and Christmas) was considered the work location. Similarly, if the frequency of visits to different census tracts was the same, centroids were identified as the work census tract. The home census tracts for 9076 Twitter users and work census tracts for 7268 Twitter users were identified by centroid calculation. All other locations were assumed to be areas visited which are unrelated to either living or working. The users whose home and work locations could not be identified were removed from the analysis resulting in a total of 31,820 Twitter users. Internal flows in census tracts were also excluded from the dataset, which retained 31,688 home-work flows in the final dataset. These flows were then aggregated at census tract level to identify census tract-to-tract Twitter flows.

3.2. Trip distribution modeling

There was a high volume of zero Twitter flows in many census tracts in our dataset; we, therefore, first developed the three models only for those ODs (7445 ODs) that have both LODES flows and Twitter flows. These models were developed once without Twitter data, and then Twitter flow was added as a separate independent variable to the models. Since Twitter flow did not improve the models and the mean square errors (MSE) were large (see section 4), we dropped the variable in our final analyses. The final models were developed for 878,132 OD dyads first without the Twitter data discussed in the previous section and then with the addition of Twitter population in home and work census tracts. The next sections present the developed models, based on our final dataset that includes both non-Twitter data and Twitter population.

3.2.1. Gravity model

The gravity model used here can be formulated as follows (Eq. (1)):

$$T_{ij} = G \frac{M_i^{\alpha} M_j^{\beta}}{f(d_{ij})} \quad (1)$$

where T_{ij} is the flow between two areas i (origin) and j (destination). In this study, the origin and destination are home and work census tracts and T_{ij} is the total number of home-work flows (commuting flows). M_i and M_j are trip production and attraction factors, respectively. $f(d_{ij})$ is the distance decay function commonly represented as a power function $f(d_{ij}) = d_{ij}^{-\gamma}$. In the original model, M_i and M_j are population of origin and destination. However, the trip production and attraction factors can be extended to other socioeconomic factors. Here, residential population, household median income, household median size, household median number of vehicles, and Twitter population at home census tract were used as trip production factors; also, employment, POI, sprawl, and Twitter population at work census tract served as trip attraction factors. The production and attraction factors and network distance between the centroids of census tracts were used to estimate the interzonal home-work flows in the gravity model. The model can be

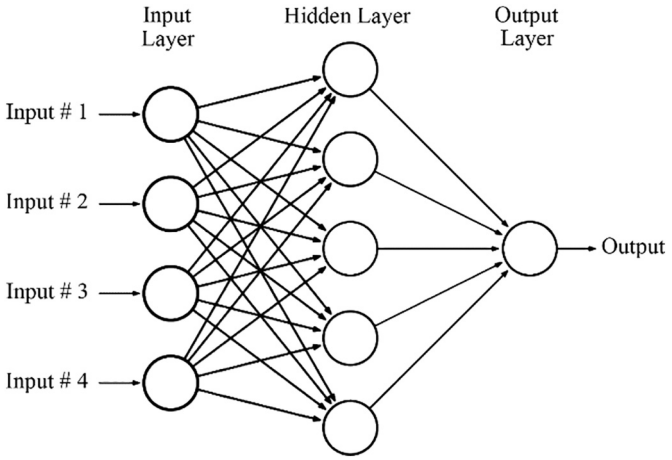

Fig. 2. Multi-layer perceptron (MLP) structure.

adjusted using a linear regression in the logarithmic scale (i.e., Beiró et al., 2016). Therefore, the final model is a linear combination of all variables in a log-log form. We quantified the estimation and compared the gravity model's performance with the ANN and RF models using the mean square error (MSE) and the coefficient of determination (R^2).

3.2.2. Artificial neural networks (ANNs)

ANNs are computational models that learn from and recognize patterns in data by mimicking the structure and functions of the nervous system in the brain (Nielsen, 1987). The fundamental building block of neural networks is the single-input neuron with three processes including the net input function, the weight function, and the transfer function. The output of a neuron with n inputs is calculated as follows (Eq. (2)) (Beale, Hagan, & Demuth, 2015):

$$a = f\sum_{i=1}^{n} (w_i p_i + b) \quad (2)$$

where, a is output, p_i is input value, w_i is the weight, b is the bias and f is a transfer function of the neuron.

A widely used ANN topology is the multi-layer perceptron (MLP) (Fig. 2) (Rumelhart & Mcclelland, 1986). It has been used in approximately 70% of ANN studies (de Oña & Garrido, 2014). The basic model of an MLP has three layers including input, hidden, and output. Each layer consists of neurons that are interconnected by weighted links that perform parallel distributed processing to solve a problem. The number of neurons in the input and output layers is usually determined by the number of predictor and predicted variables in the model (Amita, Singh, & Kumar, 2015). By adjusting the links' weights, a neural network learns the correlation between input and output. The learning process is much like a reward and punishment process so that when a desired/undesired output is generated, the weights related to the input are strengthened/reduced (Ding et al., 2013). Back-propagation (BP) is the most widely used algorithms in the learning process. We developed BP neural networks in MATLAB to predict the home-work flows using the same data used in the gravity model.

The network has one hidden layer with 10 neurons in the input layer and one neuron in the output layer, corresponding to the number of input and output variables. The number of hidden layers and their neurons could be different depending on the complexity of connections between the input and output layers. However, a neural network with one hidden layer is a universal function approximator (de Oña & Garrido, 2014). Therefore, we used one hidden layer when developing the ANN model. It is common practice to select the number of hidden neurons by trial and error (Amita et al., 2015). We tested networks with 10 (number of inputs), 20, and 50 hidden neurons. Networks of larger size are impractical because of the excessive computational requirements for their training (Mozolin et al., 2000). Since no improvement in fit was observed with more hidden neurons in the model, we used 10 neurons in the final analysis. We employed two common transfer functions, the log-sigmoid and the linear for the hidden and output layers, respectively, and the Levenberg-Marquardt algorithm for training. The data were randomly divided into 70% for training, 15% for testing, and 15% for validation; the network was trained so that the MSE is minimized. Training was used for fitting and selecting the model, testing was used for evaluating the 'model's forecasting ability, and validation was used for determining the end-point for the training process (minimizing the error) and avoiding overfitting (Srisaeng & Baxter, 2017).

3.2.3. Random forests (RFs)

The RF algorithm proposed by Breiman (2001) has been used for regression and classification, as well as for variable selection (Sulaiman, Shamsuddin, Abraham, & Sulaiman, 2011). An RF (Fig. 3) is an ensemble of independent decision trees growing in parallel on a subsample of the training data (Lagomarsino, Tofani, Segoni, Catani, & Casagli, 2017). RF has a higher degree of accuracy compared to single trees, is effective in prediction, do not overfit, and allows measuring variable importance (Breiman, 2001; Lagomarsino et al., 2017; Sulaiman et al., 2011). The RF procedure includes (1) bootstrap resampling, (2) random variable selection, (3) out-of-bag error

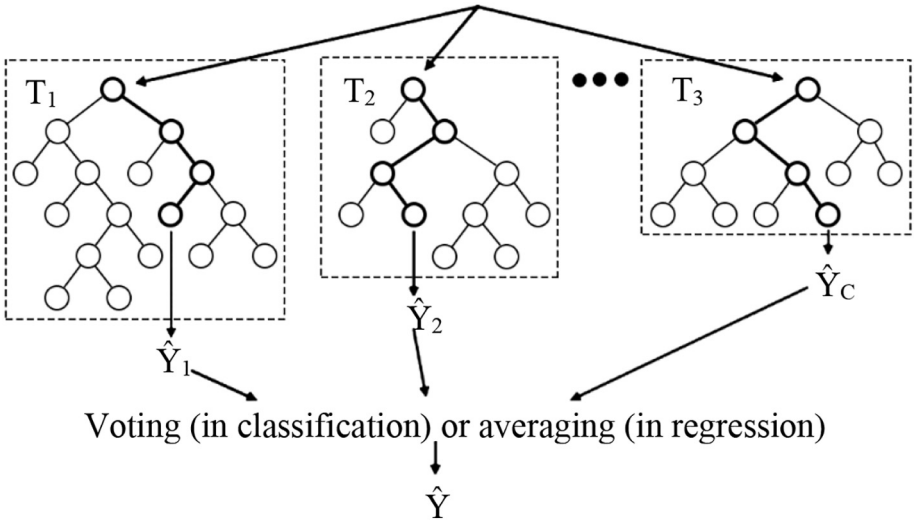

Fig. 3. Random forest structure.

estimation, and (4) full depth decision tree growing (Ahmad, Mourshed, & Rezgui, 2017). RF is an ensemble of C trees $T_1(X)$, $T_2(X)$, …, $T_C(X)$, where $X = x_1, x_2, \ldots, x_m$ is an m-dimension vector of inputs. The output results for the trees (T_1 to T_C) are $\hat{Y}_1 = T_1(X)$, $\hat{Y}_2 = T_2(X)$, …, $\hat{Y}_C = T_C(X)$. The average value of all trees' outputs ($\hat{Y}_1$ to $\hat{Y}_C$) will be the final prediction $\hat{Y}$. An RF generates C number of decision trees from N training samples (Ahmad et al., 2017). For each tree in the forest, the algorithm takes a random sample of observations with replacement from the data and uses a random subset of the predictors at each split (Sulaiman et al., 2011). The model is built from 2/3 of the data (in-bag) and excluded data named out-of-bag (OOB) are used for identifying the prediction error (Ostmann & Martínez Arbizu, 2018). The OOB data eliminates the need for a separate test set for an unbiased estimate of error and can be used to identify the importance of predictors (Breiman, 2001). The relative importance of variables is identified by random permutation of out-of-bag data across each input variable to estimate the increase in the out-of-bag error (Breiman, 2001).

We employed TreeBagger in MATLAB to generate the RF model. Three parameters need to be initialized in RF: (1) the minimum number of observations per tree leaf, (2) the number of randomly selected variables for each decision split, and (3) the number of trees. We trained a random forest with different leaf sizes of 5, 10, 20, 50, and 100. With the leaf size of 5 resulting in the minimum out-of-bag MSE, we set the minimum leaf size to 5. To develop an efficient forest, only a random subset of input variables is selected in RF to find the best splits (Ghasri et al., 2017). The Treebagger considers one third of the number of input variables for regression in each split. We used the same default value. Since increasing the number of trees in RF does not result in overfitting (Breiman, 2001), we trained random forests with 20, 50, 100, and 200 trees. No improvement was observed in the performance (decrease in OOB mean square error) of models with > 100 trees. Therefore, we used 100 trees in the final model.

4. Results

Oure initial analysis was performed based on 7445 ODs that were represented by both LODES flows and Twitter flows. The gravity, ANN, and RF models were developed first with a specification that excludes Twitter data, but includes variables: 1) network distance between ODs; 2) population, household median income, household median size, and household median number of vehicles in origin census tract; and 3) employment, sprawl, and POIs in destination census tract. Then we added the Twitter flow to the specification of each of the three models. The MSE and R^2 for the six models are reported in Table 1.

Contrary to past studies (i.e., Mozolin et al., 2000), the findings indicate the poor performance of the gravity model compared to the ANN and RF models in terms of both MSE and R^2. The lowest MSE and highest R^2 were achieved by the RF model, suggesting that a higher

Table 1
The performance of the gravity, ANN, and RF models.

Model		MSE	R^2
7, 445 ODs	Gravity without Twitter Data	110.23	0.25
	Gravity with Twitter flow	106.10	0.28
	Neural Network without Twitter Data	64.37	0.54
	Neural Network with Twitter flow	66.35	0.54
	Random Forest without Twitter Data	36.60	0.76
	Random Forest with Twitter Flow	37.78	0.75
878,132 ODs	Gravity without Twitter Data	17.33	0.09
	Gravity with Twitter Population	17.23	0.09
	Neural Network without Twitter Data	10.04	0.47
	Neural Network with Twitter Population	7.78	0.58
	Random Forest without Twitter Data	4.51	0.76
	Random Forest with Twitter Population	**4.26**	**0.78**

Note. The results for the gravity models are reported at non-logarithmic scales. Bold values indicate the best performance.

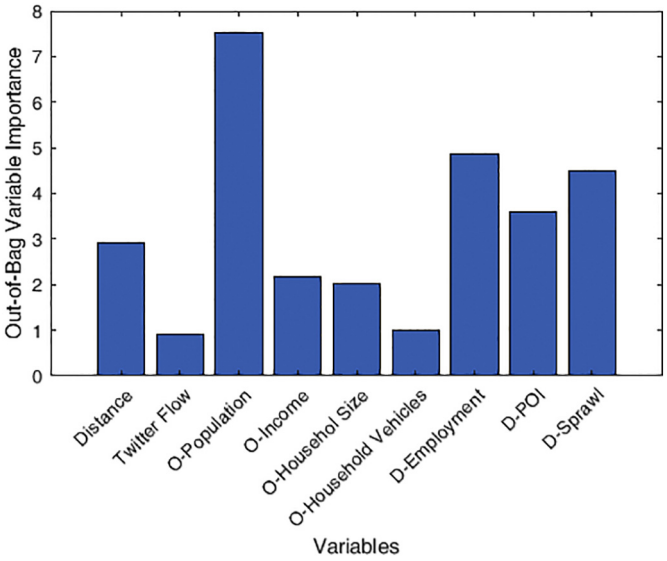

Fig. 4. Relative importance of variables in the RF model for the 7445 OD dyads.

percentage of the home-work flows can be predicted by the input variables in the RF model with a lower error. Adding Twitter data to the ANN and RF not only did not improve the models, but also increased the MSEs for 1.98 and 1.18 respectively. However, the gravity model showed a 3% increase in the R^2 and a decrease of 4.13 in MSE when adding the Twitter flows. These results are in contrast with current literature (i.e., Beiró et al., 2016; McNeill et al., 2017) that reported that flows extracted from social media such as Flickr or Twitter can be a good proxy to ground truth data or can outperform traditional models. Utilizing the relative importance of variables in the RF model, Twitter flow was among the least important variables (Fig. 4). Past studies utilizing Twitter flows have been conducted on coarse spatial granularity, such as county. Our analysis is conducted on census tracts, which may explain that the volume of Twitter flows at this geographical scale is not significant enough to influence the performance of models.

Since Twitter flow did not improve the performance of the models in a meaningful way and the MSE for models were quite large, the variable was dropped from further analysis, which increased the OD dyads to total of 878,132. First, for comparison's sake, we developed the models for these OD dyads excluding Twitter data from the specification. The Twitter population in origins and destinations were then added to the model specifications. The performance of the gravity model remained similar to the previous results (the lowest R^2 and highest MSE), while the RF model showed the best performance (Table 1). ANN and RF performances after adding Twitter population are illustrated in Fig. 5. Predicted home-work flows have large errors in the initial phase of both models due to the small sample size, but errors decrease as the number of iterations increases. Although the performance of both test and validation sets in the ANN model are similar to the training set (no overfitting), the ANN training model does not fit the input data. The MSE for the ANN model is much higher than for the RF model, suggesting the shortcomings of ANNs for trip distribution modeling even with larger datasets. Although the ANN model did not have a good performance, the largest improvement was observed in the ANN model when Twitter population was added. The inclusion of Twitter population increased the predictive power of the ANN and RF models by 11 and 2%, respectively. The MSE decreased from 10.04 to 7.78 for the ANN model and from 4.50 to 4.26 for the RF model. The R^2 of the gravity model did not change with the addition of Twitter population, but the MSE value slightly decreased by 0.10.

Fig. 6 shows the difference between the actual and predicted numbers of incoming flows per census tract (summation of incoming flows to a census tract from all other tracts) for the gravity, ANN, and

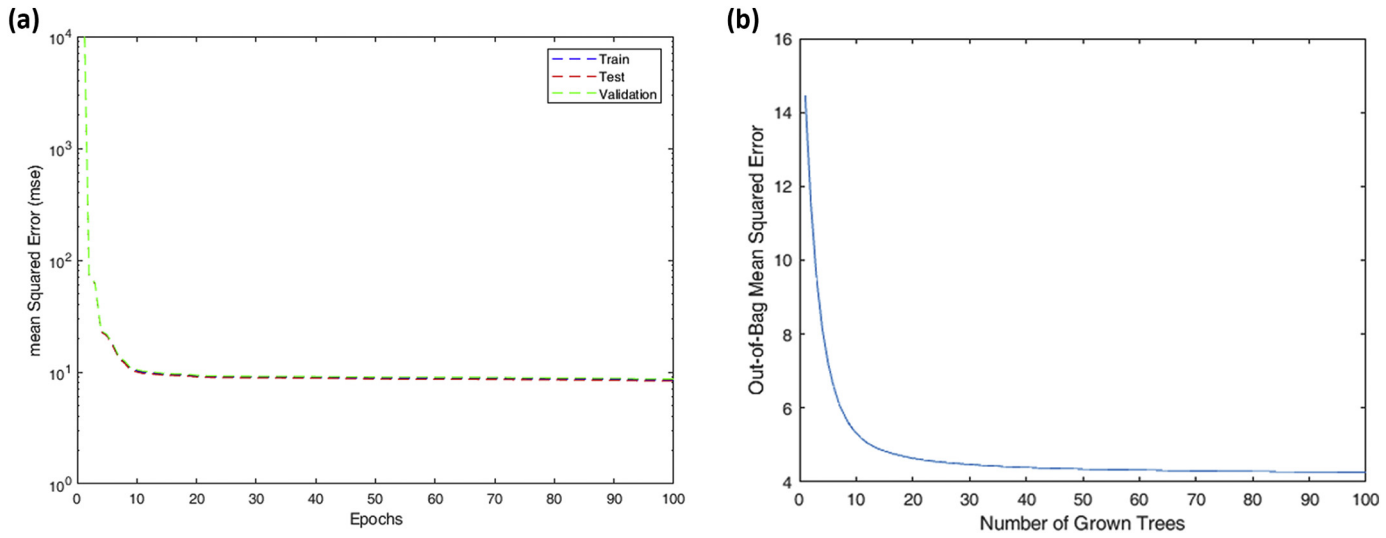

Fig. 5. Neural network (a) and random forest (b) performance (including Twitter population).

RF models developed with Twitter population. While both the gravity model and the ANN predictions show limited errors over large swaths of the city, the gravity model produces clusters of large underestimations (such as in Manhattan and Staten Island) but no extreme overestimation (Fig. 6a), and the ANN model exhibits both extreme overestimations and underestimations in rather isolated pockets spread across the city, particularly in Manhattan (Fig. 6b). The RF model has the most accurate estimation across the study area (Fig. 6c).

One important note is that model performances are different depending on the size of the dataset. Comparing the models developed without Twitter data for 7445 and 878,132 OD dyads, all models showed an improvement in performance with a sharp decrease in MSE when the size of training data increased. The results for the ANN model are in contrast with past studies (i.e., Tillema et al., 2006), where ANNs were identified to be more reliable than the gravity models when data is scarce. These results suggest a general improvement in all models with additional data.

The RF results are noteworthy from three perspectives: the models' predictive power, MSE, and the importance of predictor variables in explaining the output variable of home-work flows. Past studies of travel demand have reported traffic flows on the roads with an R^2 ranging from 0.50 to 0.75 (Apronti & Ksaibati, 2018). Although our study is about trip distribution between census tracts, the resulting R^2 of 0.77 can be regarded as a very good fit, especially at fine spatial granularity. The observed MSE is also lower compared to similar studies (i.e., Mozolin et al., 2000).

To identify the relative importance of variables in the magnitude of OD flows, the model measures how much worse the MSE becomes after permutation of out-of-bag observations across each input variable. The larger the change, the more important the variable. Population of origin census tract was identified as the most important variable contributing to the model performance in both cases whether Twitter population is included or excluded (Fig. 7). Destination employment, destination POI, destination sprawl index, and distance were among the other important variables when developing the RF model in the absence of Twitter data (Fig. 7b). With the addition of Twitter population, the model identified the Twitter population of destination as one of the important variables (Fig. 7a). We also developed an RF model using

Fig. 6. Commuting flows modeling error. These maps show the difference between the ground-truth and predicted numbers of incoming flows per census tract for the gravity (a), ANN (b), and RF (c) models. Blue marks underestimation by the models, red overestimation by the models, and pink the most accurate predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

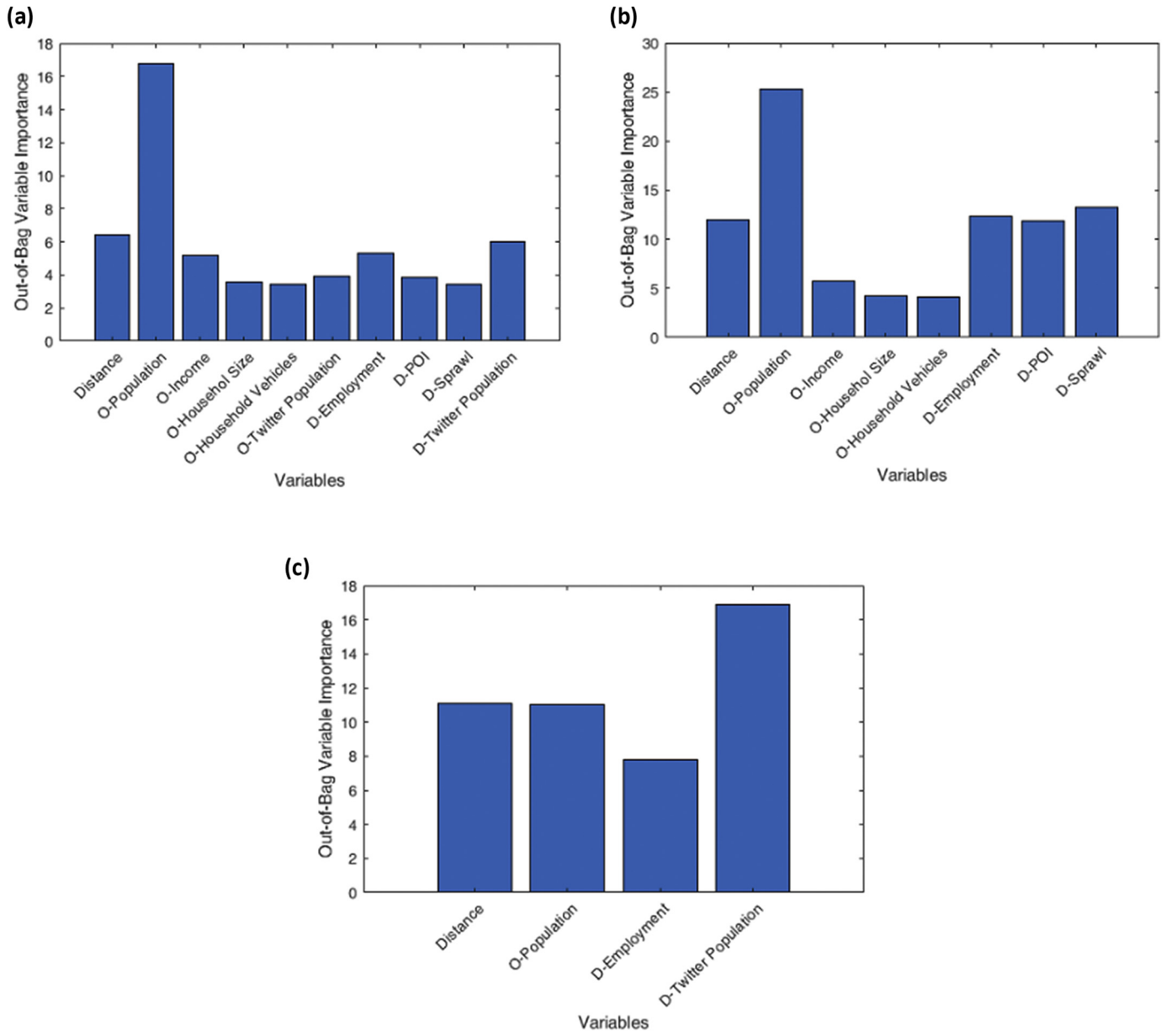

Fig. 7. Relative importance of variables for the RF models including Twitter population (a), excluding Twitter population (b), and with the four important variables (c).

only the four most important variables (Fig. 7c). The findings show that, with a limited number of input variables, the destination Twitter population becomes the most important factor contributing to the model.

5. Discussion

We compared the performance of the gravity, ANN, and RF models in commuting trip distribution at the census tract level in NYC. RF is identified as the best model, with the highest R^2 and the lowest MSE. Travel demand analysis studies have commonly used statistical methods to model different travel components (i.e., travel mode, departure time, and trip destination) (Golshani et al., 2018). With the limitation of these methods in capturing the nonlinearities in the data, machine learning methods such as ANNs have gained popularity in transportation research. In trip distribution modeling, studies have indicated that ANN could provide similar goodness of fit as a gravity model, yet the error is usually higher for the ANN models (Mozolin et al., 2000; Tillema et al., 2006; Yaldi, Taylor, & Yue, 2009, 2011). Our results show better performance of ANN models with higher R^2 and lower MSE compared to the gravity models. While Tillema et al. (2006) found better performance of ANNs (in terms of lower error) in case of data scarcity, our findings indicate a lower MSE for ANN models in both datasets (7445 and 878,132 ODs).

Despite the lower performance of ANNs in trip distribution modeling, there have been other studies pointing to promising results of ANNs for modeling traffic flows. In short-term traffic forecasting, ANNs (individual or hybrid) have mostly outperformed other methods (i.e., Stathopoulos, Dimitriou, & Tsekeris, 2008; Vlahogianni, Karlaftis, & Golias, 2007). The primary focus in these studies was the time-series prediction to test the accuracy against the traditional models. One can conclude that ANNs might show better performance when dealing with temporal aspects of the traffic data. Therefore, ANNs could be more appropriate in developing dynamic models. Developing time-series trip distribution models might be a good practice for future research.

Trip distribution modeling is a complicated and important step in

transport planning. The errors generated during this step will pass on to the other steps (Tillema et al., 2006). Developing more accurate models, therefore, is critical for planning purposes such as alleviating traffic congestion problems. While RF model had the best performance in this study, the ANN model also showed a lower MSE compared to similar studies (i.e., Mozolin et al., 2000). Past studies have mostly been conducted at coarse spatial granularity such as county that dealt with much fewer zones compared to our study. A study similar to ours was conducted by Mozolin et al. (2000) who modeled trip distribution at the census tract level. The MSE observed for the ANN model in our research was lower compared to this earlier study. The better performance of our ANN model might confirm that the appropriate selection of the variables is important in trip distribution modeling. Despite the better performance of ANN model compared to previous studies, its predictive accuracy is poor on the whole scale. The RF model not only resulted in a higher goodness of fit, but also a lower MSE. While RFs have been identified among the best ensemble methods for modeling trip generation and modal split (Ghasri et al., 2017; Rasouli & Timmermans, 2014), our results confirm their suitability in trip distribution modeling as well. In addition, RF allowed us to identify the importance of variables in the model.

Contrary to past research, Twitter flow was not an important variable in the RF model. It is our conjecture this may be due to the fine spatial granularity at which we modeled trip distribution. Past studies (Beiró et al., 2016; McNeill et al., 2017) at the county or larger scales have identified Twitter flows as a good proxy for mobility flows. A more detailed analysis for Twitter flows should be conducted probably in specific areas, where there are more flows. The RF model showed a slight improvement with the addition of Twitter population. Twitter population at destination (work census tract) was the third most influential variable after origin (home census tract) population and distance. Its importance even increased when developing the model with fewer variables. Hence, Twitter population can be a good predictor of trip distribution when data on more conventional predictors are scarce, incomplete, outdated or uncertain. This would be useful in cases where the major aim is a higher accuracy and better prediction for modeling trip distribution. When Twitter population was dropped from the model specification, sprawl index and POI were identified among the significant variables. This is particularly meaningful when the objective is to understand the mechanism behind the modeling process and identifying major factors. From a policy analysis perspective, it is essential to identify how specific factors such as land use can affect trip distribution within cities.

Twitter population may in fact be a proxy measure of attraction of a census tract, where there are higher number of points of interest and therefore, higher activities. Higher activities are usually happening at denser areas, where mixed land uses attract more population. Therefore, Twitter as a proxy for the attractiveness of these areas can be translated to a trip attraction factor in modeling trip distribution. Twitter has the potential to be a good source of data to develop population dynamics that can be used for planning purposes. These data could be instrumental in better apprehending critical attraction locations in cities and arranging sustainable transportation alternatives (Yang, Herrera, et al., 2015). Additionally, Population dynamics can be applied to evacuation planning models to forecast movements in real time. With the increasingly adverse effects of climate change such as sever hurricanes, such data are becoming more valuable for better urban management (Eshghi & Schmidtke, 2018). There would also be more potential to conduct dynamic travel demand modeling and monitoring in the near future with the increase in the number of people using social media. The use of these data coupled with artificial intelligence may lead to smart and sustainable solutions for travel demand planning and management.

6. Conclusions

A primary aim of transportation policy makers is to achieve sustainable mobility in urban areas (Kepaptsoglou et al., 2012; May, 2013). However, collecting travel demand data at high spatio-temporal resolution is the major gap that exists between the current state of the practice and an efficient urban sustainable solution (Yang, Herrera, et al., 2015). Social media may be a useful source of data to achieve this goal, yet their potential remains insufficiently investigated. The primary focus in our study was to explore the integration of social media data with traditional data in estimating trip distribution within cities and in comparing different models to identify the model with the highest performance. Given the identified need in literature for utilizing more factors in gravity modeling combined with new techniques, three models including gravity, neural network, and random forest were developed for predicting the home-work flows between census tracts of New York City. The results showed that the models performed better when the Twitter population enters the model specification as input. The random forests model showed the best performance. Twitter flow was not identified as an important factor in the models that would warrant further investigation. However, Twitter population, particularly at the work census tract, was a significant factor. The findings suggest the potential for using social media for developing dynamic models of population and trip distribution in future research to achieve more sustainable solutions for cities. However, this research has several limitations. Twitter population is not totally representative of all demographic and socioeconomic groups within a population. In addition, not all Twitter users share their location, thus resulting in a population bias (Jiang et al., 2018) that may result in overestimations or underestimations in the model results. Therefore, these biases need to be considered when using volunteered geographic information (Jiang & Thill, 2015). Despite all these limitations, social media data are still valuable especially in areas with high population density (Li, Goodchild, & Xu, 2013). Developing more advanced methods to draw a representative sample from social media (Jiang, Li, & Cutter, 2019) remains an important issue for future research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038.

Amita, J., Singh, J. S., & Kumar, G. P. (2015). Prediction of bus travel time using artificial neural network. *International Journal for Traffic and Transport Engineering, 5*(4), 410–424.

Anda, C., Erath, A., & Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences, 21*(sup1), 19–42. https://doi.org/10.1080/12265934.2017.1281150.

Apronti, D. T., & Ksaibati, K. (2018). Four-step travel demand model implementation for estimating traffic volumes on rural low-volume roads in Wyoming. *Transportation Planning and Technology, 41*(5), 557–571. https://doi.org/10.1080/03081060.2018.1469288.

ArcGIS (2018). North America detailed streets. Retrieved from https://www.arcgis.com/home/item.html?id=f38b87cc295541fb88513d1ed7cec9fd.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., & Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, 734*, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (2015). *Neural network toolbox user's guide. MathWorks.* https://doi.org/10.1002/0471221546.

Beiró, M. G., Panisson, A., Tizzoni, M., & Cattuto, C. (2016). Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science, 5*(1), 30. https://doi.org/10.1140/epjds/s13688-016-0092-2.

Berger, A. D. (2012). *A travel demand model for rural areas, (August)*.

Black, W. R. (1995). Spatial interaction modeling using artificial neural networks. *Journal of Transport Geography, 3*(3), 159–166. https://doi.org/10.1016/0966-6923(95)00013-S.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Celik, H. M. (2004). Modeling freight distribution using artificial neural networks. *Journal of Transport Geography, 12*(2), 141–148. https://doi.org/10.1016/j.jtrangeo.2003.12.003.

Ding, C., Wang, W., Wang, X., & Baumann, M. (2013). A neural network model for driver's lane-changing trajectory prediction in urban traffic flow. *Mathematical Problems in Engineering,* (2013), 1–8. https://doi.org/10.1155/2013/967358.

de Dios Ortuzar, J., & Willumsen, L. G. (2011). *Modelling transport.* John Wiley & Sons.

Eshghi, M., & Schmidtke, H. R. (2018). An approach for safer navigation under severe hurricane damage. *Journal of Reliable Intelligent Environments, 4*(3), 161–185. https://doi.org/10.1007/s40860-018-0066-1.

Ewing, R., & Hamidi, S. (2014). *Measuring urban sprawl and validating sprawl measures.* Retrieved from file:///C:/Users/n_poureb/Downloads/sprawl-report-short.pdf.

Ghasri, M., Hossein Rashidi, T., & Waller, S. T. (2017). Developing a disaggregate travel demand system of models using data mining techniques. *Transportation Research Part A: Policy and Practice, 105*(June 2016), 138–153. https://doi.org/10.1016/j.tra.2017.08.020.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., & Mohammadian, A. (2018). Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. *Travel Behaviour and Society, 10*, 21–32. https://doi.org/10.1016/j.tbs.2017.09.003.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782. https://doi.org/10.1038/nature06958.

Hamner, B. (2010). Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 1357–1359). . https://doi.org/10.1109/ICDMW.2010.128.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13 (p. 1). New York*New York, USA: ACM Press. https://doi.org/10.1145/2505821.2505823.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072.

Internet Live Stats (2019). Retrieved April 29, 2019, from www.internetlivestats.com.

Jiang, B., & Thill, J.-C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems, 53*, 1–3. https://doi.org/10.1016/j.compenvurbsys.2015.09.011.

Jiang, Y., Li, Z., & Cutter, S. L. (2019). Social network, activity space, sentiment and evacuation : What can social media tell us? *Annals of the Association of American Geographers* (February).

Jiang, Y., Li, Z., & Ye, X. (2018). Understanding demographic and socioeconomic biases of geotagged twitter users at the county level. *Cartography and Geographic Information Science, 46*(3), 1–15. https://doi.org/10.1080/15230406.2018.1434834.

Josephs, L. (2017, April). New York City needs foreign visitors because they spend four times more money than Americans. Quartz. Retrieved from https://qz.com/954413/new-york-city-needs-foreign-tourists-because-they-spend-more/.

Karduni, A., Cho, I., Wessel, G., Ribarsky, W., Sauda, E., & Dou, W. (2017). Urban space explorer: A visual analytics system for urban planning. *IEEE Computer Graphics and Applications, 37*(5), 50–60. https://doi.org/10.1109/MCG.2017.3621223.

Karimi, F., Sultana, S., Shirzadi Babakan, A., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems, 75*, 61–75. https://doi.org/10.1016/j.compenvurbsys.2019.01.001 (August 2018).

Kepaptsoglou, K., Meerschaert, V., Neergaard, K., Papadimitriou, S., Rye, T., Schremser, R., & Vleugels, I. (2012). Quality Management in Mobility Management: A scheme for supporting sustainable transportation in cities. *International Journal of Sustainable Transportation, 6*(4), 238–256. https://doi.org/10.1080/15568318.2011.587137.

Kim, J., Park, J., & Lee, W. (2018). Why do people move? Enhancing human mobility prediction using local functions based on public records and SNS data. *PLoS One, 13*, e0192698. https://doi.org/10.1371/journal.pone.0192698.

Kurkcu, A., Ozbay, K., & Morgul, E. F. (2016). Evaluating the usability of geo-located twitter as a tool for human activity and mobility patterns: A case study for New York City. *Transportation Research Board's 95th Annual Meeting.* Washington, D.C, Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781107415324A009.

Lagomarsino, D., Tofani, V., Segoni, S., Catani, F., & Casagli, N. (2017). A tool for classification and regression using random forest methodology: Applications to landslide susceptibility mapping and soil thickness modeling. *Environmental Modeling and Assessment, 22*(3), 201–214. https://doi.org/10.1007/s10666-016-9538-y.

Lenormand, M., Bassolas, A., & Ramasco, J. J. (2016). Systematic comparison of trip distribution laws and models. *Journal of Transport Geography, 51*, 158–169. https://doi.org/10.1016/j.jtrangeo.2015.12.008.

Leshem, G., & Ritov, Y. (2007). Traffic flow prediction using adaboost algorithm with random forests as a weak learner. *International Journal of Mathematical and Computational Sciences, 1*(1), 193–198. https://doi.org/10.1007/s11117-011-0122-z.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science, 40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139.

Liu, J., Zhao, K., Khan, S., Cameron, M., & Jurdak, R. (2015). Multi-scale population and mobility estimation with geo-tagged tweets. *2015 31st IEEE International Conference on Data Engineering Workshops (Vol. 2015–June)* (pp. 83–86). IEEE. https://doi.org/10.1109/ICDEW.2015.7129551.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773.

May, A. D. (2013). Urban transport and sustainability: The key challenges. *International Journal of Sustainable Transportation, 7*(3), 170–185. https://doi.org/10.1080/15568318.2013.710136.

McNally, M. G. (2007). The four step model. In D. A. Hensher, & K. J. Button (Eds.). *Handbook of transport modeling* (pp. 35–52). (2nd ed.). doi:https://escholarship.org/uc/item/0r75311t.

McNeill, G., Bright, J., & Hale, S. A. (2017). Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science, 6*(1), 24. https://doi.org/10.1140/epjds/s13688-017-0120-x.

Mozolin, M., Thill, J.-C., & Lynn Usery, E. (2000). Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological, 34*(1), 53–73. https://doi.org/10.1016/S0191-2615(99)00014-4.

National Cancer Institute (2014). *Updated urban sprawl data for the United States.* Retrieved from https://gis.cancer.gov/tools/urban-sprawl/.

Nielsen, R. H. (1987). Kolmogorov's mapping neural network existence theorem. *Proceedings of the IEEE first international conference on neural networks* (pp. 11–13). San Diego: Piscataway, NJ: IEEE.

Nimmo, B. (2017, Aug). #BotSpot: Twelve ways to spot a bot. Digital Forensic Research Lab (@DFRLab). Retrieved from https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c.

Nimmo, B. (2019). Measuring traffic manipulation on Twitter. Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/01/Manipulating-Twitter-Traffic.pdf.

NYC OpenData (2017). Points of interest. Retrieved from https://opendata.cityofnewyork.us/.

de Oña, J., & Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: A new approach to handle instability. *Neural Computing and Applications, 25*(3–4), 859–869. https://doi.org/10.1007/s00521-014-1573-5.

Ostmann, A., & Martínez Arbizu, P. (2018). Predictive models using randomForest regression for distribution patterns of meiofauna in Icelandic waters. *Marine Biodiversity, 48*(2), 719–735. https://doi.org/10.1007/s12526-018-0882-9.

Pitombo, C. S., de Souza, A. D., & Lindner, A. (2017). Comparing decision tree algorithms to estimate intercity trip distribution. *Transportation Research Part C: Emerging Technologies, 77*, 16–32. https://doi.org/10.1016/j.trc.2017.01.009.

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on twitter during natural disasters: A case study of hurricane sandy. *International Journal of Disaster Risk Reduction, 37*, 101–176 10.1016/j.ijdrr.2019.101176.

Pourebrahim, N., Sultana, S., Thill, J.-C., & Mohanty, S. (2018). Enhancing trip distribution prediction with twitter data: Comparison of neural network and gravity models. *Proceedings of the 2nd ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery, GeoAI 2018*Seattle, WA: ACM. https://doi.org/10.1145/3281548.3281555.

Rashidi, T. H., & Mohammadian, A. (2011). Household travel attributes transferability analysis: Application of a hierarchical rule based approach. *Transportation, 38*(4), 697–714. https://doi.org/10.1007/s11116-011-9339-8.

Rasouli, S., & Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *EJTIR Issue, 14*(4), 412–424.

Roy, J. R., & Thill, J. C. (2003). Spatial interaction modelling. *Papers in Regional Science, 83*(1), 339–361. https://doi.org/10.1007/s10110-003-0189-4.

Rumelhart, D. E., & Mcclelland, J. L. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. *Foundations*. *Vol. 1*Cambridge, Massachusetts: The MIT Press. Retrieved from https://www.researchgate.net/publication/200033859.

Sekhar, C. R., Minal, & Madhu, E. (2016). Mode choice analysis using random forrest decision trees. *Transportation Research Procedia, 17*, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119 (December 2014).

Shirzadi Babakan, A., Alimohammadi, A., & Taleai, M. (2015). An agent-based evaluation of impacts of transport developments on the modal shift in Tehran, Iran. *Journal of Development Effectiveness, 7*(2), 230–251. https://doi.org/10.1080/19439342.2014.994656.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 96–100. https://doi.org/10.1038/nature10856.

SimplyAnalytics (2015). Retrieved from http://simplyanalytics.com/.

Srisaeng, P., & Baxter, G. (2017). Modelling Australia's outbound passenger air travel demand using an artificial neural network approach. *International Journal for Traffic and Transport Engineering, 7*(4), 406–423. https://doi.org/10.7708/ijtte.2017.7(4).01.

Stathopoulos, A., Dimitriou, L., & Tsekeris, T. (2008). Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-Aided Civil and Infrastructure Engineering, 23*(7), 521–535. https://doi.org/10.1111/j.1467-8667.2008.00558.x.

Sulaiman, S., Shamsuddin, S. M., Abraham, A., & Sulaiman, S. (2011). Intelligent web caching using machine learning methods. *Neural Network World, 21*(5), 429–452. https://doi.org/10.14311/NNW.2011.21.025.

Sultana, S., Pourebrahim, N., & Kim, H. (2018). Household energy expenditures in North Carolina: A geographically weighted regression approach. *Sustainability, 10*(5), 1511. https://doi.org/10.3390/su10051511.

Sultana, S., & Weber, J. (2007). Journey-to-work patterns in the age of sprawl: Evidence from two midsize southern metropolitan areas*. *The Professional Geographer, 59*(2), 193–208. https://doi.org/10.1111/j.1467-9272.2007.00607.x.

Sultana, S., & Weber, J. (2014). The nature of urban growth and the commuting transition: Endless sprawl or a growth wave? *Urban Studies, 51*(3), 544–576. https://doi.org/10.1177/0042098013498284.

Thill, J.-C., & Wheeler, A. (2000a). Tree induction of spatial choice behavior. *Transportation Research Record, 1719*, 250–258.

Thill, J.-C., & Wheeler, A. (2000b). Knowledge discovery and induction of decision trees in spatial decision problems. In A. Reggiani (Ed.). *Spatial economic science: New frontiers in theory and methodology* (pp. 188–211). Heidelberg: Springer.

Tillema, F., van Zuilekom, K. M., & van Maarseveen, M. F. A. M. (2006). Comparison of neural networks and gravity models in trip distribution. *Computer-Aided Civil and Infrastructure Engineering, 21*(2), 104–119. https://doi.org/10.1111/j.1467-8667.2005.00421.x.

Tsou, M.-H., Zhang, H., & Jung, C.-T. (2017). *Identifying data noises, user biases, and system errors in geo-tagged twitter messages (tweets)*. arXiv:1712.02433 [cs.SI].

U.S. Census Bureau (2015). Longitudinal employer-household dynamics. Retrieved from https://lehd.ces.census.gov/data/.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2007). Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks. *Computer-Aided Civil and Infrastructure Engineering, 22*(5), 317–325. https://doi.org/10.1111/j.1467-8667.2007.00488.x.

Wilson, A. (1970). *Entropy in urban and regional modelling.* London: Routledge.

Wilson, A. G. (1998). Land-use / transport interaction models past and future. *Journal of Transport Economics and Policy, 32*(1), 3–26.

Yaldi, G., Taylor, M. A. P., & Yue, W. L. (2009). Improving artificial neural network performance in calibrating doubly-constrained work trip distribution by using a simple data normalization and linear activation function. *32nd Australasian transport research forum, ATRF 2009*.

Yaldi, G., Taylor, M. A. P., & Yue, W. L. (2011). Forecasting origin-destination matrices by using neural network approach: A comparison of testing performance between back propagation, variable learning rate and levenberg-marquardt algorithms. *Australasian transport research forum 2011* (pp. 1–15). .

Yang, F., Jin, P. J., Cheng, Y., Zhang, J., & Ran, B. (2015). Origin-destination estimation for non-commuting trips using location-based social networking data. *International Journal of Sustainable Transportation, 9*(8), 551–564. https://doi.org/10.1080/15568318.2013.826312.

Yang, Y. (2013). *Understanding human mobility patterns from digital traces.* Massachusetts Institute of Technology.

Yang, Y., Herrera, C., Eagle, N., & González, M. C. (2015). Limits of predictability in commuting flows in the absence of data for calibration. *Scientific Reports, 4*(1), 5662. https://doi.org/10.1038/srep05662.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies, 58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019.

Zipf, G. K. (1946). The P 1 P 2 /D hypothesis: On the intercity movement of persons. *Source: American Sociological Review, 11*.