# Enhancing Trip Distribution Using Twitter Data: Comparison of Gravity and Neural Networks

By: Nastaran Pourebrahim, Selima Sultana, Jean-Claude Thill, and Somya Mohanty

Pouebrahim, N., Sultana, S., Thill, J-C., Mohanty, S. Enhancing Trip Distribution Using Twitter Data: Comparison of Gravity and Neural Networks. In 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI'18), November 6, 2018, Seattle, WA, USA. ACM. New York, NY, USA, 4 pages. https://doi.org/10.1145/3281548.3281555.



## Abstract:

Predicting human mobility within cities is an important task in urban and transportation planning. With the vast amount of digital traces available through social media platforms, we investigate the potential application of such data in predicting commuter trip distribution at small spatial scale. We develop back propagation (BP) neural network and gravity models using both traditional and Twitter data in New York City to explore their performance and compare the results. Our results suggest the potential of using social media data in transportation modeling to improve the prediction accuracy. Adding Twitter data to both models improved the performance with a slight decrease in root mean square error (RMSE) and an increase in R-squared ($R^2$) value. The findings indicate that the traditional gravity models outperform neural networks in terms of having lower RMSE. However, the $R^2$ results show higher values for neural networks suggesting a better fit between the real and predicted outputs. Given the complex nature of transportation networks and different reasons for limited performance of neural networks with the data, we conclude that more research is needed to explore the performance of such models with additional inputs.

**Keywords:** Machine learning | neural networks | mobility | social media | transport modeling | Twitter

## Article:

## 1 INTRODUCTION

Modeling human mobility at different spatial scales and aggregation levels is a long tradition in geography, spatial science, and other disciplines. Various applications of human mobility include urban and transportation planning and management, resource allocation, prediction of migration flows, and epidemic spreading [3].Transportation demand modeling has been dominated by the

four-step modeling framework (FSM): trip generation, trip distribution, mode choice, and route choice [19]. Trip distributions between origins and destinations (OD) have long been predicted through spatial interactions or the gravity models [4, 23], and more recently through the radiation model [3]. The gravity model has been reported to be more reliable than the radiation model at small spatial scales such as commuting flows within cities [16]. However, mobility patterns within cities might not be fully predicted by the traditional gravity model where population and distance are used as key factors [14].

The emergence of information technologies in recent years has introduced a new source of data known as big data to scholars in the field of transportation [27]. Several types of geospatial big data such as taxi trajectories, mobile phone records, and social media messages have been used to study cities in ways that were not possible before [18]. Studies have utilized geolocated data to model and visualize movements in urban areas [i.e. 12]. With the growing rate of social media users, a much larger sample size is now available, which presents a challenge to effectively use these data in modeling human behavior. Therefore, the potential use of hybrid models combining the big and traditional data to improve predictions has been identified in the literature [3]. Most recently, developed models integrating social media data with the gravity model have shown promising results in predicting commuting and non-commuting flows [3, 27]. More research is needed to understand the potential of social media in predicting mobility patterns, particularly at small spatial scales. In addition, the need for using new techniques for improving prediction accuracy has been identified in the literature [3].

Over the past years, Machine Learning (ML) techniques such as Artificial Neural Networks (ANNs) have been applied in various applications of transportation research. Domains of applications include traffic operations and traffic management systems [7, 13], which provide superior results compared to traditional modeling techniques [6]. While a few studies [i.e. 21, 26] have compared the gravity and ANN models in estimating trip distribution, the potential inclusion of social media data in these models has not been considered. Given the current state of research, our study explores the development of gravity and ANN based models for commuting trip distribution combining both traditional data (population, employment, distance) and social media data (number of tweets). The paper is organized as follows: the review of related work regarding social media and human mobility is discussed in section 2, followed by a presentation of the study area, data sources and methodology in section 3. Finally, the results are presented in section 4 with concluding remarks in section 5.

## 2 SOCIAL MEDIA AND HUMAN MOBILITY

Social media data has attracted considerable attention from social and data scientists in academia [20]. A number of studies have used social media data, particularly Twitter posts, to understand population movements at different spatial scales. For example, Hawelka et al. [9] used Twitter dataset to examine mobility profiles of international travelers through mobility rate, radius of gyration, diversity of destinations and a balance of the inflows and outflows. Similarly, other studies identified the feasibility of Twitter data as a proxy for human mobility at the country and county levels [i.e. 15, 17].

While mobility patterns identified from geolocated tweets have been commonly reported through measures such as radius of gyration and user displacement [i.e. 15], a few studies have looked at these patterns from a modeling perspective. More recently, McNeil et al. [20] reported that estimated commuting flows derived from Twitter data outperform the radiation method, particularly for short trips with higher volume of commuters. Kim et al. [14] compared different variables to generate mass values for an urban traffic gravity model and identified the number of tweets as the most powerful predictor. Yang et al. [27] combined clustering, regression, and gravity modelling to estimate an origin-destination (OD) matrix for non-commuting trips using Foursquare user check-in data. They found similarity between their estimated OD matrix and the ground-truth OD matrix in the Chicago urban area. Beiro et al. [3] integrated georeferenced pictures from Flickr with a standard gravity model under a stacked regression procedure to predict air travel and daily commuting in the United States counties. Their results showed that the hybrid gravity model outperforms the traditional gravity model.

With the promising results achieved in these studies, it is evident that these modeling efforts enhance traditional approaches. However, more research is needed to fully grasp the potential of social media in predicting mobility patterns at small spatial scales using new techniques and integrating traditional and new datasets. The application of artificial neural networks (ANNs) in travel demand modelling was introduced in 1960. However, despite their success at learning and recognizing patterns, they were not used for about three decades due to their slow response to the inputs' modifications [24]. ANNs have shown great advantages in prediction, pattern identification, optimization, and signal processing due to their non-linear and flexible structure that can solve complex problems [6]. While ANNs have been used in different transportation studies [i.e. 21], their applications in understanding and modeling mobility patterns within cities need further investigation.


**Figure 1:** Study area: New York City census tracts

## 3 METHODOLOGY

### 3.1 Study Area and Datasets

This study uses New York City (NYC) to identify commuting patterns. NYC is selected due to a large number of trips generated within its boundary and the ready availability of data. We use NYC census tracts as the geographical units of modeling (Figure 1). We collected 2015 LEHD Origin-Destination Employment Statistics (LODES) for New York City from the U.S. Census Bureau as ground-truth data. The dataset reports the workers home and employment locations with other characteristics such as age, earnings, industry distributions, and local workforce

indicators. We extracted a total of 903,686 home-work flows between census tracts. We supplemented the origin-destination data with two per-census tract features of population and employment taken from Simply Analytics, a web-deriven database [25], for the year 2015. For our Twitter dataset, we collected geolocated tweets posted in New York City from June 2015 to May 2016 from the SOPHI data lake maintained by the Data Science Initiative (DSI) Center at the University of North Carolina at Charlotte. We only used the tweets with accurate location information, resulting in 2,254,289 usable tweets. We then calculated the number of tweets in each census tract.

### 3.2 Gravity Model

The gravity model, which is derived from Newton's law of gravity, is commonly used in modeling various flows such as migration and trades between geographic areas [8]. This model is the basis for estimating trip distribution in the four-step model of transportation [3]. The model is formulated as follows (Equation 1) [28]:

$$Tij = G\frac{Mi^{\alpha}Mj^{\beta}}{Dij^{\gamma}} \quad (1)$$

Tij is the volume of mobility between two areas i (origin) and j (destination). Mi and Mj are trip production and attraction values represented by the population of areas i and j. Dij represents the geographic distance between the two areas i and j. Different socio-economic factors can be added to the equation as trip production and attraction factors [14]. We use residential population at home census tract as the trip production and employment at work census tract as the trip attraction in the first gravity model. In the second model, the number of tweets in both home and work census tracts are added to the model. The Euclidean distance between the centroids of census tracts is used to estimate the inter-zonal homework flows. We recognize that using the Euclidean distance instead of actual road distance might affect the results. However, the relative accuracy of the estimates is reasonable, as our major goal is to understand the potential of Twitter data and to compare the gravity and neural network models [21]. We only include non-zero flows for the analysis. The equation can be represented as a linear model by log-transforming both sides to obtain α, β, and γ [8]. We use root mean square error (RMSE) and coefficient of determination ($R^2$) to quantify the estimation and compare the gravity models' performance with the ANN models [11].

### 3.3 Artificial Neural Networks (ANNs)

Artificial neural networks (ANNs) are computational models that learn from and recognize patterns in data, inspired by the nervous system in the brain [10]. The basic model of neural network has three layers: input, hidden, and output. Each layer is composed of processing elements named neurons or nodes that are interconnected by weighted links [7]. The number of neurons in the input layer is the number of the variables that are used to predict the output, and the neurons in the output layer correspond to the predicted variables [1]. The complexity of connection between the input and output layers determines the number of hidden layers and their neurons. The number of hidden neurons is usually decided through trial and error approach [1]. The output of a neuron with n inputs is calculated as follows (Equation 2) [2]:

$$Y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \quad (2)$$

Where Y is output, $x_i$ is input value, $w_i$ is the weight, b is the bias and f is a transfer function of the neuron. A neural network learns the correlation between input and output by adopting the links' weights [1]. The learning process can be done through different algorithms of which back-propagation (BP) is the most widely used. We develop two BP neural networks in MATLAB 2018 to predict the home-work flows with the same dataset used in the gravity models. For the first model, population, employment and distance are the input variables in the model representing three neurons in the input layer. The network has one hidden layer with 3 neurons. We developed one hidden layer because past literature suggests that most problems in the world can be solved with a single hidden layer [5]. In the second model, the number of tweets in both origin and destination are added as neurons to the input layer. The input and hidden layers have 5 neurons in this model. The output layers in both models have one neuron showing our target value: home-work flows. We use log-sigmoid and linear as transfer functions for hidden and output layers, respectively, and the Levenberg-Marquardt algorithm for training. We train the BP neural networks so that the MSE is minimized. We randomly divided the data to 70% for training, 15% for testing and 15% for validation. We present the performance of models based on root mean square error (RMSE) and coefficient of determination (R*2) values.

## 4 RESULTS AND DISCUSSION

The first gravity model used population, employment, and distance as the inputs. The RMSE and $R^2$ for this model show 0.78 and 0.06, respectively. Adding Twitter data to the model improved the performance by decreasing the RMSE to 0.68 and increasing the $R^2$ to 0.31 (Table 1). For the neural network model, the test set shows the lowest RMSE of 3.97 and an $R^2$ of 0.15. The RMSEs observed for both gravity and neural network are almost similar to the results reported in the study by Mozolin et al. [21]. They developed gravity and neural network models for commuting trip distribution between census tracts of the Atlanta Metropolitan Statistical Area. However, Tilema et al. [26] reported lower RMSE (about 2) for their developed neural networks to model trip distribution between 15 regions in Rotterdam, Netherlands. We then developed a neural network adding the number of tweets for both origin and destinations as inputs. The BP neural networks with five neurons in the hidden layer (Figure 2) and the network performance (Figure 3) are presented. The RMSE for the test set is 3.21 and the regression results show the increase of 30 percent with an $R^2$ of 0.45.

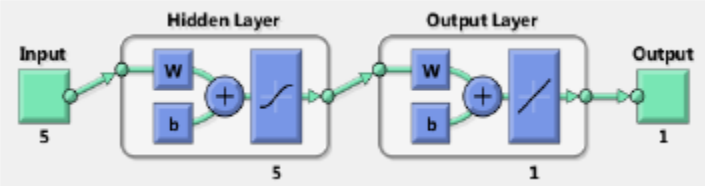


**Figure 2:** BP neural network with one hidden layer

While we did not find a desired level of RMSE or $R^2$ for trip distribution in the literature, we confirm that adding Twitter data to both models improved the performance with a slight decrease in RMSE and an increase in $R^2$ (Table 1). These results suggest the potential of using social media data in transportation modeling to improve the prediction accuracy. This is especially important for the places where there are data availability limitations [3]. The trips generated in the trip distribution models are static and do not vary based on anything other than changes to the socioeconomic factors. Adding Twitter data can be a useful step to develop dynamic models in the future. Our preliminary results show that the gravity models outperform the neural network models in terms of RMSE, a result also observed in past studies comparing the gravity and neural network models [21]. However, the $R^2$ for our ANN models shows a better fit, suggesting that a higher percentage of the home-work flows can be predicted by the input variables in the neural network models.

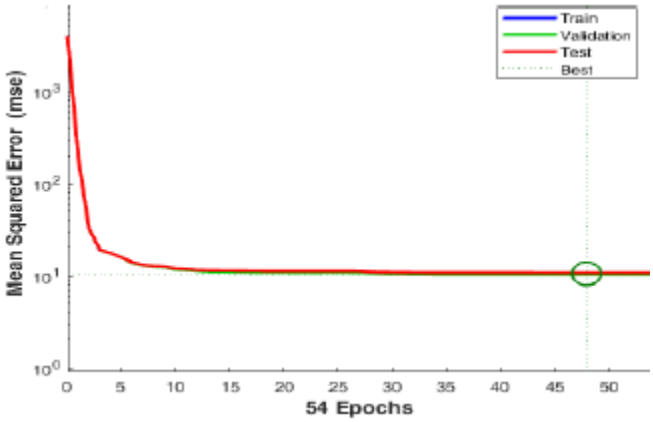


**Figure 3:** Neural network performance

**Table 1:** Comparison of gravity and neural network models

| Model | RMSE | $R^2$ |
|---|---|---|
| Gravity with 3 Variables | 0.78 | 0.06 |
| Gravity with 5 Variables | 0.68 | 0.31 |
| Neural Network with 3 Inputs | 3.97 | 0.15 |
| Neural Network with 5 Inputs | 3.21 | 0.45 |

Trip distribution is an important step in transportation modeling. Errors generated in this step are passed to the other steps that affect the model's accuracy and cause problems for transport planning [26]. Neural networks have outperformed the gravity model in the past studies estimating annual average daily traffic and in cases where data were scarce [7, 26]. In addition, neural networks' performance is dependent on the number of inputs, number of hidden layers, activation function, and learning method [22, 26]. This indicates the need to evaluate the performance of more complex networks in modeling trip distribution. A limitation of our research is the exclusion of zero home-work flows in the analysis because we were mostly interested in identifying the impact of adding social media data in both gravity and neural network models. More research is needed to identify the situations under which gravity and neural network models work well [26].

## 5 CONCLUSIONS

Understanding human mobility is an enduring topic of research in spatial sciences due to its essential role for modeling and predicting travel demand, disaster management, disease spread, and ultimately the structuring of the space economy. The primary focus in this study was to explore the potential integration of social media data with traditional data in estimating human mobility within cities rather than in developing a model with higher predictive power. Given the identified need in literature for utilizing more factors in the gravity model combined with new techniques, this research also compared the performance of the neural network and gravity models in predicting the home-work flows between census tracts of New York City. The results showed that both models performed better by adding the Twitter data as inputs. The findings also indicated a better performance of the gravity model in terms of RMSE in modeling trip distribution; however, higher $R^2$ values were observed for the neural networks. The complex nature of transportation data and neural networks make it difficult to obtain good estimation. While neural networks have the flexibility to handle a larger number of inputs, this study only considered population, employment, distance, and number of tweets. Developing more complex networks with more hidden layers, more inputs, and different networks' parameters to improve neural networks' performance warrants further investigation. Predicting trip distribution during the different time of the day or different periods of the year also remains for future research.

## REFERENCES


[1] Johar Amita, Jain Sukhvir Singh, and Garg Pradeep Kumar. 2015. Prediction of bus travel time using artificial neural network. International Journal for Traffic and Transport Engineering 5, 4 (2015), 410–424.

[2] Mark Hudson Beale, Martin T Hagan, and Howard B Demuth. 2012. Neural network toolbox user's guide. In R2012a, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098,, www.mathworks.com. Citeseer.

[3] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. 2016. Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Science 5, 1 (2016), 30.

[4] Juan de Dios OrtÃozar and Luis G Willumsen. 2011. Modelling transport. John Wiley & Sons.

[5] Juan De Oña and Concepción Garrido. 2014. Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Computing and Applications 25, 3-4 (2014), 859–869.

[6] Chenxi Ding, Wuhong Wang, Xiao Wang, and Martin Baumann. 2013. A neural network model for driver's lane-changing trajectory prediction in urban traffic flow. Mathematical Problems in Engineering 2013 (2013).

[7] Venkata Ramana Duddu and Srinivas S Pulugurtha. 2013. Principle of demographic gravitation to estimate annual average daily traffic: Comparison of statistical and neural network models. Journal of Transportation Engineering 139, 6 (2013), 585–595.

[8] Roberto Durán-Fernández and Georgina Santos. 2014. Gravity, distance, and traffic flows in Mexico. Research in transportation economics 46 (2014), 30–35.

[9] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41, 3 (2014), 260–271.

[10] Robert Hecht-Nielsen. 1987. Kolmogorov's mapping neural network existence theorem. In Proceedings of the IEEE International Conference on Neural Networks III. IEEE Press, 11–13.

[11] Inho Hong and Woo-Sung Jung. 2016. Application of gravity model on the Korean urban bus network. Physica A: Statistical Mechanics and its Applications 462 (2016), 48–55.

[12] Alireza Karduni, Isaac Cho, Ginette Wessel, William Ribarsky, Eric Sauda, and Wenwen Dou. 2017. Urban Space Explorer: A Visual Analytics System for Urban Planning. IEEE computer graphics and applications 37, 5 (2017), 50–60.

[13] Matthew G Karlaftis and Eleni I Vlahogianni. 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies 19, 3 (2011), 387–399.

[14] Jungmin Kim, Juyong Park, and Wonjae Lee. 2018. Why do people move? Enhancing human mobility prediction using local functions based on public records and SNS data. PloS one 13, 2 (2018), e0192698.

[15] Abdullah Kurkcu, K Ozbay, and EF Morgul. 2016. Evaluating The Usability of Geo-Located Twitter As A Tool For Human Activity and Mobility Patterns: A Case Study for NYC. In Transportation Research Board's 95th Annual Meeting. 1–20.

[16] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. 2016. Systematic comparison of trip distribution laws and models. Journal of Transport Geography 51 (2016), 158–169.

[17] Jiajun Liu, Kun Zhao, Saeed Khan, Mark Cameron, and Raja Jurdak. 2015. Multiscale population and mobility estimation with geo-tagged tweets. In Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on. IEEE, 83–86.

[18] Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. 2015. Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers 105, 3 (2015), 512–530.

[19] Michael G McNally. 2000. The four step model. (2000).

[20] Graham McNeill, Jonathan Bright, and Scott A Hale. 2017. Estimating local commuting patterns from geolocated Twitter data. EPJ Data Science 6, 1 (2017), 24.

[21] Mikhail Mozolin, J-C Thill, and E Lynn Usery. 2000. Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. Transportation Research Part B: Methodological 34, 1 (2000), 53–73.

[22] Caleb Robinson and Bistra Dilkina. 2018. A Machine Learning Approach to Modeling Human Migration. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. ACM, 30.

[23] John R Roy and Jean-Claude Thill. 2004. Spatial interaction modelling. Papers in Regional Science 83, 1 (2004), 339–361.

[24] Minal Srivastava and Chalumuri Ravi Sekhar. 2018. Web Survey Data and Commuter Mode Choice Analysis Using Artificial Neural Network. (2018).

[25] Selima Sultana, Nastaran Pourebrahim, and Hyojin Kim. 2018. Household Energy Expenditures in North Carolina: A Geographically Weighted Regression Approach. Sustainability 10, 5 (2018), 1511.

[26] Frans Tillema, Kasper M Van Zuilekom, and Martin FAM Van Maarseveen. 2006. Comparison of neural networks and gravity models in trip distribution. Computer-Aided Civil and Infrastructure Engineering 21, 2 (2006), 104–119.

[27] Fan Yang, Peter J Jin, Yang Cheng, Jian Zhang, and Bin Ran. 2015. Origin-destination estimation for non-commuting trips using location-based social networking data. International Journal of Sustainable Transportation 9, 8 (2015), 551–564.

[28] George Kingsley Zipf. 1946. The P 1 P 2/D hypothesis: on the intercity movement of persons. American sociological review 11, 6 (1946), 677–686.