Need A Boost? A Comparison of Traditional Commuting Models with the XGBoost Model for Predicting Commuting Flows

April Morton
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
mortonam@ornl.gov

Jesse Piburn
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
piburnjo@ornl.gov

Nicholas Nagle
Department of Geography, University of Tennessee, Knoxville, 1000 Phillip Fulmer Way, Knoxville, TN 37916, USA
nnagle@utk.edu

Abstract

Commuting models estimate the number of commuting trips from home to work locations in a given area. Since their infancy, they have been increasingly used in a variety of fields to reduce traffic and pollution, drive infrastructure choices, and solve a variety of other problems. Traditional commuting models, such as gravity and radiation models, typically have a strict structural form and limited number of input variables, which may limit their ability to predict commuting flows as well as machine learning models that might better capture the complex dynamics of the commuting process. To determine whether machine learning models might add value to the field of commuter flow prediction, we compare and discuss the performance of two standard traditional models with the XGBoost machine learning algorithm for predicting home to work commuter flows from a well-known United States commuting dataset. We find that the XGBoost model outperforms the traditional models on three commonly used metrics, indicating that machine learning models may add value to the field of commuter flow prediction.

2012 ACM Subject Classification Applied computing → Law, social and behavioral sciences

Keywords and phrases Machine learning, commuting modeling

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.51

Category Short Paper

1 Introduction

Knowing how many people commute from various home to work locations is important for solving problems in a wide variety of domains. Commonly referred to as commuting flows, these movements form a complex socio-economic network that can be used to better understand the transport of people, goods, money, information, and diseases at different spatial scales [7]. Having a better grasp of these processes is important for policy- and other decision-makers who aim to tackle a variety of issues such as reducing traffic and pollution, planning the development of new infrastructure, and preventing the spread of disease.

In response to the need for better understanding the movement of commuters, researchers have developed a suite of commuting models used for estimating population flows, planning

transportation systems, analyzing urban traffic, and many other applications [7, 8, 10, 5]. This collection of techniques has traditionally consisted of different versions of what are commonly known as gravity and radiation models [7]. In general, these models are based on simple equations with a small number of input variables that have been chosen based on the assumption that the number of trips between two locations is related to their residential and work populations, the distance between the locations and/or the number of opportunities (e.g. other jobs) between them [7].

Though useful, both gravity and radiation models are analytical models with crafted functional forms and a small number of input variables [9]. This potentially limits their ability to capture the more complex dynamics that more flexible models, such as machine learning algorithms, may be able to. To determine whether machine learning models might add value to the field of commuter flow prediction, we compare and discuss the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting home to work commuter flows from a well-known United States (U.S.) commuting dataset. We find that the XGBoost model outperforms the traditional models on three commonly used metrics, showing promise for machine learning models in the field of commuter flow prediction.

2 Related Work

The goal of commuting modeling is to predict the matrix of commuters $T = (T_{ij})_{1 \leq i,j \leq n}$ that move from every zone i to every other zone j within a set of n distinct zones. Assuming there are a total of N commuters, the estimated matrix $\tilde{T} = \tilde{T}_{ij}$ is derived by first estimating the set of probabilities $(p_{ij})_{1 \leq i,j \leq n}$ that a randomly drawn commuter from the set of N commuters moves between all zones i and j, and then drawing at random N trips from the set of estimated probabilities $(\tilde{p}_{ij})_{1 \leq i,j \leq n}$. Oftentimes, additional constraints are added to ensure that the total number of commuters m_i leaving each zone i, the total number of commuters n_j working in each zone j, or both, is preserved.

In order to estimate the probabilities $(p_{ij})_{1 \leq i,j \leq n}$, researchers have traditionally used variants of the well-known gravity and radiation laws [9]. Gravity laws are based on the assumption that the number of trips T_{ij} between two locations i and j is related to the total number of commuters m_i leaving zone i, the total number of commuters n_j working in zone j, and decays directly as a function of the distance d_{ij} between the zones [6]. The importance of the distance in predicting the probabilities is typically controlled by parameters α, β, and/or γ.

Radiation laws, on the other hand, are based on the assumption that the number of trips T_{ij} between two locations i and j depends on the total number of commuters m_i leaving zone i, the total number of commuters n_j working in zone j, and the number of intervening opportunities s_{ij} between the two zones [7]. In the commuting literature, s_{ij} is typically defined as the total number of commuters working in all zones whose centroid falls in the circle centered at i with radius d_{ij} (not including zones i or j) [7]. In some forms of this law, a parameter β is introduced to control the effect of the number of intervening opportunities between the home and work zones. Table 1 provides equations for the traditional gravity and radiation laws chosen in this study.

The XGBoost model is a subset of a broader class of models, called machine learning models, that use a set of known input and output data to "learn" a model that can then be given new input data to estimate unknown output data [1]. In the case of commuter flow modeling, one might use a set of known input variables m_i, n_j, d_{ij}, s_{ij}, and known output

Table 1 Traditional commuting laws.

Law	Equation
Gravity with exponential law	$\tilde{p}_{ij} \propto m_i n_j e^{-\beta d_{ij}}$
Extended radiation law	$\tilde{p}_{ij} \propto \frac{[(m_i+n_j+s_{ij})^\beta - (m_i+s_{ij})^\beta](m_i^\beta+1)}{[(m_i+s_{ij})^\beta+1][(m_i+n_j+s_{ij})^\beta+1]}$

variables T_{ij}, to learn the structure of a machine learning model that can then take in new values of m_i, n_j, d_{ij} and s_{ij}, to estimate unknown values of T_{ij}. The XGBoost model is well known for winning several machine learning competitions and depends on three primary parameters commonly referred to as the maximum tree depth (r), number of estimators (e), and learning rate (k) [4].

3 Methodology

To determine whether the XGBoost model might add value to the field of commuter flow prediction, we compare and discuss the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting a subset of home to work commuter flows within the Knoxville Metropolitan Statistical Area (MSA). From this point forward, we refer to the home location as the origin location and the work location as the destination location. The following subsections discuss the specific gravity, radiation, and XGBoost models chosen, as well as the data, study area, evaluation metrics, and experimental setup, in greater detail.

3.1 Models

In this study, we compare the performances of the gravity model based on an exponential distance decay function, the radiation model based on the extended radiation law, and a standard implementation of the XGBoost model. Table 1 provides the equations for both the gravity and radiation laws underlying the gravity and radiation models selected. Additionally, for both the gravity and radiation models, we ensure that the number of workers in each destination zone j is preserved by simulating all $(\tilde{T}_{ij})_{1\leq i,j\leq n}$ from the multinomial distribution $\mathcal{M}\big(n_j, \big(\frac{\tilde{p}_{ij}}{\sum_{k=1}^{n}\tilde{p}_{kj}}\big)_{1\leq i,j\leq n}\big)$. From this point forward, whenever we use the terms gravity or radiation model, we are referring specifically to the gravity and radiation models chosen in this study. The exponential distance decay function and extended radiation model were chosen because of their decent performance in a recent study conducted by [7]. The standard XGBoost model was chosen because of its flexibility and proven track record as the winner of several machine learning competitions [4].

3.2 Data and Study Area

We use each of the three models to predict commuting flows reported in a Census dataset called the 2010 Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) [3]. The 2010 LODES dataset is a partially synthetic dataset that provides residential, workplace, and origin to destination commuter flow totals for a variety of U.S. Census-defined regions. We focus our study on estimating commuting flows between origin and destination Census block groups. Additionally, we focus our study on a subset of origin and destination block groups in the Knoxville MSA. More specifically, we consider all origin and destination block group pairs within a random sample of 120 block groups in

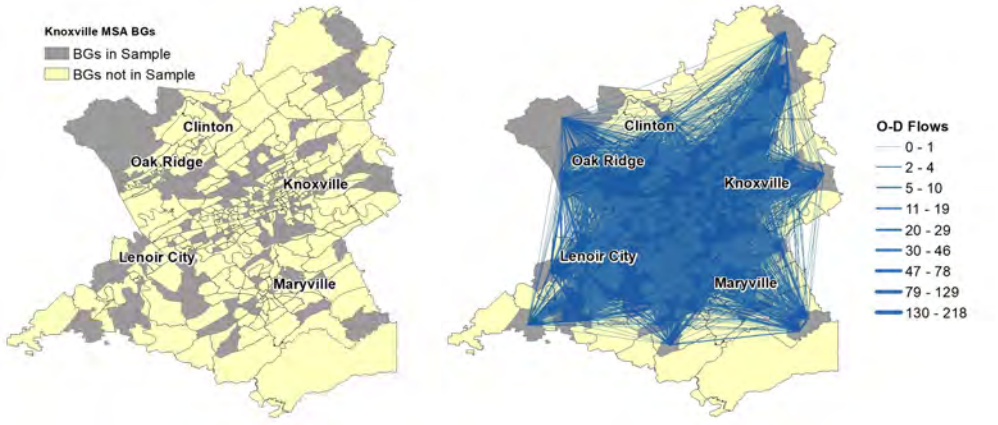

Figure 1 The spatial boundaries for all 2010 Knoxville block groups (bgs), the subset of sampled bgs in the study area, and all origin-destination (o-d) commuting flows between the sampled block groups.

Table 2 Evaluation metrics.

Metric	Equation
Common Part of Commuters (CPC)	$CPC(T,\tilde{T}) = \frac{2\sum_{i,j=1}^{n} min(T_{ij},\tilde{T}_{ij})}{\sum_{i,j=1}^{n} T_{ij}+\sum_{i,j=1}^{n} \tilde{T}_{ij}}$
Common Part of Links (CPL)	$CPL(T,\tilde{T}) = \frac{2\sum_{i,j=1}^{n} (\mathbb{1}_{T_{ij}>0}\cdot\mathbb{1}_{\tilde{T}_{ij}>0})}{\sum_{i,j=1}^{n} \mathbb{1}_{T_{ij}>0}+\sum_{i,j=1}^{n} \mathbb{1}_{\tilde{T}_{ij}>0}}$
Root Mean Squared Error ($RMSE$)	$RMSE(T,\tilde{T}) = \sqrt{\frac{1}{n}\sum_{i,j=1}^{n} (T_{ij}-\tilde{T}_{ij})^2}$

the Knoxville MSA. In total, there are $n = 14,280$ block group pairs within this subset, and $N = 15,288$ commuters who travel these routes. Figure 1 provides a visual map of the study area and data.

We use the LODES dataset to determine m_i, n_j, and T_{ij}, and another dataset, called the 2010 U.S. Census Block Group Shapefiles [2], to obtain the distances d_{ij} and intervening opportunities metrics s_{ij} for all origin block groups i and destination block groups j in the study area. Whenever we calculate a distance for a set of locations, we use the haversine formula to determine the great-circle distance between them.

3.3 Evaluation Metrics

To evaluate how well each of the models perform, we use three metrics commonly used in the commuting modeling literature. The first two, known as the Common Part of Commuters (CPC) and Common Part of Links (CPL) metrics, measure the similarity between the true commuting flow network and a predicted network. The third metric, known as the Root Mean Squared Error ($RMSE$), measures the prediction accuracy (how similar the true flow counts are to the predicted flow counts). Table 2 provides the equations for each metric.

3.4 Experimental Setup

To select the optimal hyperparameters and then compare the winning models, we split our data into training, validation, and testing sets via nested cross validation. More specifically,

we first split our data into 10 unique training and testing set pairs via 10-fold cross validation. We refer to each of these training/testing set pairs as outer folds. We then further split the training sets of each outer fold into 10 more unique training and validation sets via a second round of 10-fold cross validation.

For our gravity and radiation models, we choose the optimal $\beta \in [0, 0.1, \cdots, 1]$ for each outer fold by first simulating one possible $\tilde{T}_{ij}$ from the models corresponding to each β on the training sets of each inner fold. We then select the β that corresponds to the model with the highest average *CPC* score over all inner folds. Once the optimal βs are selected for each outer fold, we use the winning models to compute one possible $\tilde{T}_{ij}$ on the testing sets of each outer fold.

For the XGBoost model, we choose the optimal maximum tree depth r, number of estimators e, and learning rate k, by first using a randomized grid search to simulate 100 random samples (r, e, k) from the Cartesian product of $r \in [2, 3, \cdots, 7]$, $e \in [25, 26, \cdots, 275]$, and $k \in [0.1, 0.2, \cdots, 0.5]$. We then find the optimal combination (r, e, k) for each outer fold by first simulating $\tilde{T}_{ij}$ from the models corresponding to each of the 100 parameter combinations (r, e, k) on the training sets of each inner fold, and then selecting the (r, e, k) set that corresponds to the model with the highest average *CPC* score over all inner folds. Once the optimal parameter combinations are selected for each outer fold, we next use the optimal models to compute $\tilde{T}_{ij}$ on the testing sets of each outer fold. During each simulation, we round the output data $\tilde{T}_{ij}$ to the nearest non-negative integer.

4 Results

Figure 2 shows box plots of the *CPC*, *CPL*, and *RMSE* scores produced by each model type for all outer folds. In addition, each of these figures shows the actual values for each metric over each outer fold, randomly adjusted, or "jittered", on the y-axis to prevent overlap. More specifically, we see in Figure 2 that all *CPC* and *CPL* scores produced by the XGBoost model are higher than all *CPC* and *CPL* scores produced by the gravity model, which are in turn higher than all of the *CPC* and *CPL* scores produced by the radiation model. Since all scores, rather than just all median scores, are higher in the XGBoost model than both other models, we are confident that the XGBoost model outperforms the gravity and radiation models on the *CPC* and *CPL* metrics. On the other hand, though we see that the median *RMSE* of the XGBoost model is also the best, or lowest, median *RMSE* among all three models, not all of the XGBoost model's *RMSE* scores are lower than scores coming from the other models. For example, the XGBoost *RMSE* score from one of the 10 testing sets is worse than all of the gravity model's *RMSE* scores and worse than eight of the radiation model's *RMSE* scores. This suggests that, though the XGBoost model produces a network with more similar structure to the ground truth network, it may also produce flow counts that are very far apart from one another.

5 Conclusion and Future Work

In this paper, we compared and discussed the performance of a standard gravity and radiation model with the XGBoost machine learning algorithm for predicting origin/destination commuter flows for a subset of block groups in the Knoxville MSA. We parameterized each model using two well known Census datasets and then evaluated and compared each model using the *CPC*, *CPL* and *RMSE* metrics.

Overall, we found that the XGBoost model far outperformed the gravity and radiation models on both the *CPC* and *CPL* metrics, indicating that it was able to re-create the

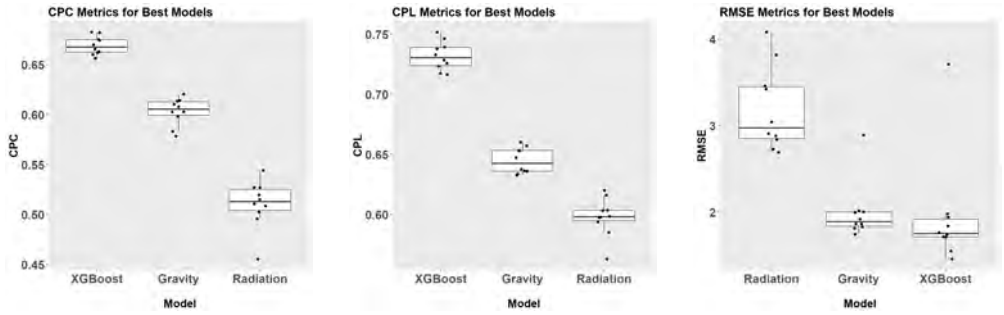

Figure 2 Box plots and horizontally jittered *CPC*, *CPL* and *RMSE* scores for the best performing models on each testing set.

original network better than the traditional models. However, we also discovered that the XGBoost model sometimes led to higher *RMSE* scores than both the gravity and radiation models, despite having the lowest median *RMSE* value. This may indicate that, given certain training/testing set combinations, the XGBoost model has the potential to produce estimates that are very far off from the ground truth flows. Thus, despite the fact that the XGBoost model re-creates the overall flows better than the gravity and radiation models, certain (though likely rare) links may have larger errors.

Though this study does indicate that the XGBoost model likely adds value to the field of commuter flow prediction, there are a few limitations and opportunities worth noting. For example, in a follow-up study it may be worth comparing more complex commuting models with the XGBoost model to determine if it still performs better. Additionally, one might want to add other machine learning models to the framework to determine if they add additional value on top of the XGBoost model. Furthermore, there may be other non-conventional input variables worth considering in the machine learning models that may further improve their performances.

6 Copyright

References

1 Ethem Alpaydin. *Introduction to machine learning*. MIT Press, 2014.

2 United States Census Bureau. Block group shapefiles for Tennessee [data file], 2010. URL: `https://www.census.gov/geo/maps-data/`.

3 United States Census Bureau. LEHD Origin-destination employment statistics [dataset], 2010. URL: `https://lehd.ces.census.gov/data/`.

4 Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

5 Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani. The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924, 2007.

6 Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. VSP, 1990.

7 Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

8 Celik H Murat. Sample size needed for calibrating trip distribution and behavior of the gravity model. *Journal of Transport Geography*, 18(1):183–190, 2010.

9 Caleb Robinson and Bistra Dilkina. A machine learning approach to modeling human migration. *arXiv preprint arXiv:1711.05462*, 2017.

10 Jan Rouwendal and Peter Nijkamp. Living in two worlds: a review of home-to-work decisions. *Growth and Change*, 35(3):287–303, 2004.