Urban Trip Generation Forecasting Based on Gradient Boosting Algorithm

Zhishuai Li
State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
lizhishuai2017@ia.ac.cn

Gang Xiong
Beijing Engineering Research Center of Intelligent Systems and Technology, Institute of Automation, Chinese Academy of Sciences
Beijing, China
Guangdong Engineering Research Center of 3D Printing and Intelligent Manufacturing, Cloud Computing Center, Chinese Academy of Sciences
Dongguan, China
gang.xiong@ia.ac.cn

Yu Zhang
Traffic Planning Department, Beijing Municipal Institute of City Planning and Design
Beijing, China
zy_jts@aliyun.com

Meng Zheng
Traffic Planning Department, Beijing Municipal Institute of City Planning and Design
Beijing, China
sd_zhengmeng@163.com

Xisong Dong
Beijing Engineering Research Center of Intelligent Systems and Technology, Institute of Automation, Chinese Academy of Sciences
Beijing, China
xisong.dong@ia.ac.cn

Yisheng Lv*
State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
yisheng.lv@ia.ac.cn

2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI) | 978-1-6654-3337-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/DTPI52967.2021.9540082

***Abstract*—The four-step transportation model plays an important role in urban planning. The quality of the first phase, i.e. trip generation, determines the performance of the global course. The majority of trip generation forecasting models highly rely on mathematical derivation and have many predictor variables during the prediction, which leads to low accuracy of results and requires laboriously hand-crafted design of input vectors. This paper is the first to introduce the gradient boosting decision tree (GBDT) algorithm for trip generation prediction, and harmonizes such a powerful machine learning method with traditional urban planning requirements to achieve better prediction performance. Unlike the commonly used linear regression method, GBDT can automatically perform feature selection and model the non-linear relationships between input and output variables. Experimental results on real-world residential travel census in Beijing prove that the GBDT model significantly outperforms the baseline and can forecast the trip generation more accurately.**

***Index Terms*—Four-step model, Trip generation, Gradient boosting decision tree, Linear regression**

The research was supported by National Key R&D Program of China (2020YFB2104001), National Natural Science Foundation of China under Grants U1909204, 61773381, U1811463 & U19B2029, and Chinese Guangdong's S&T project (2019B1515120030, 2020B0909050001).

*Corresponding author.

I. Introduction

Urban traffic demand prediction is an important application to evaluate the capacity of road traffic facility in city planning and intelligent transportation systems [1]. The four-step model serves as a dominated framework for demand prediction, and can be calibrated through travel household surveys data [2]–[5]. As its name implies, the model consists of four stages: trip generation, trip distribution, mode choice, and route assignment. Significantly, the accuracy of the first phase, i.e. trip generation, influences the effectiveness of the four-step model and becomes the backbone of travel demand modeling.

Trip generation predicts the travel frequency for particular purposes in each traffic analysis zones. Practically, it can be treated as a regression problem, which models the latent correlation with population demographics, land use, and additional socio-economic factors. It has two sub-tasks, which are the predictions for trip production and trip attraction, respectively. Representative methodologies used for trip generation are ordinary least squares linear regression (OLSLR), cross-classification methods, etc [6] [7], which have been applied to empirical studies. However, there are some drawbacks, evidenced by: 1) the OLSLR model assumes that the input

and output follow the linear functional form [6], while ignoring the non-linear relationship. And the parameters are limited by the identifiable attributes. 2) The cross-classification approach needs the cell-by-cell calculation, leading to the incomprehensive prediction. Therefore, an effective model should be designated to prevent or eschew these disadvantages.

The recent advances in machine learning may open a new chapter in modeling prediction tasks [8]–[14]. The decision tree-based models, such as classification and regression tree (CART) [15], can accordingly explain how a target can be predicted based on input features. Furthermore, the gradient boost algorithm [16] assembles multiple weak learners for getting stronger predictions in both classification and regression tasks. This provides us with novel insights. To response the above-mentioned issues, in this paper, we introduce a powerful gradient boost decision tree (GBDT) model [17] with several CARTs for trip generation. The GBDT is trained from previous household census data by gradient descent approach and yields the prediction in planned years. It can accurately model the non-linear correlation between input and target, and automatically shield the redundant variables while alleviating the hand-crafted feature selection engineering. Our contributions lie in that we are the first to harness the power of GBDT for trip generation and illustrate the feasibility of bridging the connection between the advanced machine learning methods and conventional traffic planning models.

The rest of the paper is organized as follows. Section II provides the formulation and methodology. Section III states the experimental results. The concluding remarks and future directions are discussed in Section IV.

II. Formulation and Methodology

A. Problem Statement

Trip generation aims to predict the number of trips, per unit time by purpose, that is generated by and attracted to each zone in study areas. Hence, it has two sub-tasks, namely trip production prediction and trip attraction prediction.

By the urban residential travel census and land use information, the origin-destination matrix between pairwise traffic analysis zones (TAZ) is easily captured. Then, the traffic volume of residents trip production and attraction for each TAZ in history years can be observed. In practice, the important and valuable factors related to trip generation include the populations of different types, demographic, occupation, income, ownership of transportation tools, etc. The trip production/attraction P_i and A_i for TAZ i are calculated by

$$P_i = \mathcal{F}_\alpha\left(o_i, u_i, w_i, \ldots\right), \tag{1}$$

$$A_i = \mathcal{F}_\beta\left(o_i, u_i, w_i, \ldots\right), \tag{2}$$

where o_i, u_i, w_i represent the input factors, and α, β are the parameters of respective trip generation models. For brevity, in the following, these attributes are treated as input vector $\mathbf{x}$, and y is used to represent the production or attraction for respective TAZ. So the objective is converted to find optimal parameters θ^* for model $\mathcal{F}$, which can be formulated as

$$\theta^* = \arg\min_{\theta} L(\mathcal{F}(\mathbf{x}), y), \tag{3}$$

where L is the evaluated metric, which can be mean square error. The θ^* are trained by the data from previous years, then the inference can be applied for planning year.

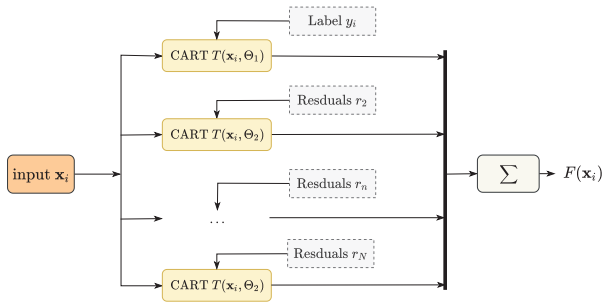

Fig. 1. The diagram of GBDT model, which consists of N CARTs and makes the prediction by adding the output from all leaners. The dashed boxes represent the ground truth used to train each CART respectively.

B. The CART sub-learner for GBDT

The CART is adopted as weak learner, while the GBDT integrates multiple CARTs to construct a single strong learner and obtain better predictive performance.

The CART model works by repeatedly partitioning the data into multiple sub-regions according to its features, so that the results in each region are as homogeneous as possible. The objective is to find the optimal split feature j and split value s from input $\mathbf{x}$. Since CART is a binary tree, only two sub-regions are generated in each division. By traversing each dimensional split feature j and split value s, the loss function l is calculated for the CART at each region, i.e.

$$l = \min_{j,s}[\min_{c_1} \sum_{\mathbf{x}_i \in \mathcal{R}_1} (y_i - c_1)^2 + \min_{c_2} \sum_{\mathbf{x}_i \in R_2} (y_i - c_2)^2], \tag{4}$$

where $\mathcal{R}_1 = \left\{\mathbf{x}^{(j)} \mid \mathbf{x}^{(j)} \leq s\right\}, \mathcal{R}_2 = \left\{\mathbf{x}^{(j)} \mid \mathbf{x}^{(j)} > s\right\}$ denote two subregions divided by value s under j-th feature, y_i is the label of input $\mathbf{x}_i$, M_t is the number of samples in region $\mathcal{R}_t$, c_1, c_2 are the mean of samples in respective regions $\mathcal{R}_1, \mathcal{R}_2$, that is,

$$c_t = \frac{1}{M_t} \sum_{\mathbf{x}_i \in \mathcal{R}_t} y_i, t = 1, 2. \tag{5}$$

For the data in respective sub-regions, the area should be divided by the above steps unceasingly, until the number of leaves of CART equals threshold J.

$$T(\mathbf{x}, \Theta) = \sum_{m=1}^{J} c_m I\left(\mathbf{x} \in R_m\right), \tag{6}$$

where Θ is the hyperparameters of CART, the $\mathbf{x}$ is divided into J regions (i.e. the threshold), and each region has a fixed output value c_m, $I\left(\mathbf{x} \in R_m\right) = 1$.

C. The GBDT ensemble Model

In this paper, CART is the basic learner of GBDT, and each one has the same depth and number of leaves. Each newly added weak learner fits the residuals from the last step so that the model can improve, and the GBDT thus becomes a strong regressor by integrating trees using a gradient boosting technique. Fig. 1 shows the diagram of the GBDT algorithm, and each weak learner indicates a CART.

Note that existing trees in the model do not change when a new tree is added for predicting the residuals. Therefore, in the gradient boosting method, each new weak leaner is established to reduce the residual error of the previous model following the negative gradient direction. The mean square error is adopted as loss function for the GBDT, i.e.

$$L = \frac{1}{M}\sum_{i}^{M}\left(y_i - f_1\left(\mathbf{x}_i\right)\right)^2, \tag{7}$$

where M is the number of samples, $f_1\left(\mathbf{x}_i\right) = T\left(\mathbf{x}_i, \Theta_1\right)$, Θ_1 represents the parameters of the first CART T. By contrast, loss l in CART is used to determine the optimal Θ of the trees, while L here is used to calculate the residual prediction error. For the sample $(\mathbf{x}_i, y_i)$, the negative gradient value for the residual can be calculated as

$$r_i^2 = -\frac{\partial L\left(y_i, f_1\left(\mathbf{x}_i\right)\right)}{\partial f_1\left(\mathbf{x}_i\right)}. \tag{8}$$

Then the input $\mathbf{x}_i$ and r_i^2 are constructed as a new sample, and the second CART $T\left(\mathbf{x}_i, \Theta_2\right)$ is added to predict r_i^2, while the prediction of the GBDT is the sum of the two trees, i.e., $f_2\left(\mathbf{x}_i\right) = T\left(\mathbf{x}_i, \Theta_2\right) + f_1\left(\mathbf{x}_i\right)$. For the residual r_i^N from the $(N-1)$-th CART, a new sample $\left(\mathbf{x}_i, r_i^N\right)$ can be constituted and used to train the N-th CART, while r_i^1 can be regarded as y_i. The training process of GBDT is demonstrated in Algorithm 1. And the total prediction for input $\mathbf{x}_i$ is

$$\mathcal{F}\left(\mathbf{x}_i\right) = T\left(\mathbf{x}_i, \Theta_N\right) + f_{N-1}\left(\mathbf{x}_i\right) = \sum_{j=1}^{N} T\left(\mathbf{x}_i, \Theta_j\right), \tag{9}$$

where N is the number of basic learners, i.e. CARTs.

Algorithm 1: The training of GBDT model for urban trip generation prediction.

Inputs: Train set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)\}$; The number of basic learners: N; Loss function: L; The number of leaves J.

Outputs: The trained GBDT model $\mathcal{F}\left(\mathbf{x}_i\right)$.

Initialize the first tree: $\Theta_1 = \arg\min_{\Theta_1} \sum_i^M L\left(y_i, f_1(\mathbf{x}_i)\right)$;

for t = 2 *to* N **do**

  Calculate the negative gradient of the model under the t-th iteration: $r_i^t = -\frac{\partial L(y_i, f_{t-1}(\mathbf{x}_i))}{\partial f_{t-1}(\mathbf{x}_i)}, i = 1, \ldots, M$;

  Train the t-th CART by $\left(\mathbf{x}_i, r_i^t\right), i = 1, \ldots, M$, the split regions are denoted as $R_{tj}, j = 1, 2 \ldots, J$;

  For the region $R_{tj}, j = 1, 2 \ldots, J$, find the optimal c_j^t according (4)(5);

  Build the the t-th CART: $f_t(x) = \sum_i^J c_i^t I\left(x \in R_{ti}\right)$

Return $\mathcal{F}(x) = \sum_{t=1}^{N} \sum_{j=1}^{J} c_j^t I\left(x \in R_{tj}\right)$

III. Experiments and Results

In this section, we analyze the results from the GBDT-based trip generation method and compare the model with a commonly used method in the trip generation, i.e. OLSLR.

A. Data Description

The experimental data is from Beijing residential household travel survey in 2017. The covered attributes include the information of populations, tourists, hospitals, occupations, and educations. The data contains the respective volume of trip production and attraction for $2,006$ traffic analysis zones. That is, the size of samples is $2,006$. Trip generation is divided into two types of families with and without cars under four purposes, and finally, eight types of trips are generated respectively, including CA-HBW (home-based work with cars), CA-HBS (home-based school with cars), CA-HBO (home-based other trips with cars), CA-NHB (non-home-based trip with cars), and NCA-HBW (home-based work without cars), NCA-HBS (home-based school without cars), NCA-HBO (home-based other trips without cars) and NCA-NHB (non-home-based trip without cars).

B. Experimental Setting

To conduct the experiment, we use the sci-kit learn library with Python programming language. And the experimental hyperparameters are set as follows:

- The number of weak learners is 100.
- The loss function is the mean square error.
- The maximum depth of the decision tree is 4.
- The minimum sample number of leaves is 2.
- The ratio between train set and test set is $8:2$.

The evaluated metrics in this experiment are root mean square error (RMSE) and mean absolute error (MAE), which can be calculated by:

$$\mathrm{RMSE} = \left[\frac{1}{M}\sum_{i=1}^{M}\left(y_i - \hat{y}_i\right)^2\right]^{1/2}, \tag{10}$$

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|y_i - \hat{y}_i\right|, \tag{11}$$

where M is the number of samples, $\hat{y}_i$ is the predicted value, namely $\hat{y}_i = \mathcal{F}(\mathbf{x}_i)$, y_i is the ground truth.

C. Results and Analysis

We perform predictions for trip production and attraction associated with eight purposes, respectively. Fig. 2 shows the RMSE and MAE metrics between predicted and true values in eight purposes of respective OLSLR and GBDT models. It is evident that the GBDT model (orange bar) significantly outperforms the OLSLR method (blue bar) in

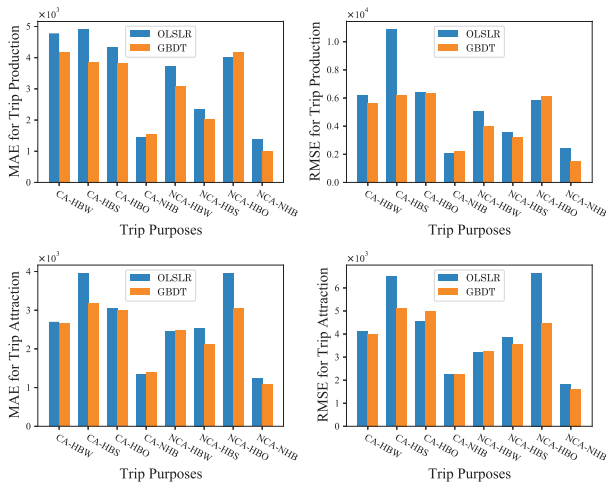

Fig. 2. The evaluated errors for the two models across eight trip purposes. CA and NCA are the short of travel with cars and without cars respectively. Lower values indicate better performance.

TABLE I
THE COMPARISON BETWEEN TWO METHODS ACROSS EIGHT PURPOSES

Methods	mean Production		mean Attraction	
	RMSE ($\times 10^2$)	MAE ($\times 10^2$)	RMSE ($\times 10^2$)	MAE ($\times 10^2$)
OLSLR	33.03	26.48	51.96	41.23
GBDT	29.34	23.28	43.56	36.04

general, especially for the trip attraction prediction of CA-HBS (about $10\% \sim 25\%$ improvement). In detail, for the attraction of NCA-HBW prediction, the performance of the two models is comparable. While the OLSLR model performs slightly better than the GBDT model in the trip production prediction for NCA-HBO, and all of them achieve promising performance for the trip production prediction of non-home-based trips.

Moreover, we demonstrate the prediction errors of mean production and attraction across the eight purposes in Table I. It can be seen that the GBDT model is superior to the OLSLR model in all indicators (about 12% overall improvement). Another unshowable advantage of GBDT is that it does not require the operation of hand-crafted feature selection, since it works by traversing all the split features and values judging by the loss function. All of those prove that the introduced GBDT model is better than the OLSLR method for the task.

IV. CONCLUSION

Trip generation forecasts the number of trips that begin from or end in each TAZ, which is the first phase of the four-step travel demand prediction. This paper is the first to introduce the GBDT algorithm for solving the problem of trip generation prediction tasks in urban planning and harmonizes such a powerful machine learning method with traditional urban planning requirements to achieve better prediction performance. In essence, the GBDT model integrates weak learners with CART via the gradient descent approach, which can adjust its structure according to data characteristics. Hence it fits well the non-linear relationship between input and output in the task, and does not need hand-crafted feature selection. By such a data-driven method, the pattern of input vectors can be extracted effectively. We also demonstrate the superiority of the introduced GBDT model by comparing it with the conventional OLSLR model. Our future work intends to evaluate and develop effective algorithms for remaining phases in four-step model and thus achieve better performance for the overall urban travel demand prediction.

REFERENCES

[1] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 630–638, 2010.

[2] G. Xiong, F. Zhu, X. Liu, X. Dong, W. Huang, S. Chen, and K. Zhao, "Cyber-physical-social system in intelligent transportation," *IEEE/CAA Journal of Automatica Sinica*, vol. 2, no. 3, pp. 320–333, 2015.

[3] F. Zhu, Y. Lv, Y.-y. Chen, X. Wang, G. Xiong, and F.-Y. Wang, "Parallel transportation systems: Toward IoT-enabled smart urban traffic control and management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4063–4071, 2019.

[4] M. Zhong, R. Shan, D. Du, and C. Lu, "A comparative analysis of traditional four-step and activity-based travel demand modeling: A case study of tampa, florida," *Transportation Planning and Technology*, vol. 38, no. 5, pp. 517–533, 2015.

[5] M. G. Dowd, "Modeling inundation impacts on transportation network performance: A GIS and four-step transportation modeling analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[6] J. S. Chang, D. Jung, J. Kim, and T. Kang, "Comparative analysis of trip generation models: Results using home-based work trips in the seoul metropolitan area," *Transportation Letters*, vol. 6, no. 2, pp. 78–88, 2014.

[7] K. M. Currans, G. Abou-Zeid, K. J. Clifton, A. Howell, and R. Schneider, "Improving transportation impact analyses for subsidized affordable housing developments: A data collection and analysis of motorized vehicle and person trip generation," *Cities*, vol. 103, p. 102774, 2020.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[9] Y.-y. Chen, Y. Lv, Z. Li, and F.-Y. Wang, "Long short-term memory model for traffic congestion prediction with online open data," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 132–137.

[10] Z. Li, G. Xiong, Y.-y. Chen, Y. Lv, B. Hu, F. Zhu, and F.-Y. Wang, "A hybrid deep learning approach with gcn and lstm for traffic flow prediction," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1929–1933.

[11] Z. Li, G. Xiong, Y. Tian, Y. Lv, Y.-y. Chen, P. Hui, and X. Su, "A multi-stream feature fusion approach for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[12] X. Zhao, Z. Li, Y. Zhang, and Y. Lv, "Discover trip purposes from cellular network data with topic modelling," *IEEE Intelligent Transportation Systems Magazine*, 2020.

[13] L. Li, S. Wang, and F.-Y. Wang, "An analysis of taxi drivers route choice behavior using the trace records," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 576–582, 2018.

[14] D. Kang, Y. Lv, and Y.-y. Chen, "Short-term traffic flow prediction with lstm recurrent neural network," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[15] C. D. Sutton, "Classification and regression trees, bagging, and boosting," *Handbook of statistics*, vol. 24, pp. 303–329, 2005.

[16] N. Duffy and D. Helmbold, "Boosting methods for regression," *Machine Learning*, vol. 47, no. 2, pp. 153–200, 2002.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.