Using Deep Gravity Model for BART Ridership Prediction

Kaihang ZHANG

Department of Civil and Environmental Engineering, University of California, Berkeley, CA, United States, 94704. Email: kaihangzhang@berkeley.edu

ABSTRACT

Transportation demand generation has long been a hot research area. It aims at generating trip counts between locations by taking socio-economic data as input. With this trip demand information in hand, transportation authorities can manage the public resource in a more cost-effective way, especially in the background of the post-pandemic situation. The benefit is particularly important for public transportation systems because they rely highly on management. Unfortunately, demand trip count data is not always available. Therefore, demand generation is one of the possible ways to get traffic demand. In this report, we show how can the state-of-the-art method be implemented into real-world data. We applied the Deep Gravity model which has been shown in previous works to have appropriate performance for the trip generation task. We tested the model on the Bay Area Rapid Transit (BART) system and found the model can achieve a very good performance.

1. INTRODUCTION

During the COVID-19 pandemic, public transit ridership has been significantly influenced by the virus and the relevant policies such as lockdown. As the post-pandemic era begins, the public transportation market is recovering and leads to a remarkable increase in transit ridership. The public transportation demand information can be helpful to repair the airline market in a more cost-effective way. However, OD-level transit ridership demand data (i.e., trip counts between each origin and destination stations) are not always available, especially in some developing regions. By contrast, socio-economic data (e.g., population of the origin and destination cities) provided by local governments or international organizations are easier to obtain. In this case, we can use predictive models to generate the demand data from socio-economic data.

The approaches to realize the trip generation purpose can be categorized into two types. Conventional and data-driven model. The most representative conventional model is the Gravity Model (GM). The gravity model was firstly proposed by (Zipf, 1946) and is often used to generate traffic flows based on population and OD distance. The model is deterministic and efficient. Besides, the idea is straightforward and easy to understand. Therefore, a lot of work in industry is based on GM. Nevertheless, the model is restricted to only one variable as input (such as population) and may underestimate the complicated relationship between traffic demand and socio-economic characteristics. Data driven models can tackle this problem in a better way because they depend highly on case-specific data. Thanks to the blooming of machine learning, many applications of data driven models have been implemented in the field of transportation, including trip generation. Recently, Simini et al. (2021) proposed a multiplayer-perception-based gravity model, i.e., the Deep Gravity (DG) model, which inputs both the census information of OD pairs and OD distances, and outputs trip counts between OD pairs. This model can achieve better performance in terms of generate traffic flows more similar to the real flows. Though being tested to be efficient in three different locations (New York, Italy, and England) by Simini et al. (2021), the model has not been proved to be effective under public transportation systems.

In this paper, we implemented the DG model in the Bay Area Rapid Transit (BART) system and compared its performance to the traditional GM model. The remainder of this paper is organized as follows. In Section 2, we introduce the two models applied in this paper. In Section 3, we introduce the smart location dataset and the BART ridership dataset. In Section 4, we present the results of these models. In Section 5, we discuss the results we obtained and draw a conclusion. Finally, we provide our insights for future work in Section 6.

2. TECHNICAL APPROACH

a. Gravity Model

We take the Gravity Model (GM) as the base model in this report. According to (Zipf, 1946), the number of trips between two locations is respectively proportional to the population of them, and is negatively related to the distance between these two communities. Formatively,

$$T_{i,j} \propto \frac{P_i P_j}{d_{i,j}} \tag{1}$$

, where $T_{i,j}$ denotes the trip counts between location i and location j, P_i and P_j represents the population of location i and location j respectively, $d_{i,j}$ represents the distance between two locations. In our project, we use the singly constrained Gravity Model, which fixes the total outflow originating from each location. Hence, the target of the model is to estimate the number of trips to each destination.

b. Deep Gravity Model

The Deep Gravity Model (DG) implemented here is inspired by Simini et al. (2021), which is a typical fully connected multilayer perceptron model which takes the census information of both locations as well as the distance as input. It outputs the trip counts between the two locations. The model layout is shown in Figure 1, there are 15 hidden layers in total, the first 6 of which have 256 dimensions while the last 9 have 128 dimensions.

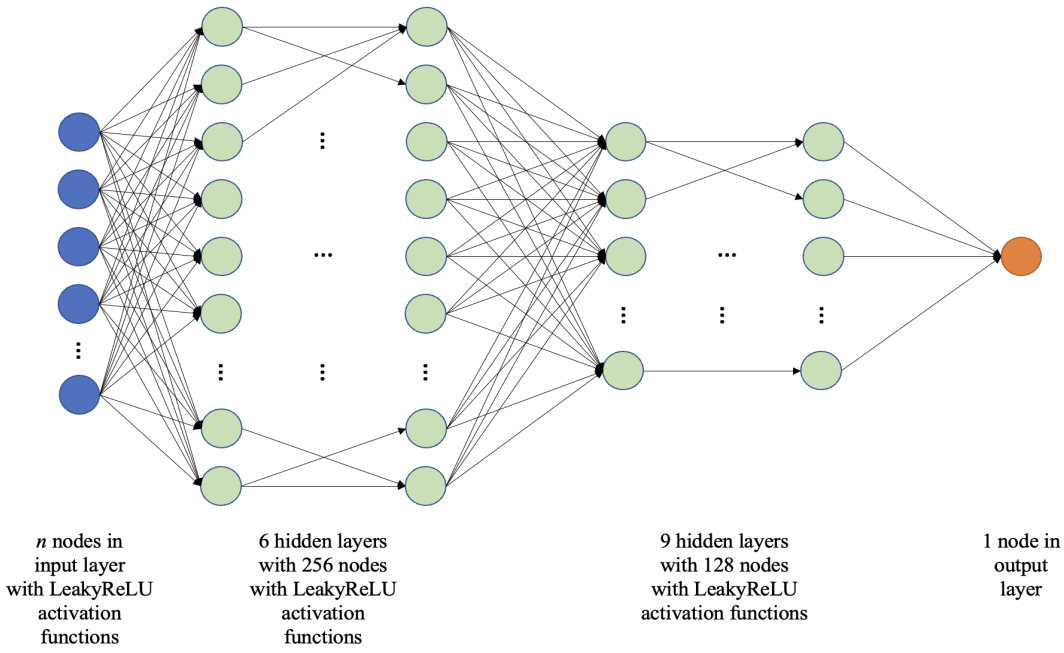


Figure 1. Deep Gravity Model

3. DATA

The BART ridership dataset provides us with the OD trip data in hour level and month level. In this report, we will use the hourly data collected in 2019 and then aggregate the data to a

annual level. Figure 2 is a quick overview of the data structure, where only the first ten rows are shown. As we can see, there are five columns, i.e., date, hour, origin station, destination station, and trip count[1]. The station names are shown in abbreviated forms, and the look-up table can be found on the BART website[2].

	A	B	C	D	E
1	1/1/19	0	12TH	12TH	3
2	1/1/19	0	12TH	16TH	4
3	1/1/19	0	12TH	ANTC	1
4	1/1/19	0	12TH	BAYF	1
5	1/1/19	0	12TH	CIVC	2
6	1/1/19	0	12TH	COLM	1
7	1/1/19	0	12TH	COLS	1
8	1/1/19	0	12TH	CONC	1
9	1/1/19	0	12TH	DALY	1
10	1/1/19	0	12TH	DELN	2

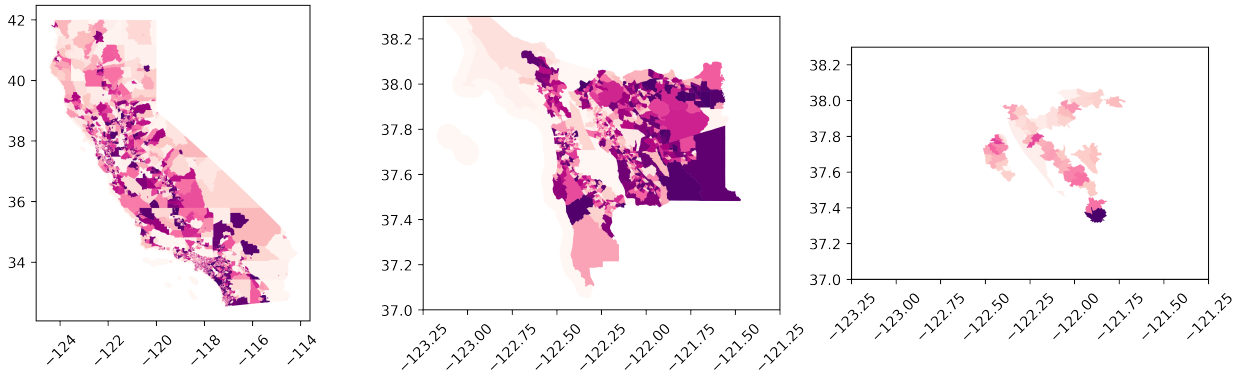
Figure 2: BART OD trip data (shown in Microsoft Excel)

Now that we are familiar with the trip data, we will introduce the socio-economic census data in this and the next paragraph. The smart location dataset is a very powerful data source that consists of over one hundred features of each census block group (CBG) (a unit of census data, subdivision of census tract). The data are available across the United States and are easy to access via a GIS software. The features it contains include total population, total geometric area, number of households owning 0/1/2+ cars, total employment, etc. We preliminarily chose 18 of them, the attributes and their meanings can be found in Appendix A. A full description and guideline of the dataset can be found in the documentation provided by United States Environmental Protection Agency (Chapman et al., 2021).

With a goal of analyzing the BART system, the region we focus on is the bay area itself. To be more specific, the BART catchment area. As mentioned before, the Smart Location Dataset is a nationwide dataset, hence the first step is to extract the area we are interested in. Figure 3 illustrates different data levels, where the total population data of each block group are shown as a demonstration. The darker the color is, the more people there are in that area. The three figures also present our data processing steps: we first extract the state level data, then the San Francisco–

[1] Bay Area Rapid Transit (BART), "Hourly Ridership by Origin-Destination Pairs," *README*. http://64.111.127.166/origin-destination/READ%20ME.txt
[2] Bay Area Rapid Transit (BART), "BART Legacy API Documentation," *Station Abbreviations*. https://api.bart.gov/docs/overview/abbrev.aspx

Oakland–Berkeley Bay Area, finally the BART catchment area. In Figure 3(c), we preliminarily assumed that the BART catchment area is a circle with a radius of 5.5 km (3.5 miles).



(a) California (b) Bay Area (c) BART catchment area

Figure 3. Smart location data

4. RESULTS

The dataset was splitted into two parts, namely the training data and the testing data. The training data is 70% of the entire data set while the testing data is the rest of them. All the results shown in this section are the model results from the testing data. The results of the generated flow and real flow is evaluated by four metrics, one of them is the Common Part of Commuters and the other three are three errors. They are defined in Equation (2) through Equation (5).

$$CPC = \frac{2\sum \min(y^g, y^r)}{\sum y^g + \sum y^r} \times 100\% \tag{2}$$

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}(y^r - y^g)^2}{N}} \times \frac{1}{y^r_{max} - y^r_{min}} \times 100\% \tag{3}$$

$$NMAE = \frac{1}{N}\sum_{i=1}^{N}|y^g - y^r| \times \frac{1}{y^r_{max} - y^r_{min}} \times 100\% \tag{4}$$

$$\text{SMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y^g - y^r|}{(|y^g| + |y^r|)/2} \times 100\% \tag{5}$$

, where y^g represents the generated flow and y^r represents the real flow. It is easy to understand the error equations because they represent how the generated trips are away from the real trip

counts. As for CPC, we can have some insights on what it means. CPC is a metric that can measure the difference between the generated flow data and the ground truth data. It is originally defined as $\frac{2|X \cap Y|}{|X|+|Y|}$, representing the similarity of two samples. Thus, if two data samples are exactly the same, the CPC value will be 100%, and the lower the more different they are.

Since the machine learning model is highly stochastic, we execute the DG model 25 times and take the average values of the results. The evaluation metrics of both models are shown in Table 1. We also provide an illustration of the generated flow and the real flow. Since they are flow of different OD pairs, we show in Figure 4 that the results in a scalar plot and sort the value just for better visualization. Additionally, we also show the spatial visualization of how the trips between locations are on the map as shown in Figure 5. In Figure 5, blue lines mean low flow which are below 30% of the maximum flow; red lines mean the medium flow which are between 30% and 60% of the maximum flow; and yellow lines represent the high flow which are higher than 60%.

Table 1. Model Performance

Model\Metrics	CPC		NRMSE		NMAE		SMAPE	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
GM	35.05%	/	13.96%	/	7.78%	/	119.17%	/
DG	69.74%	0.117	8.00%	0.146	4.51%	0.008	78.42%	0.136

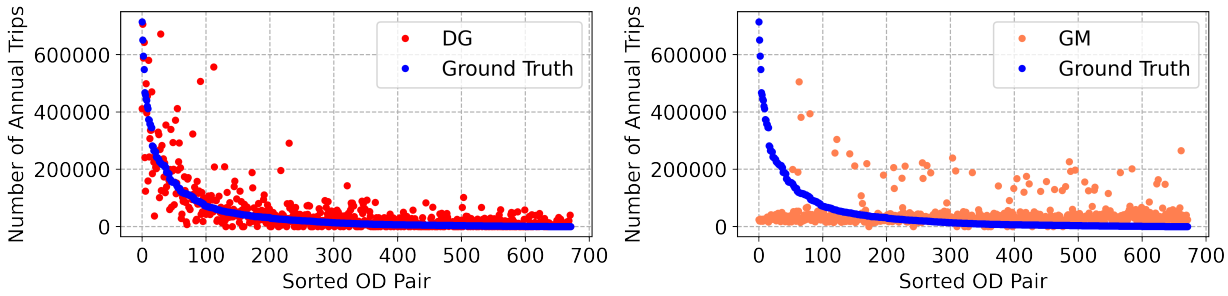
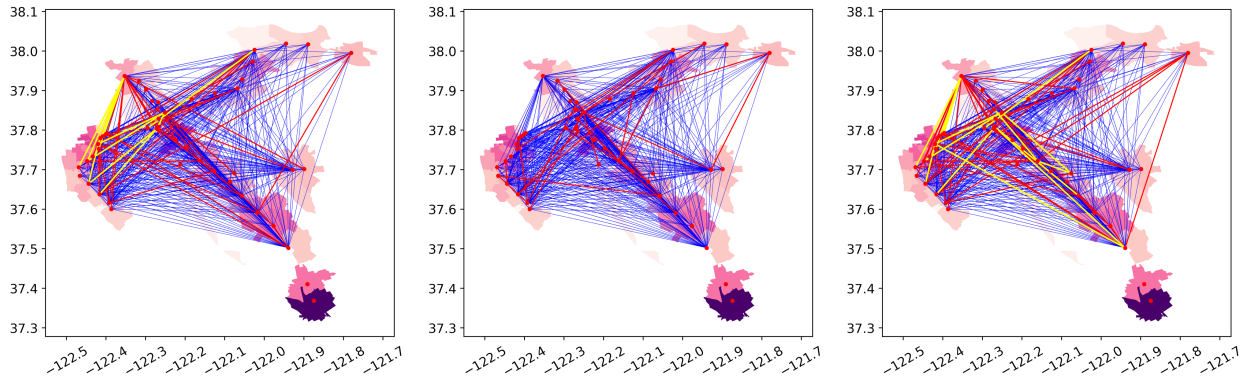


Figure 4. Trip Generation Visual Comparison Between Two Models



(a) Real Flow (b) GM Generation (c) DG Generation

Figure 5. Spatial Visualization of Trip Generation

5. DISCUSSION and CONCLUSION

As we can see from the previous section, the Deep Gravity model (DG) has better performance in all the evaluation metrics. CPC value can measure how the generated flow be similar to the real flow. In Simini et al. (2021)'s paper, their best result has a CPC value of 70%, which is comparable with what we get from the BART ridership scenario. This means that it is practicable to apply the DG model in the public transportation system at least for BART. Moreover, the CPC value is much higher than that of GM, showing that the DG model can better generate the flow compared with GM. As for the other three error metrics, DG model also holds the lower error than those of GM. Therefore, we can conclude that the data-driven DG model outperforms the conventional GM.

Apart from the numerical values, we can also get some insights from Figure 4 and Figure 5. From Figure 4, we can tell that the results of DG model can fit the real data adequately, while the results of GM can hardly fit the real data. Combined with Figure 5, where GM can hardly generate high flow trip counts, we can conclude that the gravity model is suitable for public transportation ridership generation. On the other hand, DG can provide convincing results however the flow is.

6. FUTURE WORK

Our future work can be focused on testing the model transferability on different transportation systems. That is, to see how the model would perform if we directly transferred the model to another similar transit system. If the model could not provide appropriate results, would

some popular data processing techniques work? For example, the Principal Component Analysis (PCA) technique, which can reduce the data dimension and thereafter making the feature data satisfy a nearly *i.i.d.* condition; or Transfer Component Analysis (TCA), which can realize domain adaptation of two feature spaces and therefore making the feature space of two systems to be similar.

REFERENCE

Chapman, J., Fox, E. H., Bachman, W., Frank, L. D., Thomas, J., & Reyes, A. R. (2021). *SMART LOCATION DATABASE - TECHNICAL DOCUMENTATION AND USER GUIDE (Version 3.0)*. US EPA. https://www.epa.gov/sites/default/files/2021-06/documents/epa_sld_3.0_technicaldocumentationuserguide_may2021.pdf

Simini, F., Barlacchi, G., Luca, M., & Pappalardo, L. (2021). A Deep Gravity model for mobility flows generation. Nature Communications, 12(1), 6576. https://doi.org/10.1038/s41467-021-26752-4

Zipf, G. K. (1946). The P1 P2/D Hypothesis: On the Intercity Movement of Persons. American Sociological Review, 11(6), 677–686.

Appendix

A. Attributes Table for SLD

Number	Short Name	Variable Name	Number	Short Name	Variable Name	Number	Short Name	Variable Name
1	TotPop	Total Population	7	E_MedWageWk	# of workers earning more than $1250/month but less than $3333/month	13	D4B050	Proportion of CBG employment within 0.5 mile of fixed guideway transit stop
2	AutoOwn0	Number of households that 0 automobiles	8	E_HiWageWk	# of workers earning $3333/month or more	14	D4C	Aggregate frequency of transit service within 0.25 miles of CBG boundary per hour during evening peak period
3	AutoOwn1	Number of households that 1 automobiles	9	D1B	Gross population density	15	D4D	Aggregate frequency of transit service [D4c] per square mile
4	AutoOwn2p	Number of households that 2 automobiles	10	D1C	Gross employment density	16	D5AR	Jobs within 45 minutes auto travel time, time-decay (network travel time) weighted
5	TotEmp	Total employment	11	D3A	Total road network density	17	D5AE	Working age population within 45 minutes auto travel time, time-decay (network travel time) weighted
6	E_LowWageWk	# of workers earning $1250/month or less	12	D4A	Distance from the population-weighted centroid to nearest transit stop	18	NatWalkInd	Walkability index