Investigating the effect of weather on bus origin-destination patterns: A case study from Changsha, China

Extended abstract

Tianli Tang, Ronghui Liu, Charisma Choudhury

(Institute for Transport Studies, University of Leeds, UK)

1 Background

Weather has a significant impact on the transport system, both on demand (Zhou et al., 2017) and supply side (Böcker et al., 2013). This is evident in recent literature which demonstrate that there is a strong correlation between weather conditions public transport ridership (Singhal et al., 2014;). However, these researches do not capture the full complexity of the impact on weather on public transport. In particular, to the best of our knowledge, there is not any work which estimates the impact of weather on origin-destination (OD) patterns in the context of bus services. Since OD matrices are a key input in the public transport planning problem, we aim to fill in this research gap in this paper.

Current methods for OD estimation for public transport are largely based on the trip features (Barry et al., 2002; Hou et al., 2012) and less on the impacts from other extended conditions. Traditionally, the ridership is recorded manually using sample survey, point check and ride check (Ceder, 2007). At present, however, automatic data system (ADS) is utilised extensively to collect the information from passengers and vehicles to estimate the ridership, and this method is more automatic, faster and cheaper. The smart card (SC) data from ADS has been widely used as the most attractive resources to estimate the OD matrix for bus ridership (Bagchi and White, 2005). However, the existing literature on OD estimation based on SC data are not applicable for the following group of passengers:

- who use a different mode of transport,
- who take bus lines running by different companies,
- who are not commuters or regular passengers, or
- who do not return to their origin stops.

Our research focuses to address this research gap as well by proposing a novel OD estimation methodology.

2 Objective

To develop a comprehensive analyses of deriving bus OD matrices from SC data (for all users), investigate the variations in the OD patterns due to weather and develop a mathematical relationship between OD matrices and different weather conditions. The weather conditions include precipitation and temperature in particular.

3 Case Study Site and Data Cleaning

The case study site is a sub-set of the bus network in the city of Changsha in central south of China. The study network includes 9 bus lines running through the city centre. One-week SC and GPS data, from 2nd to 8th April in 2016, has been calculated. Changsha's SC data is a typical example of 'open' AFC system, where passengers swipe cards only at boarding. The data records passengers' ID, boarding time, boarding line and boarding vehicle without any boarding or alighting stops. The bus operation rules and data we have:

- Passengers are requested to swipe their card when boarding but not alighting; the SC data records such information
- The GPS data of each bus
- Weather records

Before going to the estimation process, the characteristics of the data are analysed. In total, the SC logs have 730,738 records. 69,595 (10% of total) trips are formulated repetitive data. Meanwhile, 104,755 (14% of total) trips cannot be matched to the GPS data, and 105,350 (14% of total) trips are away from the normal network because of the temporary change. We consider these three kinds of trips be invalid data, and take the remaining 62% trips of total as cleaned data to be used in the estimation and analysis. All the cleaned data are categorised to five types:

- Trips in a chain: 68,421 (15%),
- Segments in transfer trips excluding last segments: 53,714 (12%),
- Last segment of transfer trips: 34,867 (8%),
- Other trips: 230,028 (51%), and
- Unknown origin trips: 64,008 (14%).

When doing the stop estimation, unknown origin trips cannot be estimated due to the missing origin stops, and the trips in type 'last segment of transfer trips' and 'other trips' are not suitable for the trip-chaining or transfer-trip methods. Moreover, 211,905 (47% of cleaned data) trips are the sole trip, and this percentage is much higher than 18.6% in Chicago (Zhao et al., 2007) and 4.9% London (Gordon et al., 2013). This is due to the factor that the SC data this study used is form one of the operating bus companies, but passengers might make the transfer between bus services of two different bus companies.

4 OD Estimating Methodology

We present here the method of estimating passengers' alighting stops based on machine learning (ML). Since the training samples and testing samples of ML require the destination stops of each trip, only trips in the first two types, trips in a chain and segments in transfer trips excluding last segments, are used in this study. To simplify the problem of interest and clarify the data analysis process, following assumptions are made:

- Each passenger owns only one card, and each card can be used by its owner, although in practice the same card may be shared with family members and friends;
- The trips that require transfer among different lines are regarded as the separate trips on those lines.

Since the SC data only records the boarding time and vehicle (and thus the line), the boarding and alighting stops, as well as alighting time, are unknown from the SC data. To identify this unknown information, the following two-step process are adopted.

Step 1: inferring the boarding stops via data fusion. Combining with the geographical information from bus GPS data, it can be inferred when a vehicle reaches and leaves each bus stop. Then, passengers' boarding stops can be estimated.

Step 2: inferring the alighting stops via classification. Trips on a particular line will be divided into several classes, which represent different destination stops. And this classification model is built by ML. The ML algorithm used is Gradient Boosting Decision Tree (GBDT) proposed by Friedman (2001), which is to use many decision trees to calculate the results and then makes the conclusion by the accumulation of the conclusion of each tree. The aim of each calculation is to decrease the residual in the last calculation. Figure 1 shows how the GBDT works to train the raw model.

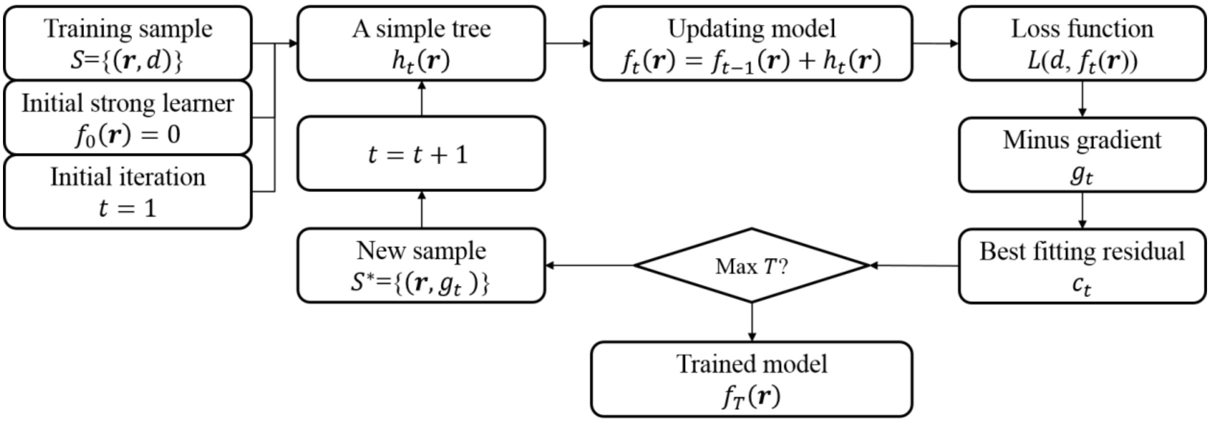

Figure 1. The flow chart of the training process in GBDT

The inputs of GBDT are the training samples, the selected loss function and the maximum number of iteration. Each training sample includes a feature vector, representing a trip by many feature variables, and the true destination stop of this trip. In the feature vector, three kinds of characteristics are selected to describe trips:

- Travel characteristics: date, time, day of week, line, origin stop, vehicle type, etc.
- Traveller characteristics: card ID, card type, frequent destination, all destinations, total number of trips, etc.

- Weather conditions: weather event, precipitation, temperature, dew point, air pressure, humidity, wind speed, etc.

Considering the rule of GBDT, the feature vectors are coded by One-Hot Encoding. According to the initial training sample, a simple tree is generated and then followed by the calculations of the loss function, minus gradient and the best fitting residual. A new sample for a new decision tree is then generated, where the real destination stops are replaced by the minus gradient, and the calculation follows. The process iterates for a certain times and stops. For each iteration, the new tree is to decrease the residual in the last action. And the simple tree in each circulation is accumulated as the final trained model that is the output of this algorithm. After that, this trained model will be applied to all kinds of trips, expressed as the feature vectors, and the possible destination stops of all trips can be estimated.

5 Initial Results

Followed by the methodology, the first step is origin stop estimation. After cleaning the data, 64,008 (14% of cleaned data) trips cannot be assigned their origin stops without a certain reason, while the rest (387,030 trips, 86% of cleaned data) find their origin stops correctly. In sum, the accuracy of the origin stop estimation method is over 85%. Then, the destination stop estimation is carried on. As mentioned at the beginning of methodology section, in this case, the destination stops of only 122,135 trips can identified by the trip chains and transfer trips, which are used as the dataset to test the GBDT method. 70% trips are chosen randomly as the training samples, and the rest for verification. Conclusively, the accuracy of the trained model is above 65%, and if the acceptable error is the error less than 3 stop, the accuracy is over 75%.

The impact of weather analyses is at its early stage now, but given the wide variation of temperature and precipitation in the area and the levels of variation in the OD matrices, we envision it will produce interesting results.

6 Conclusion

This study proposes a method to estimate the historical OD matrix for bus ridership, where the ML is used for the first time on this problem. One of the advantages of the proposed method is its applicability to all kinds of trips rather than to just regular passengers. It is because the destination stops can be inferred by the feature variables of each trip without identifying everyone’s daily trip sequence. Meanwhile, weather parameters are employed in the model, so the estimation is not only based on the trip characteristics but also referring to weather impacts. Combining these two contributions and innovations, the proposed method fills in gaps in the OD estimation for all trips and the weather effects on OD estimation. It is believed that the accurately estimated OD matrix can provide the planner with an all-sided, detailed and reliable data basis for all bus planning stages. Only the historical OD matrix can we measure based on this method, which is the first step in the whole planning cycle. In the further investigation, this historical ridership will be introduced to the ridership prediction and the following bus planning stages considering the weather. It will provide a better bus service in varying weather conditions.

7 References

Bagchi, M. and White, P.R. 2005. The potential of public transport smart card data. *Transport Policy*. **12**(5), pp.464-474.

Barry, J. et al. 2002. Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board.* (1817), pp.183-187.

Böcker, L. et al. 2013. Impact of Everyday Weather on Individual Daily Travel Behaviours in Perspective: A Literature Review. *Transport Reviews.* **33**(1), pp.71-91.

Ceder, A. 2007. *Public Transit Planning and Operation: Theory, Modeling and Practice*. Oxford, UK: Butterworth-Heinemann.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics.* pp.1189-1232.

Gordon, J. et al. 2013. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board.* **2343**, pp.17-24.

Hou, Y. et al. 2012. Origin-destination Matrix Estimation Method Based On Bus Smart Card Records. *Computer and Communications.* **30**(6), pp.109-114.

Singhal, A. et al. 2014. Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice.* **69**, pp.379-391.

Zhao, J. et al. 2007. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering.* **22**(5), pp.376-387.

Zhou, M. et al. 2017. Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies.* **75**, pp.17-29.