A Novel Passenger Travel-time and Destination Distribution Model for Day-ahead Forecasting

Yiren Liu[a], Lishuai Li[a], Yining Dong[a], Yang Zhao[b], and S Joe Qin[c,*]

[a] City University of Hong Kong, Hong Kong SAR, China
yirenliu2@cityu.edu.hk, lishuai.li@cityu.edu.hk, yining.dong@cityu.edu.hk
[b] Sun Yat-sen University, Shenzhen, China
zhaoy393@mail.sysu.edu.cn
[c] Lingnan University, Hong Kong SAR, China
joeqin@ln.edu.hk
[*] Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 26, 2024

Keywords: Subway systems, passenger flow, travel time distribution, flow patterns, short-term forecasting

1 INTRODUCTION

This study delves into the mobility of inbound passengers, focusing on their travel time and destinations. We construct a matrix that records the travel time and destination for passengers arriving at each station during specific time slots. Using this matrix, we establish a travel-time and destination distribution (TDD) matrix, which computes the proportion of passengers' mobility from each origin station. By quantifying the similarity of current time slot TDD with slot-ahead, day-ahead, and week-ahead, we unveil in-depth underlying mechanisms of passengers' periodic travel patterns, revealing predictable and regular mobility behavior. Therefore, we propose a TDD model to predict passengers' TDD one-day ahead. By combining this prediction with real-time passenger data from an originating station, we can accurately estimate the passenger flows to other stations and the corresponding travel durations. Consequently, the short-term passenger outflow at each destination station can be estimated. We validate our model using actual subway passenger data and demonstrate its superior performance compared to other well-known models, such as Auto-regressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Historical Average.

The traditional ARIMA and SARIMA models usually neglected the volatility in passenger flow influenced by uncertain real-time flow change Chen *et al.* (2019). In the metro ridership prediction task, how to conduct predictions incorporating real-time information is dramatically important Zhang *et al.* (2021). To capture real-time information, Pereira *et al.* (2015) gathers information from the Internet to predict transport arrivals under special events scenarios. This study introduced a groundbreaking approach for short-term outflow prediction by introducing the TDD matrix. This matrix effectively handles real-time passenger inflow, enabling accurate capturing of timely changes. Furthermore, by incorporating travel time as an extra dimension, the TDD matrix enhances the metric's value by reflecting a broader range of information regarding passenger mobility.

2 METHODOLOGY

The studied problem in this paper includes all subway lines and stations in Shenzhen Metro Cooperation (SMC) in 2013, comprising a complete system of 5 lines and 118 stations. The dataset utilized in this study was obtained from the Automatic Fare Collection (AFC) system of SMC in China, covering a period of 42 days from October 14 to November 24, 2013. Detailed description of the original AFC data, such as station names and transaction timestamps, can be found in Tang *et al.* (2019).

After initial processing of the original AFC data, each data record is structured to represent a single passenger trip based on Card ID, originating gate ("tap-in" action), destination gate ("tap-out" action), and the corresponding timestamps. The passenger data are aggregated with a time slot of a 15-minute interval, yielding 66 time slots each day, denoted as t. We use ℓ to represent travel time, which is the number of time slots from origin to destination. Since no travel-time records in the collected data exceeded two hours, $\ell \in [0, 8]$. Other notations, including the proposed TDD matrix, are listed below:

- **Passenger flow:** The inflow, denoted as $O(i; d, t)$, indicates the number of originating passengers from Station i during Time slot t on Date d. The outflow, represented by $D(j; d, t)$, refers to the number of passengers destined for Station j during Time slot t on Date d.

- **OD matrix:** The origin-based OD matrix, denoted with $o_j(i; d, t)$, captures the number of passengers who enter from Station i and exit Station j on Date d based on the entering time t.

- **Travel-time and Destination:** Variable $o_{j\ell}(i; d, t)$ quantifies the number of passengers from inflow $O(i; d, t)$ entering Station i and exiting from Station j on Date d with travel time ℓ.

- **Proportion of Travel-time and Destination:** $p_{j\ell}(i; d, t)$ represents the proportion of passengers originating from Station i to Station j during Time slot t on Date d destined with travel time ℓ, i.e., $p_{j\ell}(i; d, t) = \frac{o_{j\ell}(i;d,t)}{O(i;d,t)}$.

- **Travel-time and Destination Distribution Matrix:** $\mathbf{P}(i; d, t) \in \Re^{n \times 9}$ denotes the travel-time distribution of inflow passengers $O(i; d, t)$.

The OD pair from Pingzhou (PZ) to Gaoxinyuan (GXY) attracts the highest number of passengers during the morning hour, Figure 1 illustrates the TDDs of this OD pair for rush-hour slots on each day of two consecutive weeks. The vertical axis represents the origin time slot, and the horizontal axis depicts the travel time in minute with ticks of a 15-minute interval. Each TDD distribution has a number indicating the number of passengers of the OD pair in the time slot. This figure shows clear patterns of similarities or distinctions for week-to-week (W2W), day-to-day (D2D), or time-slot to time-slot (T2T) repeatability, which are summarized as:

i) The W2W repeatability is obvious for the same time-slot of the day and same day of the week. ii) The D2D repeatability is also noticeable for the same time-slot, especially for weekdays. However, iii) The T2T repeatability varies from one time slot to the next. Rush-hour slots are similar, but they differ significantly from the post-rush-hour slot.

To quantify these W2W, D2D, and T2T similarities, let $\mathbf{p}_j(i; d, t) = \texttt{vec}\{p_{j\cdot}(i; d, t)\}$ denote the vector of a TDD, N_t be the number of slots per day, N_d be the number of days of the data, and N_w be the number of weeks of the data. The W2W, D2D, and T2T similarities are defined and calculated as follows.

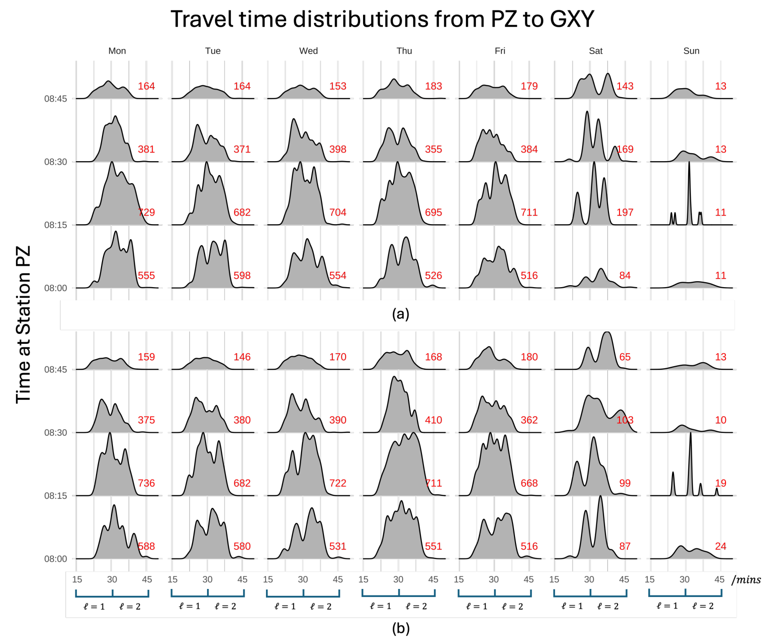

Figure 1 – *The distribution of trip destination-time between PZ and GXY in early morning of two weeks (a) October 14 to October 20, and (b) October 21 to October 27. The numbers represent the total number of passengers of the departing time slot.*

1. For each day of the week, calculate the W2W similarity as

$$\frac{\sum_{w=1}^{N_w-1} \mathbf{p}_j^\top(i,d+7w,t)\mathbf{p}_j(i,d+7w-7,t)}{\sum_{w=1}^{N_w-1} \|\mathbf{p}_j(i,d+7w,t)\|\|\mathbf{p}_j(i,d+7w-7,t)\|}$$

2. For each day of the week, calculate the D2D similarity as

$$\frac{\sum_{w=0}^{N_w-1} \mathbf{p}_j^\top(i,d+7w,t)\mathbf{p}_j(i,d+7w-1,t)}{\sum_{w=0}^{N_w-1} \|\mathbf{p}_j(i,d+7w,t)\|\|\mathbf{p}_j(i,d+7w-1,t)\|}$$

   where $d = 0, 1, ..., 6$ represents seven days of a week, $t = 1, 2, ..., 66$ represents time slots of a day.

3. Calculate the T2T similarity as

$$\frac{\sum_{d=1}^{N_d} \mathbf{p}_j^\top(i,d,t)\mathbf{p}_j(i,d,t-1)}{\sum_{d=1}^{N_d} \|\mathbf{p}_j(i,d,t)\|\|\mathbf{p}_j(i,d,t-1)\|}$$

   for rush-hour $t = 8:00 - 8:45$ and post rush-hour $t = 8:45 - 9:45$.

The calculated W2W, D2D, and T2T similarities from the Shenzhen subway data show that i) W2W and D2D similarities are much higher than the T2T similarities for all days of the week, especially for weekdays; ii) D2D similarities on Monday are less than those on other weekdays; and iii) T2T similarities are the least among all, with non-rush-hour similarity less than that of the rush-hour.

Based on the preceding analysis of the subway data we propose a TDD prediction model relying on the dominant D2D and W2W correlation in travel-time distributions as

$$\begin{gathered}\hat{\mathbf{P}}(i;d,t) = f(\{\mathbf{P}(i;d-a,t)\}_a, \{\mathbf{P}(i;d-7b,t)\}_b) \\ a = 1, 2, \ldots, n_d.\ b = 1, 2, \ldots, n_w.\end{gathered} \tag{1}$$

where n_d and n_w denote the daily and weekly look-back lengths, respectively. The function f maps input matrices to output. The model is individually constructed for each station i at time slot t. Since the TDD matrices in (1) are distributions with all non-negative elements, we adopt a linear model and the following optimization problem,

$$\begin{aligned} \hat{\mathbf{P}}(i; d, t) &= \sum_{a=1}^{n_d} \alpha_a \mathbf{P}(i; d-a, t) + \sum_{b=1}^{n_w} \beta_b \mathbf{P}(i; d-7b, t) \\ s.t. \quad & \alpha_a \geqslant 0, \beta_b \geqslant 0, \ \forall a \in [1, n_d], \forall b \in [1, n_w] \\ & \sum_{a=1}^{n_d} \alpha_a + \sum_{b=1}^{n_w} \beta_b = 1 \end{aligned} \tag{2}$$

where $\{\alpha_a\}$ and $\{\beta_b\}$ represent the coefficients to be estimated and the last equality guarantees that the prediction is a distribution matrix. The model (2) is effectively a weighted average of the past TDDs where the weights are estimated for each originating station and each time slot. As both the objective and inequality constraints are convex and the equality constraint is affine, (2) could be solved with a convex optimization program Boyd & Vandenberghe (2004).

The model for $\mathbf{P}(i; d, t)$ is actually a weighted average model for day-to-day and week-to-week relations. The slot-to-slot relation varies significantly depending on the time of the day, and therefore is not used in the model. With the predicted $\hat{\mathbf{P}}(i; d, t)$ and real-time inflow $O(i; d, t)$, the passenger outflows at destined stations can be estimated based on equation (3).

$$D(j; d, t) = \sum_i \sum_\ell O(i; d, t-\ell) p_{j\ell}(i; d, t-\ell) \tag{3}$$

3 RESULTS

The dataset is split into a training set spanning five weeks from October 14 to November 17, and a testing set encompassing from November 18 to November 24. To evaluate the proposed TDD model for estimating passenger flow, we present several commonly used time series models as benchmarks, such as Auto-regressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Historical Average, and assess their performance by comparing their Root Mean Squared Error (RMSE). SARIMA demonstrates superior performance over ARIMA and Historical Average in terms of test RMSE across all stations. Hence, our focus shifts to comparing the results between SARIMA and TDD models. Out of the 118 stations examined, the TDD model achieves the best test RMSE results for 102 stations, accounting for 86.4% of the total. SARIMA exhibits an average test RMSE of 25.85 across all stations, while TDD model achieves a mean value of 24.04, showcasing an improvement of nearly 10%.

References

Boyd, Stephen P, & Vandenberghe, Lieven. 2004. *Convex optimization*. Cambridge university press.

Chen, Enhui, Ye, Zhirui, Wang, Chao, & Xu, Mingtao. 2019. Subway passenger flow prediction for special events using smart card data. *IEEE Transactions on Intelligent Transportation Systems*, **21**(3), 1109–1120.

Pereira, Francisco C, Rodrigues, Filipe, & Ben-Akiva, Moshe. 2015. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, **19**(3), 273–288.

Tang, Liyang, Zhao, Yang, Cabrera, Javier, Ma, Jian, & Tsui, Kwok Leung. 2019. Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro. *IEEE Transactions on Intelligent Transportation Systems*, **20**(10), 3613–3622.

Zhang, Jinlei, Che, Hongshu, Chen, Feng, Ma, Wei, & He, Zhengbing. 2021. Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method. *Transportation Research Part C: Emerging Technologies*, **124**, 102928.