Wang, Kaipeng; Wang, Pu; Huang, Zhiren; Ling, Ximan; Zhang, Fan; Chen, Anthony

A Two-Step Model for Predicting Travel Demand in Expanding Subways

Published in:
IEEE Transactions on Intelligent Transportation Systems

DOI:
10.1109/TITS.2022.3166669

Published: 01/10/2022

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
A Two-Step Model for Predicting Travel Demand in Expanding Subways

Kaipeng Wang, Pu Wang®, Member, IEEE, Zhiren Huang®, Ximan Ling®, Fan Zhang, and Anthony Chen®

Abstract—In many cities, subways are expanding with new or extended lines being built and put into operations. The prediction of future travel demand in subway with the planned expansion is of significant importance because such information is crucial for new line planning and new network operations. In this study, we identify the determinant features from potential influential factors of passenger travel demand and develop a two-step model for predicting passenger travel demand in expanding subways. The proposed model is tested in an actual subway with a new line being put into operations, and achieves higher prediction accuracy than the benchmark models.

Index Terms—Subway, travel demand prediction, new lines, line extension.

I. INTRODUCTION

SUBWAY is characterized with the good features of high capacity, high speed, and high punctuality. Developing efficient and reliable subways is regarded as one of the most effective ways for mitigating the ubiquitous traffic congestion in big cities. Most big cities in the developed countries have built well-connected subway networks, while for many big cities in the developing countries the subway networks are still in their infancy or expanding stage. The expansion of a subway network is achieved by building new subway lines or extending existing subway lines. In the process of network expansion, the travel demand of passengers will also have considerable changes. More passengers will use the subway due to the enhanced network accessibility, and passengers can reach more regions through the expanded subway network. Predicting the passenger travel demand in subways with the planned expansion is an important issue because the travel demand information is crucial for planning new subway lines or line extensions and for deploying new network operation schedules.

Different from an expanding subway network, a stable subway network does not have structural change (i.e., no new lines or line extensions) and the full spectrum of passenger travel demand information (i.e., travel demand of each OD pair) can be recorded by the smartcard data. Consequently, the smartcard data have been widely used as the fundamental data for predicting short-term passenger travel demand. The employed models included the autoregressive integrated moving average models [1]–[3], the Kalman filter models [4], the neural network models [5]–[10], the decision trees models [11], and the support vector models [12], [13]. These short-term models in general focus on the prediction of passenger travel demand during a short period (e.g., 15 minutes). The smartcard data were usually split into two parts. The first part of data was used to calibrate or train the models, and the second part of data was used to validate the predictions.

Predicting future passenger travel demand in an expanding subway network is particularly challenging. Because the subway network structure is going to change with new lines and/or line extensions, and most importantly no historical travel demand information is available for the new line or the extended line. Hence, when predicting travel demand for subways with expansion, the socioeconomic data, the transportation environment data, and the travel survey data were usually used as model inputs, and the four-step modeling approach was often employed to build the prediction models [14], [15].

Many socio-economic and transportation environmental factors may affect passenger travel demand in an expanding subway network. Previous works have investigated this issue at the station level. The passenger flows of a subway station were found to be affected by socioeconomic factors such as land use [16]–[18], population [19]–[21], human mobility [22], [23], gross domestic product (GDP), household income, housing prices, and car ownership [21]. The passenger flows of a subway station were also found to be affected by transportation environment factors such as bus service [19],

Authorized licensed use limited to: AALTO UNIVERSITY. Downloaded on May 03,2022 at 08:05:08 UTC from IEEE Xplore. Restrictions apply.
accessibility [20], road segment length [18] and walking environment [24]. Moreover, the identified socioeconomic and transportation environment factors have been used to develop passenger flow prediction models for subway stations [19], [20], [25]. Many of these developed models are based on simple multiple regression analysis.

The investigations of the models for predicting travel demand in expanding subways are very limited. Existing methods are in general based on the four-step travel demand forecasting approach. For example, Yang [15] used the daily travel survey data and the four-step model to predict the average daily origin-destination (OD) matrix of Harbin Metro Line 1. Similarly, Zhao et al. [14] followed the four-step model by calculating the attraction index of station and the growth factor of passenger flow to predict the passenger travel demand of Beijing Metro Line 4. However, the predicted passenger travel demand was not validated directly; instead only the transfer passenger flow and line passenger flow were validated. Moreover, the four-step models adopted in previous studies were usually based on travel survey data, which require not only a significant amount of time but also a considerable cost to collect. In this study, we use the expanding Shenzhen Metro as a case study. We develop a two-step model (K-means and XGBoost combined machine-learning model, termed as KXGBoost) that only employs the online available data to achieve reliable prediction for the future passenger travel demand of a subway network with expansion. The used online data include demographic data (e.g., population distribution), POI data (e.g., number of shopping services around a station), transit infrastructure data (e.g., number of exits of a station), and transit service data (e.g., train departure time interval). In the proposed prediction model, the determinant features of passenger travel demand are first identified using the recursive feature elimination (RFE) algorithm. Based on the identified features, 13,806 OD pairs (118 stations times 117 stations) are split into a number of clusters using the K-means algorithm, and the optimal number of clusters is determined by the mean silhouette coefficient. Next, an XGBoost prediction model for predicting the passenger travel demand is trained for each cluster of OD pairs, and the optimized set of parameter values are obtained. Finally, when predicting the future passenger travel demand, each OD pair is first clustered into an appropriate cluster, and then the passenger travel demand of the OD pair is predicted using the XGBoost model trained for the cluster that the OD pair belongs to.

The smartcard data of passengers were collected for a long period during which a new subway line (Line 11) was put into operations. Hence, the obtained smartcard data recorded the change of passenger travel demand after the change of network structure, and the smartcard data are used to validate the developed travel demand prediction model. The model proposed in this study can be easily extended to other cities because all input data are readily available for most cities. The proposed model also represents a time and cost-efficient alternative to the four-step methods based on the costly and time-consuming travel surveys.

II. DATA

A. Smartcard Data of Passengers

We use the expanding Shenzhen Metro as a case study. In August 2016, there were 6 lines (namely Lines 1, 2, 3, 4, 5, and 11) and 132 stations in Shenzhen Metro (Fig. 1 a). TABLE I shows the general properties of the 6 subway lines, which include the number of stations, the date when the line was put into operations, the date when the line was recently extended, the length of the line, and the operation speed of the line. The smartcard data were provided by the Shenzhen Transportation Authority and recorded the trips of more than 3 million subway passengers in Shenzhen. In the smartcard data, the time, card ID, station ID, gate ID, and transaction type (tap-in or tap-out) were recorded each time a passenger entered or exited a station. We define that a passenger makes a trip from $i$ to $j$ when the passenger enters a station $i$ and exits a station $j$. Passenger travel demand during any specific time window can be estimated using the smartcard data. Firstly, the records missing both the station ID and the gate ID were first filtered out (40 in total). Secondly, the trips started and ended at the same station (567,983 in total) were filtered out. Thirdly, the trips with tap-in or tap-out time beyond the operation hours (587 in total) and the trips lasted for more than 3 hours were filtered out (17,597 in total). Finally, 8636197 subway passenger trips were used for further analysis. In this study, the smartcard data are used to train and validate the travel demand prediction model.

We define the passenger travel demand $T(i, j)$ from station $i$ to station $j$ as the average daily number of passenger trips from $i$ to $j$. The travel demand can be well approximated by a power-law distribution: $p(T(i, j)) = \lambda \cdot T(i, j)^{-\beta}$, where $\lambda = 2.20 \times 10^5$ and $\beta = 2.51$. Specifically, the travel demands of about 80% OD pairs are less than 200 passengers/day, while 1.6% OD pairs have more than 1,000 passengers/day. Besides, the in-flow of a station is defined as the average number of passengers entering the station per day $f_{in}(i) = \sum_{j \neq i} T(i, j)$, and the out-flow of a station is defined as the average number of passengers exiting the station per day $f_{out}(i) = \sum_{j \neq i} T(j, i)$, where $N_{s}$ is the total number of stations.

B. Factors Potentially Influencing Passenger Travel Demand

Passenger travel demand in subway is influenced by various factors, ranging from demographic factors [26] to built-environment factors [27], [28] and transit service factors [29]. Inspired by these studies, a set of station-level and OD-level features are considered to develop the travel demand prediction model for the expanding subway (TABLE II). Different from previous models that mostly relied on time-consuming and costly travel surveys, the data of the selected features are readily available for most cities.

The service area of a subway station is estimated using the catchment area (CA), the radius of which is determined by the maximum distance that most passengers prefer to walk to a station. In this study, the radius of CA is set to 0.5 km, which is a widely used parameter setting in previous studies [30], [31]. To measure the effect of stations...
competing for passengers, the extended catchment area (ECA) of a subway station is defined as a 1 km-radius circular area centered at the station (Fig. 1 a). If station \( j \) is within the ECA of station \( i \), the CAs of \( i \) and \( j \) have an overlapping area, implying that \( j \) may attract passengers within the CA of \( i \).

According to the attributes of the data, we divide the features into three classes, namely, social and economic features, transportation environment features and competition effect feature. These features all could have an influence on travel demand of subway passengers.

1) Social and Economic Features: \( \text{Pop}(i), N_{ci}(i), N_{as}(i), N_{ds}(i), N_{ss}(i), N_{c}(i), N_{hs}(i), N_{fjs}(i), N_{recs}(i), N_{c}(i). \) The population distribution \( \text{Pop}(i) \) data are obtained from the WorldPop project (www.worldpop.org). This population dataset was released in July 2015. In this dataset, the population is estimated for each high-resolution grid with an area of approximately \( 0.1 \times 0.1 \text{ km}^2 \). The population of each grid within the CA of a station is aggregated to estimate the population within the CA.

We make a reasonable assumption that the population distribution is relatively stable in Shenzhen from 2015 to 2016.

POI (Point of Interest) refers to a specific point location on the map. A POI can represent a house, a school, a bus station, or a shopping mall. Each POI is documented with the information of name, coordinate, category, etc. The POI information can reflect the land use patterns of a city, and different POIs have different passenger flow attraction characteristics. The number of POIs within the CA of a subway station is used to quantify the social and economic properties of the station’s service area.

2) Transportation Environment Features: \( N_{bl}(i), N_{bs}(i), N_{se}(i), N_{sl}(i), t_{di}(i), T_Y(i), \langle t(i,j) \rangle \). Bus transit is an important mode of public transportation. The related bus transit features are the number of bus lines connecting a subway station \( N_{bl} \) and the number of bus stops \( N_{bs} \) connecting a subway station. The values of these two features can reveal the public transportation demand at a specific site. The data of the two bus transit features are generally available from the public transport agencies. In this study, the data of \( N_{bl} \) and \( N_{bs} \) are obtained from the official website (https://www.szmc.net) of Shenzhen Metro Group Co. Ltd. The bus lines and bus stops connecting a subway station can also be identified using online maps (e.g., Amap) when no data is released on the

<table>
<thead>
<tr>
<th>Line</th>
<th># of stations</th>
<th>Date put into operations</th>
<th>Date recently extended</th>
<th>Length (km)</th>
<th>Speed (km/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1</td>
<td>30</td>
<td>28 December 2004</td>
<td>15 June 2011</td>
<td>41.04</td>
<td>80</td>
</tr>
<tr>
<td>Line 2</td>
<td>29</td>
<td>28 December 2010</td>
<td>28 June 2011</td>
<td>35.78</td>
<td>80</td>
</tr>
<tr>
<td>Line 3</td>
<td>30</td>
<td>28 December 2010</td>
<td>28 June 2011</td>
<td>41.66</td>
<td>85</td>
</tr>
<tr>
<td>Line 4</td>
<td>15</td>
<td>28 December 2004</td>
<td>16 June 2011</td>
<td>19.96</td>
<td>80</td>
</tr>
<tr>
<td>Line 5</td>
<td>27</td>
<td>22 June 2011</td>
<td>N/A</td>
<td>40.00</td>
<td>80</td>
</tr>
<tr>
<td>Line 11</td>
<td>18</td>
<td>28 June 2016</td>
<td>N/A</td>
<td>51.94</td>
<td>80</td>
</tr>
</tbody>
</table>
The official website of subway. For the stations in Shenzhen Metro, we use the number of lines passing through a station are usually expected to have higher passenger flows. Thus, suburbs or residential areas. Secondly, the transfer stations urban areas usually have more exits than stations in the passenger flows. Big stations and transfer stations in central area of station has been put into operations grow with time [32]. Hence, the number of years that a station passenger flows of new subway stations were observed to as a candidate feature. Thirdly, the average train departure interval at a station \( t_{di} \) determines the maximum number of passengers that a line can transport per unit time. Fourthly, passenger flows of new subway stations were observed to grow with time [32]. Hence, the number of years that a station has been put into operations \( T_Y(i) \) is also considered. Finally, the estimated travel time \( t(i, j) \) from station \( i \) to station \( j \) is also included. The values of this feature are estimated in METHODS section B. The data of all subway features above is also included. The values of this feature are estimated in

<table>
<thead>
<tr>
<th>Feature</th>
<th>Level</th>
<th>Description</th>
<th>Classes</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Pop(i))</td>
<td>Station</td>
<td>The population within the CA of station ( i )</td>
<td></td>
<td>WorldPop (2015)</td>
</tr>
<tr>
<td>(N_{cr}(i))</td>
<td>Station</td>
<td>Number of commercial residences within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{as}(i))</td>
<td>Station</td>
<td>Number of accommodation services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{ds}(i))</td>
<td>Station</td>
<td>Number of domestic services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{ss}(i))</td>
<td>Station</td>
<td>Number of shopping services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_c(i))</td>
<td>Station</td>
<td>Number of corporations within the CA of station ( i )</td>
<td></td>
<td>Society and economy</td>
</tr>
<tr>
<td>(N_{hs}(i))</td>
<td>Station</td>
<td>Number of healthcare services within the CA of station ( i )</td>
<td></td>
<td>Open platform</td>
</tr>
<tr>
<td>(N_{fis}(i))</td>
<td>Station</td>
<td>Number of financial insurance services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{secs}(i))</td>
<td>Station</td>
<td>Number of science, education, and cultural services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{bs}(i))</td>
<td>Station</td>
<td>Number of bus stops connecting station ( i )</td>
<td></td>
<td>Shenzhen Metro Group Co. Ltd. (2016)</td>
</tr>
<tr>
<td>(N_{se}(i))</td>
<td>Station</td>
<td>Number of exits of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{st}(i))</td>
<td>Station</td>
<td>Number of subway lines passing through station ( i )</td>
<td></td>
<td>Transportation environment</td>
</tr>
<tr>
<td>(t_{di}(i))</td>
<td>Station</td>
<td>Train departure interval at station ( i )</td>
<td></td>
<td>Shenzhen Transportation Authority</td>
</tr>
<tr>
<td>(T_Y(i))</td>
<td>Station</td>
<td>Number of years station ( i ) was put into use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(l(i,j))</td>
<td>OD pair</td>
<td>Estimated travel time from station ( i ) to station ( j )</td>
<td>Competition effect</td>
<td>Obtained through calculation</td>
</tr>
<tr>
<td>(\varepsilon(i,j))</td>
<td>OD pair</td>
<td>The efficiency of using station ( i ) to arrive station ( j )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some factors related to subway are also important. The related features are the number of exits of a station \( N_{se}(i) \), the number of lines passing a station \( N_{st}(i) \), train departure interval at a station \( t_{di}(i) \), the number of years that a station has been put into operations \( T_Y(i) \), the estimated travel time \( t(i, j) \) from station \( i \) to station \( j \). Firstly, the number of exits of a station \( N_{se}(i) \) is a good indicator to the volume of passenger flows. Big stations and transfer stations in central urban areas usually have more exits than stations in the suburb or residential areas. Secondly, the transfer stations are usually expected to have higher passenger flows. Thus, we use the number of lines passing through a station \( N_{st}(i) \) as a candidate feature. Thirdly, the average train departure interval at a station \( t_{di}(i) \) determines the maximum number of passengers that a line can transport per unit time. Fourthly, passenger flows of new subway stations were observed to grow with time [32]. Hence, the number of years that a station has been put into operations \( T_Y(i) \) is also considered. Finally, the estimated travel time \( t(i, j) \) from station \( i \) to station \( j \) is also included. The values of this feature are estimated in METHODS section B. The data of all subway features above are available on the official website (https://www.szmc.net) and the Wikipedia pages of Shenzhen Metro Group Co. Ltd. (https://en.wikipedia.org/wiki/Shenzhen_Metro).}

3) **Competition Effect Feature:** \( \varepsilon(i,j) \). Subway stations may compete for passengers when they are near to each other. We define the neighboring stations of a station as \( s'(i) \in ECA(i) \) where \( ECA(i) \) is the extended catchment area of station \( i \). The number of neighboring stations ranges from 1 to 6 for the studied Shenzhen Metro stations, with the number equals to 1 representing station \( i \) is the only station inside its own ECA (i.e., no neighboring station competes for passengers). Next, we use the travel efficiency \( \varepsilon(i,j) \) from station \( i \) to station \( j \) to describe the competition effect. The travel efficiency \( \varepsilon(i,j) \) is defined as:

\[
\varepsilon(i,j) = \min_{s' \in ECA(i)} \{\forall (t(i', j)| s' \in ECA(i))\} \frac{t(i, j)}{t(i', j)}
\]

where \( t(i, j) \) represents the estimated travel time from station \( i \) to station \( j \), \( s' \) represents the neighboring station within the ECA of station \( i \), and \( t(i', j) \) represents the estimated travel time from station \( i \) to station \( j \). The value of efficiency \( \varepsilon(i,j) \) ranges from 0 to 1, and \( \varepsilon(i,j) = 1 \) means that using \( i \) is the most efficient way to travel to \( j \) compared to using \( i \)'s neighboring stations. Using \( XSJ \) to \( DF \) as an example (Fig. 1 a), there are two neighboring stations within the ECA of \( XSJ \), namely \( SSJ \) and \( CL \). According to (1), \( \varepsilon(XSJ, DF) = \frac{t(CL, DF)}{t(XSJ, DF)} = 0.85 \), indicating that using \( XSJ \) is not the most efficient way to travel to \( DF \).

**III. METHODS**

**A. The Framework of the Prediction Model**

The developed model for predicting passenger travel demand in expanding subways mainly includes the following five modules (please also see the flowchart of Fig. 2):

1) The Estimation of Travel Time and Travel Efficiency: The actual travel time between stations in an expanding subway is unknown when the new line has not been put into operations. Hence, we estimate the travel time between each pair of OD

<table>
<thead>
<tr>
<th>Feature</th>
<th>Level</th>
<th>Description</th>
<th>Classes</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Pop(i))</td>
<td>Station</td>
<td>The population within the CA of station ( i )</td>
<td></td>
<td>WorldPop (2015)</td>
</tr>
<tr>
<td>(N_{cr}(i))</td>
<td>Station</td>
<td>Number of commercial residences within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{as}(i))</td>
<td>Station</td>
<td>Number of accommodation services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{ds}(i))</td>
<td>Station</td>
<td>Number of domestic services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{ss}(i))</td>
<td>Station</td>
<td>Number of shopping services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_c(i))</td>
<td>Station</td>
<td>Number of corporations within the CA of station ( i )</td>
<td></td>
<td>Society and economy</td>
</tr>
<tr>
<td>(N_{hs}(i))</td>
<td>Station</td>
<td>Number of healthcare services within the CA of station ( i )</td>
<td></td>
<td>Open platform</td>
</tr>
<tr>
<td>(N_{fis}(i))</td>
<td>Station</td>
<td>Number of financial insurance services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{secs}(i))</td>
<td>Station</td>
<td>Number of science, education, and cultural services within the CA of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{bs}(i))</td>
<td>Station</td>
<td>Number of bus stops connecting station ( i )</td>
<td></td>
<td>Shenzhen Metro Group Co. Ltd. (2016)</td>
</tr>
<tr>
<td>(N_{se}(i))</td>
<td>Station</td>
<td>Number of exits of station ( i )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N_{st}(i))</td>
<td>Station</td>
<td>Number of subway lines passing through station ( i )</td>
<td></td>
<td>Transportation environment</td>
</tr>
<tr>
<td>(t_{di}(i))</td>
<td>Station</td>
<td>Train departure interval at station ( i )</td>
<td></td>
<td>Shenzhen Transportation Authority</td>
</tr>
<tr>
<td>(T_Y(i))</td>
<td>Station</td>
<td>Number of years station ( i ) was put into use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(l(i,j))</td>
<td>OD pair</td>
<td>Estimated travel time from station ( i ) to station ( j )</td>
<td>Competition effect</td>
<td>Obtained through calculation</td>
</tr>
<tr>
<td>(\varepsilon(i,j))</td>
<td>OD pair</td>
<td>The efficiency of using station ( i ) to arrive station ( j )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some factors related to subway are also important. The related features are the number of exits of a station \( N_{se}(i) \), the number of lines passing a station \( N_{st}(i) \), train departure interval at a station \( t_{di}(i) \), the number of years that a station has been put into operations \( T_Y(i) \), the estimated travel time \( t(i, j) \) from station \( i \) to station \( j \). Firstly, the number of exits of a station \( N_{se}(i) \) is a good indicator to the volume of passenger flows. Big stations and transfer stations in central urban areas usually have more exits than stations in the suburb or residential areas. Secondly, the transfer stations are usually expected to have higher passenger flows. Thus, we use the number of lines passing through a station \( N_{st}(i) \) as a candidate feature. Thirdly, the average train departure interval at a station \( t_{di}(i) \) determines the maximum number of passengers that a line can transport per unit time. Fourthly, passenger flows of new subway stations were observed to grow with time [32]. Hence, the number of years that a station has been put into operations \( T_Y(i) \) is also considered. Finally, the estimated travel time \( t(i, j) \) from station \( i \) to station \( j \) is also included. The values of this feature are estimated in METHODS section B. The data of all subway features above are available on the official website (https://www.szmc.net) and the Wikipedia pages of Shenzhen Metro Group Co. Ltd. (https://en.wikipedia.org/wiki/Shenzhen_Metro).
using the Dijkstra shortest path algorithm (see section B for
detail). Using the estimated travel time, we calculate the travel
efficiency \( \varepsilon(i, j) \) of each OD pair.

2) Data Pre-Processing: To ensure the values of each feature
are within the same scale, we use the z-score method [33]
to normalize each feature, with mean value of 0 and variance
of 1. The passenger flows of different OD pairs vary widely,
which may affect the performance of machine learning model.
Hence, we use the logarithmic passenger flow instead of the
actual passenger flow in the model training process.

3) Feature Selection: In order to avoid model overfitting, we use the recursive feature elimination (RFE)
algorithm [34] combined with cross-validation to identify
the determinant features of passenger travel demand. Here,
the RFECV of the sklearn library is used to identify
the features (https://scikit-learn.org/stable/modules/generated/
sklearn.feature_selection.RFECV.html).

4) K-Means Clustering: Using the data samples similar to
the predictive object to train a machine learning model
may achieve better performance of prediction [35]. In this study,
we cluster the OD pairs using the K-means algorithm and train
the passenger travel demand prediction model for each cluster
of OD pairs respectively. The mean silhouette coefficient [36]
is employed to determine the optimized number of clusters.

5) XGBoost Model: An XGBoost model is trained for each
cluster of OD pairs. When predicting future passenger travel
demand, each OD pair is first clustered into an appropriate
cluster, and then the passenger travel demand of the OD pair
is predicted using the XGBoost model trained for the cluster
that the OD pair belongs to. Finally, the predicted logarithmic
passenger flow is converted to its normal scale using an
exponential function.

B. The Estimation of Travel Time and Travel Efficiency

To estimate the travel time between each pair of OD,
we regard each subway station as a node and each subway
section as a link. In order to take the transfer time into account,
we split each transfer station into several virtual nodes that
belong to different lines passing through the transfer station.
Each pair of virtual nodes at a transfer station are connected
by a virtual link, and the transfer time is the weight of the
virtual link. The transfer time between each pair of \( n \) virtual
nodes of a transfer station is set to the same value.
Si et al. [37] proposed that transfer requires additional
efforts. Thus, the perceived transfer time is estimated by
multiplying the expected transfer time with the magnification
factor \( a \):

\[
\bar{E}_r = aE_r \tag{2}
\]

where \( \bar{E}_r \) is the perceived transfer time at station \( r \), \( E_r \) is the
officially estimated transfer time at station \( r \), which is obtained
from the official website (https://www.szmc.net) of Shenzhen
Metro Group Co. Ltd. Consequently, the perceived travel time
is calculated using (3):

\[
\langle t(i, j) \rangle = \sum_{s}^{n} t_s + \sum_{r}^{m} \bar{E}_r \tag{3}
\]

where \( \langle t(i, j) \rangle \) is the perceived travel time based on path
from station \( i \) to station \( j \), \( t_s \) is the officially estimated travel
time of subway section \( s \) obtained from the official website
(https://www.szmc.net) of Shenzhen Metro Group Co. Ltd.
We assume that passengers select the path with the shortest
perceived travel time. Consequently, we assign each passenger
trip to the subway network, obtaining the route and the
estimated travel time of each trip:

\[
\langle t(i, j) \rangle = \sum_{s}^{n} t_s + \sum_{r}^{m} E_r \tag{4}
\]

where \( \langle t(i, j) \rangle \) is the estimated travel time from station \( i \) to
station \( j \), \( t_s \) is the officially estimated travel time of subway
section \( s \), \( E_r \) is the officially estimated transfer time at transfer
station \( r \).
We use the traversal method to determine the appropriate
value of \( a \). To achieve this, we compare the estimated travel
time with the average actual travel time obtained from the
smartcard data (only using the data collected on working days in April 2016), and select the value of \( \alpha \) that minimizes the root mean square error (RMSE). Using the selected value of \( \alpha \), we estimate the travel time and travel efficiency using (4) and (1) for each OD pair in the expanding subway. At the same time, we estimate the travel time and travel efficiency for each OD pair in the original subway. The travel time feature and the travel efficiency feature are used to train the passenger travel demand prediction model.

C. K-Means Clustering Algorithm

The K-means algorithm is a widely used clustering method. Here, the identified determinant features are used to generate the feature space, within which the feature vector of each OD pair is a data sample. The OD pairs are clustered according to their distances in the feature space. The clustering of OD pairs mainly includes the following steps:

- **Step 1:** Randomly select K samples \( \{\mu_1, \mu_2, \ldots, \mu_k\} \) from the OD pairs as the initial cluster centers. Each cluster is denoted as \( C_i \), where \( i \in \{1, 2, \ldots, k\} \).
- **Step 2:** Calculate the Euclidean distance between a data sample and each cluster center, and assign the data sample to the cluster with the smallest Euclidean distance.
- **Step 3:** Update the cluster center \( \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \) where \( |C_i| \) is the number of OD pairs assigned to the \( i \)th cluster. Next, repeat step 2 and step 3 until the maximum number of iterations is reached or the convergence condition is satisfied.
- **Step 4:** Repeat steps 1-3 using different K values. Use the mean silhouette coefficient to determine the optimized value of K. Here, the value of K is tested from 2 to 30 (i.e., K = 2, 3, 4, \ldots, 30).

D. XGBoost Model

The XGBoost model is a machine learning model based on tree integration and characterized with the features of fast speed and good robustness [38], [39]. The basic idea of the XGBoost model is to combine several tree models with low accuracy into one integrated model with high accuracy, which is usually employed in the prediction tasks [40], [41].

We combine K-means and XGBoost to develop a passenger travel demand prediction model for expanding subways. We take advantage of K-means clustering to train the XGBoost model for each cluster of OD pairs. In addition, five-fold cross-validation is used to optimize the model parameters, which include the learning rate, the number of trees, and the maximum tree depth [42]. The root mean square error (RMSE) of the predicted passenger flows is measured, and the parameters that minimize the average RMSE of the five-fold cross-validation are selected as the optimized set of parameters.

When predicting passenger travel demand, each OD pair is first clustered into an appropriate cluster, and then the passenger travel demand of the OD pair is predicted using the XGBoost model trained for the cluster that the OD pair belongs to.

IV. RESULTS

A. Description of the Experiments

We use the Shenzhen Metro as a case study. The Shenzhen Metro Line 11 was put into operations on June 28, 2016. The passenger travel demand prediction model is developed and applied as follows:

Firstly, the passenger travel time data on working days in April 2016 (inferred from the smart card data) are employed to determine the value of transfer magnification factor \( \alpha \). In order to ensure the reliability of the travel time data, we only use the OD pairs with sufficient number of passenger trips. Therefore, we estimate the required minimum number of passenger trips in April, \( n_{\text{min}} \), using the method proposed by Cochran [43]:

\[
 n_{\text{min}} = \frac{z^2 \cdot p(1-p)}{e^2}
\]  

where \( z^2 \) is the selected critical value of the desired level of confidence, \( p \) is the estimated proportion of an attribute in the population or the maximum variability of the population, and \( e \) is the desired level of precision or margin of error. Here, we set \( z = 1.96 \) (95% confidence level), \( p = 50\% \), \( e = 5\% \) [44], and obtained \( n_{\text{min}} = 384.16 \). That means the passenger volumes of the selected OD pairs should at least reach 384.16 passengers. We find the qualified OD pairs representing 74% of all OD pairs, and only the qualified OD pairs are used to determine the magnification factor \( \alpha \). As the magnification factor \( \alpha \) increases from 0.2 to 5 with a tolerance of 0.2, we find the root mean square error (RMSE) of the estimated travel time reaches the minimum (392.5) when \( \alpha \) equals to 3.6 (Fig. 3 a). Therefore, the transfer magnification factor \( \alpha \) is set to 3.6, implying the perceived transfer time is about 3-4 times of the actual transfer time. Then, we estimate the travel time and travel efficiency of each OD pair in the original subway and the expanding subway.

Secondly, each feature is normalized using the z-score method, and the passenger flow of each OD pair in the original subway network (before Line 11 was put into operations) is converted to its logarithmic value using the \( \log_{10} \) function. For the OD pairs with no passenger flow, the logarithmic transformation cannot be applied. To solve this, we set the logarithmic passenger flow of these OD pairs to -10, indicating that the passenger flow of these OD pairs is very tiny and approaches zero.

Thirdly, the multiple linear regression model is used in RFE algorithm to identify the determinant features. Six determinant features are identified, and five-fold cross-validation is used to determine the feature importance. The importance of a feature is defined as the increase in the root mean square error (RMSE) when the feature is not used for prediction. Sort by importance, the six determinant features are the estimated travel time \( t(i, j) \), the number of catering services within the CA of the origin station \( N_{cs}(i) \), the number of catering services within the CA of the destination station \( N_{cs}(j) \), the number of bus lines connecting the destination station \( N_{bl}(j) \), the number of bus lines connecting the origin station \( N_{bl}(i) \), and the travel efficiency \( e(i, j) \). It is not surprisingly that \( t(i, j) \) has an important impact on passenger travel demand,
because travel time is always one of the most important factors determining passengers’ path choice behavior [37], and travel time also serves as the core parameter in many travel demand prediction models [8], [45]. Interestingly, the number of catering services around a subway station has a great impact on passenger flow. This may suggest that catering industry may promote passenger mobility and induce larger passenger travel demand. For example, Laojie station is surrounded by various catering services (Fig. 1 c), its in-flow \( f_{in} \) and out-flow \( f_{out} \) exceeded 37,000 passengers/day. Another interesting finding is that the connection between subway lines and bus lines, which is manifested in \( N_{bl}(i) \) and \( N_{bl}(j) \), also plays an important role. That means in a place with many bus passengers the travel demand of subway passengers is also large. For example, Shenzhen University station is connected by many bus lines (Fig. 1 d), its in-flow \( f_{in} \) and out-flow \( f_{out} \) exceeded 42,000 passengers/day. Finally, \( e(i, j) \) has less impact on passenger travel demand than the other five determinant features.

Fourthly, we cluster the OD pairs using the K-means algorithm. The mean silhouette coefficient reaches the maximum when the value of K is 3, which suggests the optimized number of clusters is 3 (Fig. 3 b). The first cluster contains 4,691 OD pairs, the second cluster contains 942 OD pairs, and the third cluster contains 8,173 OD pairs (a total of 13,806 OD pairs). Fig. 4 shows the distributions of the six determinant features for each cluster of OD pairs. The origin and destination for most OD pairs in cluster 2 are within a short distance (Fig. 4 e). In addition, the origins of these OD pairs are mostly distributed in the downtown areas, where many bus lines and catering services are located nearby (Fig. 4 a & c). In contrast, the origin and destination for a considerable fraction of OD pairs in cluster 1 and cluster 3 are far from each other (Fig. 4 e), and the origin stations are usually located in the peripheral area (Fig. 4 a & c). However, OD pairs in cluster 1 are characterized with larger \( N_{bl}(j) \) and \( N_{cs}(j) \) compared with OD pairs in cluster 3 (Fig. 4 b & d). This suggests that the destinations of cluster 1 OD pairs are primarily located at the downtown area, whereas the destinations of cluster 3 OD pairs are primarily located at the peripheral area. Therefore, \( N_{bl}(i) \), \( N_{cs}(i) \), \( t(i, j) \), and \( e(i, j) \) can be used as the representative features of cluster 2 OD pairs, whereas \( N_{bl}(j) \) and \( N_{cs}(j) \) can be used as the representative features of cluster 1 and cluster 3 OD pairs. Meanwhile, five-fold cross-validation is used to identify the optimized values of the learning rate, the number of trees, and the maximum tree depth of the XGBoost model. Following Ogunleye & Wang [46] and Farrugia et al. [42], we set the range of the learning rate to 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, and 0.3, the range of the number of trees to 100, 125, 150, 175, 200, 225, 250, 275, and 300, the range of the maximum tree depth to 2, 3, 4, 5, 6, 7, and 8. All possible combinations of these parameter values are tested and the optimized set of parameter values are used for training the XGBoost model for each cluster of OD pairs.

Finally, identify the cluster of each OD pair in the expanding subway (with Line 11 being put into operations), and apply the XGBoost model trained for the cluster to predict the passenger travel demand of the OD pair. Among them, 5,411 OD pairs belong to the first cluster, 1,071 OD pairs belong to the second cluster, and 10,810 OD pairs belong to the third cluster (a total of 17,292 OD pairs). The predicted logarithmic passenger flow is converted to its normal scale using the \( exp2 \) function.

We use three error indicators to evaluate the prediction accuracy of the developed model: the mean absolute error (MAE), the root mean square error (RMSE), and the mean absolute percentage error (MAPE).

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (6)
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad (7)
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \quad (8)
\]
where $y = \{y_1, y_2, \ldots, y_l, \ldots, y_n\}$ represents the actual average daily number of passenger trips of each OD pair $T(i, j)$, $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_l, \ldots, \hat{y}_n\}$ represents the predicted average daily number of passenger trips of each OD pair $\hat{T}(i, j)$, and $n$ is the number of OD pairs in the expanding subway (Line 11 has been put into operations).

**B. Results Analysis**

We compare the predicted passenger flow of each OD pair in the expanding subway (with Line 11 having been put into operations) with the actual passenger flow value (Fig. 5). The results indicate that passenger travel demand can be predicted using the developed model with a pretty high accuracy (Fig. 5, TABLE III). The predicted passenger flows are highly correlated with their actual values (Pearson correlation coefficient PCC = 0.96). In addition, the in-flow $f_{in}(i) = \sum_{j \neq i} \hat{T}(i, j)$ and the out-flow $f_{out}(i) = \sum_{j \neq i} \hat{T}(j, i)$ of a subway station $i$ can be derived from the predicted travel demand $\hat{T}(i, j)$ (i.e., the average daily number of passenger trips from station $i$ to station $j$). We find both in-flow and out-flow of a station can be well predicted using the developed model (Fig. 6).

In order to show the advantage of the proposed prediction model, several machine learning models are used for comparisons, which include extreme gradient boosting (XGBoost), multilayer perceptual neural network (NN), K nearest neighbor (KNN), and support vector machine (SVM) (TABLE III). We use passenger flow data collected in the original network (Line 11 had not been put into operations) to train the comparative models. Five-fold cross-validation is used to determine the optimized values of the parameters of each model. For extreme gradient boosting (XGBoost), the optimized parameter values are the same as the proposed model. For multilayer perceptual neural network (NN), Heaton [47] suggests that the optimal number of nodes in the hidden layer is usually between the size of the input layer and the size of the output layer, and a hidden layer is sufficient to solve most problems. Therefore, we use a three-layer multi-layer perceptual neural network (NN), where the number of hidden layer neurons is selected from 1 to 6. For K nearest neighbor (KNN), we follow the suggestion of Lall & Sharma [48] that the number of nearest neighbor samples K can be selected from 1, 5, 10, 20, and the square root of the sample size. For support vector machine (SVM), we follow the suggestion of Chen & Li [49] that the range of penalty coefficient is $2^{-5}, 2^{-4}, 2^{-3}, \ldots$, and $2^{12}$ and the range of gamma is $2^{-12}, 2^{-11}, 2^{-10}, \ldots$, and $2^5$. As shown in TABLE III, the proposed prediction model (XGBoost) shows the best performance in predicting passenger travel demand in the studied expanding subway. Finally, the prediction accuracy for each cluster of OD pairs is shown in TABLE IV.

Although passenger flows, in general, can be accurately predicted for stations in our study, we still observe that the in/out passenger flows of a few stations are underestimated or overestimated. For example, the passenger flows of Chegongmiao station of Line 11 are underestimated. One possible reason is that the station is located near a commercial street, where a lot of tourists visit. The passenger flows of some stations are overestimated, for example, the actual passenger flows of the North Airport station of Line 11 are smaller than the predicted passenger flows. One possible explanation is that the terminal T4 of Shenzhen airport had not been put into operations in April 2016, and the proposed model does not incorporate such detailed information. In future studies, incorporating more specific features to capture the special mobility properties of a region and random variation.
of travel demand in special events [50] can further enhance the prediction accuracy.

V. Conclusion

In this study, we propose a two-step model (KXGBoost) to predict passenger travel demand in an expanding subway with a new line being put into operations. We discover six determinant features that have great impact on passenger travel demand, and use the identified features to cluster the OD pairs. The XGBoost model trained for each cluster of OD pairs is applied to predict passenger travel demand. We find the prediction accuracy of the developed model is higher than other comparison models. The predicted passenger travel demand shows great consistency with the actual passenger travel demand. In addition, both in-flow and out-flow of subway stations can be accurately predicted. Finally, the identified determinant features on passenger travel demand can provide insights for better planning and operations of a subway network with expansion.

The prediction is based on open data, which can be easily obtained. The wide availability of the data used in the model makes the model easily transferable to many other cities in the world. The prediction accuracy of the model can be further enhanced by improving the quality of the input data. Another way to improve the prediction accuracy is to incorporate more features that can capture specific mobility properties of a region and random variation of travel demand in special events. The determination of the CA of a station can be further improved by considering the phenomenon that a station’s passenger attraction ability decreases as the walking distance increases. The XGBoost model is employed to generate the passenger travel demand prediction model because the XGBoost model has many good properties such as strong generalization ability, fast parallel calculation, and accurate loss function. In future studies, other classes of machine-learning models can also be tested in the proposed model for the prediction of future travel demand of a subway network with expansion.

Acknowledgment

Kaipeng Wang would like to thank Bao Guo, Hu Yang, and Minglun Li for providing valuable discussions.

References


