Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro

Liyang Tang, Yang Zhao, Javier Cabrera, Jian Ma, and Kwok Leung Tsui

***Abstract*— Forecasting short-term traffic flow has been a critical topic in transportation research for decades, which aims to facilitate dynamic traffic control proactively by monitoring the present traffic and foreseeing its immediate future. In this paper, we focus on forecasting short-term passenger flow at subway stations by utilizing the data collected through an automatic fare collection (AFC) system along with various external factors, where passenger flow refers to the volume of arrivals at stations during a given period of time. Along this line, we propose a data-driven three-stage framework for short-term passenger flow forecasting, consisting of traffic data profiling, feature extraction, and predictive modeling. We investigate the effect of temporal and spatial features as well as external weather influence on passenger flow forecasting. Various forecasting models, including the time series model auto-regressive integrated moving average, linear regression, and support vector regression, are employed for evaluating the performance of the proposed framework. Moreover, using a real data set collected from the Shenzhen AFC system, we conduct extensive experiments for methods validation, feature evaluation, and data resolution demonstration.**

***Index Terms*— Short-term passenger flow forecasting, multivariate linear regression, SVR (support vector regression), feature extraction, time series.**

Manuscript received October 17, 2017; revised May 8, 2018; accepted October 15, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0800100, in part by the National Natural Science Foundation of China under Grant 11471275 and Grant 71473207, in part by the Research Grants Council Theme-based Research Scheme under Grant T32-101/15-R, in part by the Major Research Program of the National Natural Science Foundation of China under Grant 91546103, and in part by the Anhui Provincial Natural Science Foundation under Grant 1708085QG162. The Associate Editor for this paper was Y. Chen. *(Corresponding author: Yang Zhao.)*

L. Tang is with the Key Laboratory of Public Safety Emergency Information Technology of Anhui Province, 38th Research Institute, China Electronic Technology Group Corporation, Beijing 230000, China, and also with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: liyatang@cityu.edu.hk).

Y. Zhao is with the Centre for Systems Informatics Engineering, City University of Hong Kong, Hong Kong (e-mail: yangzhao9-c@my.cityu.edu.hk).

J. Cabrera is with the Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ 07102 USA (e-mail: xavier.cabrera@gmail.com).

J. Ma is with the Department of Safety Engineering, School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610000, China (e-mail: majian@mail.ustc.edu.cn).

K. L. Tsui is with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: kltsui@cityu.edu.hk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. Forecasting short-term passenger flow: an empirical study on Shenzhen metro. The total size of the file is 5.2 MB.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2879497

I. Introduction

With the emergence of intelligent transportation systems in recent decades, forecasting short-term traffic flow—predicting traffic condition in the immediate future in a quantitative manner [1]—has been recognized as a critical topic in transportation research [2]. Indeed, accurate short-term traffic forecasting could facilitate dynamic traffic control proactively by monitoring the present traffic and foreseeing its immediate future. For example, motivations and benefits include alleviating road traffic congestion [3], [4], informing travelers or drivers about traffic condition [5], [6], and providing real-time traffic monitoring and management [7], [8].

From the perspective of data acquisition in transportation domain, existing efforts in short-term traffic flow forecasting fall in: a) sensor-based trace data in road transportation; and b) Automatic Fare Collection (AFC) data in public transit systems. In road transportation (e.g. highway, freeway and motorway), trace data is generated by detectors or sensors such as Global Positioning System (GPS), WiFi; and traffic flow is defined as the number of vehicles passing through a road section per time unit [9]. For example, Asakura *et al.* [10] studied incident detection through traffic flow data collected by probe vehicles equipped with on-board GPS equipment; Xu *et al.* [11] tried to predict travel time with vehicle trajectory data in an express road section; and Fernandez-Ares *et al.* [12] proposed to collect real-time vehicles movement information using Bluetooth signals. Methods for short-term traffic flow forecasting range from parametric to nonparametric models, and univariate to multivariate algorithms. Extensive literatures on methods for short-term traffic flow forecasting in road transportation could be found in [13] and [14].

In contrast, AFC in public transit is initially designed for fare collection instead of traffic monitoring, and thus captures the entrances or exits of station gates. Specifically in the context of metro transportation, traffic flow refers to the volume of arrivals at metro stations during a given period of time, also termed as *passenger flow*. Some research works have been done on explanatory data analysis using AFC data [15]. For example, Liu *et al.* [16] provided preliminary visualization using Shenzhen smart card data to understand urban mobility pattern; Zhong *et al.* [17] performed a comparative study on variability in regularity in the urban mobility patterns among London, Singapore and Beijing using AFC data; Briand *et al.* [18] tried to regroup passengers based on their continuous temporal activities using

AFC data; Ma and Wang [19] developed a data-driven online transit performance monitoring platform combining AFC and Automated Vehicle Location (AVL) together to monitor transit network performance.

In practice, forecasting short-term traffic flow in metro transportation is even more challenging, since metro traffic flow is significantly influenced by the heterogeneity and uncertainty of individual traveling behavior whereas the AFC data does not reflect traffic condition immediately. There exist some efforts on short-term passenger flow forecasting using AFC data. For instance, Leng *et al.* [20] proposed a metro-net oriented probability tree method based on Origin and Destination (OD) information to predict passenger flow; Sun *et al.* [21] constructed a hybrid method of wavelet and Support Vector Machines (SVM) to predict Beijing subway passenger flow especially in morning and evening peak hours; Ding *et al.* [22] proposed a hybrid method of AutoRegressive Integrated Moving Average (ARIMA) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) modeling to improve subway short-term passenger flow forecasting; etc.

However, there remain two major limitations in existing efforts along the line of short-term passenger flow forecasting in metro transportation. First, most research works heavily depend on temporal characteristics in historical metro passenger flow; however, there exist many other influence factors such as spatial features and weather effect which could have an impact on metro traveling behavior as well. Second, data resolution including forecasting step and forecasting horizon is another significant concern in forecasting performance.

On one hand, *forecasting step* refers to the granularity of data aggregation in modeling stage, also termed as analysis interval. As suggested in 1985 Highway Capacity Manual (HCM 1985) [23], at least 15-min intervals should be used in traffic flow data analysis. Indeed, different analysis intervals have been employed in previous research, such as 5-min [24]–[26], 10-min [27], 15-min [21], [22], [28] and 30-min [29], without explicit explanation and validation. In fact, the more detailed granularity is employed in data aggregation for traffic flow computation, the more real-time information is captured and responded; but meanwhile the traffic condition could be more likely to be unstable during such a short period [30]. On the other hand, *forecasting horizon* denotes the length of predicted future during the forecasting stage, i.e. single-step or multi-step ahead. Above existing efforts mainly focus on single-step ahead $(t+1)$ forecasting. However, few research works have fully examined the impact of data resolution including forecasting step and horizon on predicted results.

In this paper, we focus on short-term passenger flow forecasting using AFC data in metro transportation. Along this line, we comprehensively investigate the effect of three types of features, and propose a data-driven forecasting framework. Main contributions of this paper are summarized as follows:

- Wepropose a data-driven framework for short-term traffic flow forecasting, consisting of three stages: a) traffic data profiling—preparing traffic flow data from AFC data; b) feature extraction—extracting temporal, spatial and external features; and c) predictive modeling—constructing predictive models with extracted features. Proposed data-driven framework is not limited to metro transportation, but can be extended to other forecasting applications.
- We comprehensively investigate three types of features, including: a) temporal features from passenger flow time series data, b) spatial features based on Origin-Destination (OD) pattern upon metro network and passenger traveling network, and also c) external weather factor to evaluate rain effect on metro passenger flow.
- We evaluate the performance of the proposed forecasting framework through extensively comparative experiments in terms of: a) model comparison, including time series model ARIMA, multi-variate linear regression and Support Vector Regression (SVR); b) feature evaluation, including spatiotemporal and environmental features; and c) forecasting step and horizon.
- With empirical demonstration of the proposed forecasting framework using Shenzhen Metro AFC data, key findings we obtained include: a) OD based spatial features indeed improve the accuracy of short-term passenger flow forecasting, which suggests that metro traffic flow is not only related to the passenger volume at certain station itself, but also concerned with the traffic condition of the whole transit network; b) external environmental factor such as weather condition contributes to short-term passenger flow forecasting, indicating that heterogeneous data integration improves traffic forecasting; and c) smaller forecasting step contributes to satisfactory prediction for a longer future; otherwise, larger forecasting step performs well for $t+1$ prediction, but falls off linearly when forecasting horizon grows.

The remainder of this paper is organized as follows. Section II formulates the problem and proposes the general data-driven framework for short-term traffic flow forecasting. Section III demonstrates feature extraction with Shenzhen Metro AFC data. Then extensive experiments are conducted in Section IV. Finally, Section V concludes this work.

II. Preliminary and Methodology

A. Data Description

The dataset was extracted from the AFC system of Shenzhen Metro Cooperation[1] in China over 48 days from Oct 14 to Nov 30 in 2013, consisting of more than 140 millions transactions in total. As shown in Figure 1, the daily number of transactions falls off drastically on Nov 12 out of a sudden because of Typhoon Haiyan. To eliminate the effect of extreme cases and focus on the general pattern of passenger flow, we exclude Nov 12. By 2013, there are 5 metro lines and 118 stations in Shenzhen. Table I lists the Shenzhen Metro AFC data fields used in this paper.

B. Notations and Problem Statement

A smart card transaction can be described as a quadruple $(cid, s, t, flag)$, where cid denotes the cardholder ($card\_id$),

[1]Shenzhen Metro Cooperation: http://www.szmc.net/

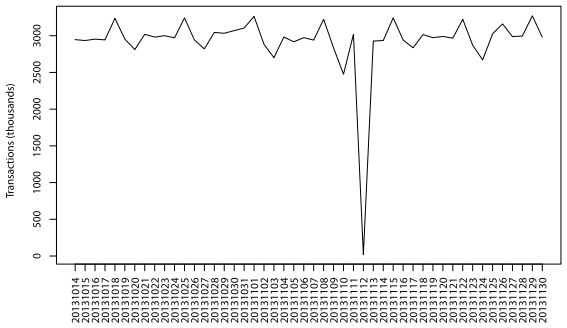


Fig. 1. Daily transactions of all stations in Shenzhen Metro AFC data.

TABLE I
LIST OF AFC DATA FIELDS

Field	Description
transaction_id	Identifying a transaction
card_id	Identifying a passenger
line_id	Identifying a metro line
station_name	Name of metro station
transaction_type	Indicate either in or out of station
transaction_timestamp	Datetime Timestamp of transaction

s denotes station (*station_name*), t denotes the timestamp of transaction (*transaction_timestamp*), and $flag$ denotes either in or out of s (*transaction_type*). Given observations at current t at station s, let $x_s^{(t)}$ be the observed volume (i.e. the volume of passenger arrivals at s) during t-th time interval. The objective is to predict $x_s^{(t+1)}$ during next time slot.

Given a m-size sequence of observed passenger volume at s up to t, define *traffic data profile* of station s as $X_s^{(t)} = \{x_s^{(t)}, x_s^{(t-1)}, x_s^{(t-2)}, \ldots, x_s^{(t-m)}\}$, where m denotes the length of observed period, and $x_s^{(t)}$ is the sum of transaction count at s during t-th time slot.

Intuitively, $x_s^{(t)}$ not only depends on the temporal pattern of traffic data profile of s (i.e. previous observations in $X_s^{(t)}$), but also closely relates to other stations destined for s, since traffic flow could be caused by passengers coming from other stations who have not left s (i.e. no tap-out transaction generated) yet. Suppose metro network is $G = (S, E)$, where $S = \{s_1, s_2, \ldots, s_N\}$ is the set of stations, and N is the number of stations. $(s_i, s_j) \in E$ denotes the OD (Origin-Destination) path from station s_i to s_j, where s_i is termed as *origin* station, and s_j is *destination* station. The assumption here is that passengers traveling from origin station o contribute to $x_s^{(t)}$ if OD path (o, s) exists. That is, if $o, s \in S$ and $(o, s) \in E$, then $x_s^{(t)}$ is related to $x_o^{(t)}$.

C. Overview of Forecasting Framework

Given target station $s \in S$, the proposed data-driven framework for forecasting traffic flow $x_s^{(t+1)}$ at the next time period can be described as following (as shown in Figure 2):

Stage 1: Traffic data profiling. Construct traffic data profile $X_s^{(t)}$ from smart card data, where $x_s^{(t)}$ is aggregated number of arrivals within t-th time slot.

Stage 2: Feature extraction. Features in transportation data could be itemized into three categories:

- Temporal features: Include daily pattern, weekly pattern and autocorrelation evidence of $X_s^{(t)}$;

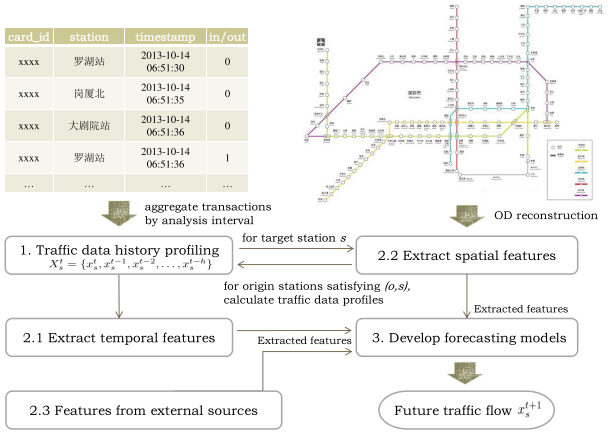


Fig. 2. General data-driven framework for short-term traffic flow forecasting, consisting of traffic data profiling, feature extraction, and predictive modeling.

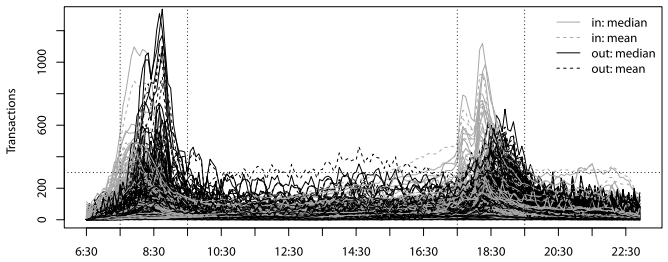


Fig. 3. Transactions of all stations within each time slot: each black (tap-out) or gray (tap-in) line denotes the median (solid) or mean (dotted) over the volume of one station during 47 days $*$ 198 time slots.

- Spatial features: To extract spatial features, we first reconstruct OD with destination as station s and extract associated time and distance characteristics; then for origin station o satisfying $(o, s) \in E$, we calculate traffic data profile $X_o^{(t_o)}$, where t_o is determined by the time spent for OD path (o, s).
- Relevant external features from other data sources: Here we introduce weather data as external feature. Indeed this stage provides support for multi-source data fusion.

Stage 3: Predictive modeling. Build predictive models with extracted features, where model selection and parameter tuning are involved. In this paper, several forecasting models, including ARIMA, multi-variant linear regression and SVR, are embedded in the proposed framework. The performance of the proposed method is evaluated in terms of varying features, forecasting steps and forecasting horizons.

III. FEATURE EXTRACTION

In this section, we comprehensively explore temporal, spatial as well as weather features for short-term passenger flow forecasting, i.e. Stage 2 of the proposed framework.

A. Temporal Features

1) Time of Day: According to the operation of Shenzhen metro system in 2013, the mutual opening hours of all 5 lines are 6:30-23:00. Divide this period by 5-min aggregation interval, we get $16.5\text{h} * 60/5 = 198$ time slots per day.

TABLE II
TIME PERIOD SEGMENTS

Time segment	Slot index range	Description
6:30-7:30	[79,91)	Non-rush hours
7:30-9:30	[91,115)	Rush hours
9:30-17:30	[115,211)	Non-rush hours
17:30-19:30	[211,235)	Rush hours
19:30-23:00	[235,277)	Non-rush hours

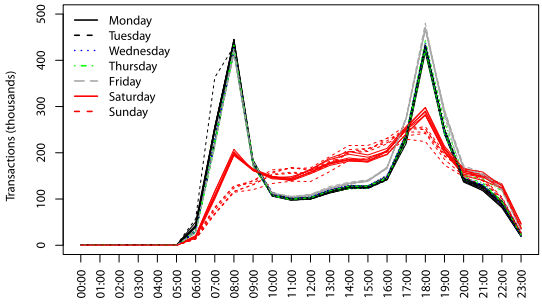


Fig. 4. Hourly transactions of weekdays of 118 stations during 47 days. Each point is aggregated number of transactions of 118 stations within one hour interval per day, each line represents hourly aggregated trend of one single day, and red lines denote weekends.

Figure 3 shows the distribution of transactions for all stations within each time slot, where each data point is aggregated by 5-min interval per day. The black line denotes the median (solid) or mean (dotted) over tap-out volume of one station during 47 days, while the gray line refers to the tap-in volume. We can observe that the transactions exhibit daily pattern with morning peak and evening peak, where the rush hours are estimated as 7:30-9:30 and 17:30-19:30 respectively. In order to differentiate rush hours and non-rush hours, we divide the daily 198 time slots into five segments, as shown in Table II, where slot index is calculated by $hour * 12 + int(minute/5) + 1$.

2) Day of Week: The investigated dataset covers 47 days from Oct 14 (Monday) to Nov 30 (Saturday) exclusive of Typhoon day (Nov 12, Tuesday). We thus have 7 Mondays, Wednesdays, Thursdays, Fridays and Saturdays, but 6 Tuesdays and Sundays.

Figure 4 plots hourly transactions aggregated over all stations during the whole investigated period, where each line represents hourly transactions of one single day, and red lines denote weekends. We observe that passenger flow exhibits weekly pattern: 1) passenger flow is significantly different between workdays (Mon-Fri) and weekend (Sat-Sun); 2) there exists slight fluctuation over days for different workdays.

3) Autocorrelation: We also investigate autocorrelation of traffic data profile $X_s^{(t)} = \{x_s^{(t)}, x_s^{(t-1)}, x_s^{(t-2)}, \ldots, x_s^{(t-m)}\}$. Intuitively, passenger flow at current t is correlated to its own history. Specifically, autocorrelation is calculated between $X_s^{(t)}$ and previous traffic data profile up to l time slots earlier, i.e. $X_s^{(t-l)}$, using Pearson correlation coefficient [31]. Since autocorrelation of passenger flow time series decreases over time, here we use first two lags where the autocorrelation coefficient is larger than 0.9 (please refer to S3 in the supplementary document).

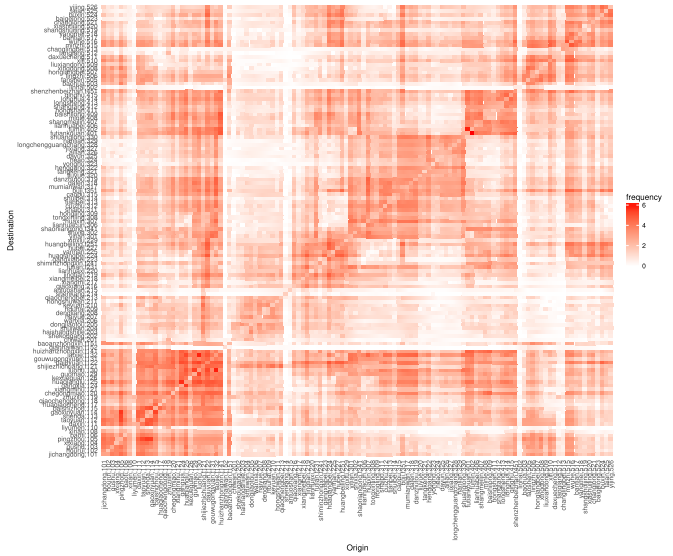


Fig. 5. Heat map of OD matrix during one day, where the frequency is transformed by $log(1 + 0.1 * x_s^{(t)})$. The order of stations are arranged by lines, and the intensity of color denotes the frequency of OD path.

B. Spatial Features

It is believed that time series data could be decomposed into three components: seasonal, trend and noise [32]. That is, given observed passenger flow data X, we have $X = S+T+R$, where S, T, R denote seasonal, trend and noise components respectively (refer to S1 in the supplementary document). Typically, forecasting models with temporal features is able to capture S, but not sufficient for explaining T and R, contributing to inaccurate prediction results. The possible explanation is that S relates to the temporal history with inherent time, day and week effects, while T and R might be non-temporal. Here we assume *propagation effect*—passengers at one station could be disseminated to other stations over metro network through riding on trains—would be a possible reason of T and R.

To deal with the propagation effect on passenger flow, we extract spatial features describing passenger traveling patterns by aggregating OD information throughout the metro network. The intuition behind is that passenger flow of origin stations contributes to the arrivals at destination stations. Therefore, traffic data profile of target station $X_s^{(t)}$ is concerned with that of origin station $X_o^{(t_o)}$ if $(o, s) \in E$.

As mentioned earlier, $(o, d) \in E$ denotes OD path from origin o to destination d. One single OD path (o, d) could be produced by multiple passengers. For each individual passenger i, a one-way trip from o to d is called *OD trip*.

For the sake of narrative and understanding, instead of Chinese station names, we define identifiers for stations according to the following rules: (1) non-transfer stations are composed of 3 digits, where the first digit denotes the line number, and the rest 2 digits denote the sequential number of station; (2) transfer stations begin with character t followed by 3 digits, where the first 2 digits denote the intersection of two lines, and the last digit means the sequential number of intersections between them. For example, *215* means the 15th station of Line 2, and *t122* means the transfer station is the second intersection of lines 1 and 2. In this way, the encoded identifiers indicate the line and station information literally.

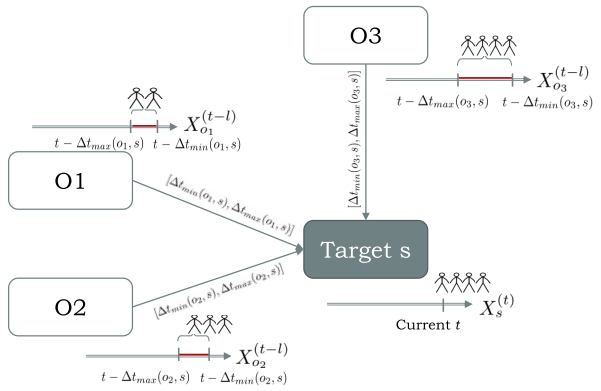


Fig. 6. Illustration example of Algorithm 1.

Take one single day (Oct 14) as an example. There are over 1.3 millions OD trips, and more than 13 thousands OD paths. Figure 5 illustrates the heat map of OD matrix of all stations, where the frequency of OD path is transformed by $log(1+0.1*x_s^{(t)})$. The axis labels consist of both station name and station identifier, x-axis denotes origin station, and y-axis denotes destination station. We can observe that Line 1 is the hottest line, and (t132, 130), (130, t132) are the most frequent OD paths on Oct 14, which are indeed related to the border between Shenzhen and Hong Kong.

Recall that passenger flow refers to the volume of arrivals at s during a given period of time, and $X_o^{(t_o)}$ contributes to $X_s^{(t)}$ if $(o,s) \in E$. Specifically, passengers from origin o at time t_o might be dispersed as the arrivals at station s at t, and $t_o < t$. Therefore, the key is to determine origin station o and time t_o for constructing $X_o^{(t_o)}$.

Algorithm 1 elaborates the procedure of extracting spatial features given station s and current time t. First, find top K frequent origin stations $\{O|o \in O\}$ satisfying $(o,s) \in E$ during historical days (Lines 1-11). Note that frequency calculation of top origins differentiates rush hours and non-rush hours during different time segments. Second, for each OD path (o,s), calculate the trip time distribution of individual trips, and the range of t_o, where $t_o = t - l_o$ and l_o is the time difference between stations s and o determined by trip time of (o,s) (Lines 13-14). The intuition behind is that the trip time of individuals is not a constant number, but depends on walking time, waiting time, and train scheduling, etc. Note that OD path is directed and the duration of OD trips differs among individuals (please refer to S2 in the supplementary document). Specifically, here we use the Interquartile Range (IQR), i.e. the 25th and 75th percentiles of the OD duration distribution, as the lower and upper bounds of trip time spent from origin to destination, notated as $\Delta t_{min}(o,s)$ and $\Delta t_{max}(o,s)$ respectively (Line 14). Last, calculate traffic data profiles for o, i.e. $X_o^{(t-l_o)}$ (Lines 15-17). Ultimately we extract $\sum_{j=1}^{K} l_{o_j}$ features for $X_o^{(t-l_o)}$.

Figure 6 illustrates an example for further understanding, where the horizontal line with an arrow denotes timeline of traffic data profile $X_s^{(t)}$. Suppose o_1, o_2, o_3 are top 3 origin stations to s. The distribution of time spent from o_j to s is calculated for each single trip, and thus the maximum and minimum time durations are calculated. Then, for each origin station o_j, the length of historical traffic data profile (i.e. l) is determined by $[t - \Delta t_{max}(o_j,s)]$ and $[t - \Delta t_{min}(o_j,s)]$, as highlighted red in the timeline. Accordingly, the highlighted traffic profiles of origins are used as features for forecasting passenger flow of target s. An intuitive interpretation is that passengers during the highlighted time period at o might be traveling to s considering the trip time and time slot difference, and therefore history traffic data profile $X_o^{(t-l_o)}$ affects $x_s^{(t)}$. Thus $X_o^{(t-l_o)}$ is associated with metro network and OD pattern.

TABLE III
EXTRACTED FEATURES FOR STATION s

Feature	Type	Description
time_of_day	*range[1,198]*	time slot index within a day
day_of_week	*range[1,7]*	day of week, i.e. Mon-Sun
$X_s^{(t-l)}$	*vector*	history traffic data with time lag l
rain	*0,1*	rain or no rain
$X_o^{(t-l_o)}$	*matrix*	history traffic data of origin stations

C. External Feature

Another possible explanation of T and R components could be external environmental factors. Besides temporal and spatial features directly extracted from AFC data, relevant features from secondary data through multiple external sources have great potential in improving passenger flow forecasting. Indeed it is the essence of data-driven approach. Here we expand feature set by including history weather information from *www.tianqi.com*, which indicates rain or not by date.

In summary, extracted features are listed in Table III, where the first three kinds of features are temporal, and last one is spatial feature. In fact, $X_o^{(t-l_o)}$ is the temporal traffic data weighted by the frequency of OD path (o,s), where o indicates spatial relationship and $(t - l_o)$ refers to temporal history. In this way, we fuse temporal and spatial patterns together.

IV. EXPERIMENTS

In this section, we empirically present the performance of different forecasting models with extracted features using Shenzhen Metro AFC data described in Section II-A.

A. Forecasting Models and Metrics

In this work, we employ three typical forecasting methods: ARIMA, linear regression and SVR in Stage 3 of the proposed forecasting framework. As the most common used time-series forecasting model, ARIMA employs history observations to describe unobserved variables [33], that is:

$$x_s^{(t)} = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i x_s^{(t-i)} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \tag{1}$$

where $\sum_{i=1}^{p} \phi_i x_s^{(t-i)}$ describes auto regression, $\sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$ denotes moving average, p is number of autoregressive terms, q is number of lagged forecasting errors, ϕ is the slope coefficient, θ is moving average parameter, and ε denotes the residuals. Parameters are determined through maximum likelihood estimation, and p, q are selected based on autocorrelation function (ACF) and partial autocorrelation function (PACF).

Algorithm 1 Extracting Spatial Features From Top-K Origins

Input:
Target station s;
Current time t;
Concerned number of other stations K.
Output:
Set of other stations $\{O|o \in O\}$, and $|O| = K$;
Lagged passenger flow of o, $X_o^{(t-l_o)}$.

```
Find all stations destined for s, notated as S;
for each historical day do
  for each time segment do
    for each station j ∈ S do
      Calculate frequency of OD path (j, s), notated as n_j;
    end for
  end for
  Sort n_j (j = 1, 2, ..., N) in descending order;
end for
Determine the segment of current time t;
Select top K stations with higher n_j for all days at t, notated as O;
for each station o ∈ O do
  Estimate trip time of OD path (o, s);
  Calculate Δt_min(o, s) and Δt_max(o, s) for lower and upper bounds respectively;
  for l_o in range(Δt_min(o, s), Δt_max(o, s)) do
    Calculate lagged passenger flow of station o: X_o^(t-l_o) = {x_o^(t-l_o), x_o^((t-l_o)-1), x_o^((t-l_o)-2), ..., x_o^((t-l_o)-m)}.
  end for
end for
```

Linear regression predicts independent variable $x_s^{(t)}$ using one or more predictor variables (i.e. features in Table III). The formula could be generalized as:

$$x_s^{(t)} = \beta_0 + \sum_{j=1}^{d} \beta_j \mathbf{X}_j + \epsilon, \tag{2}$$

where β_0 is the intercept, $\mathbf{X}$ denotes input feature vector, β_j is the slope, d is the number of features, and ϵ is the error term.

Basically, SVR uses the same principles as the SVM for classification, but is with an alternative loss function [34]. SVR can be formulated as a minimization problem as following:

$$\begin{aligned} \min\ & \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{s.t. } & y_i - f(\mathbf{X}_i, \omega) \le \varepsilon + \xi_i^* \\ & f(\mathbf{X}_i, \omega) - y_i \le \varepsilon + \xi_i \\ & \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{3}$$

where $\mathbf{X}$ is the input feature vector, $f(\mathbf{X})$ is the ε-insensitive loss function defined as $f(\mathbf{X}) = \sum(\alpha_i - \alpha_i^*)K(\mathbf{X}_i, \mathbf{X})$, $0 \le \alpha_i, \alpha_i^* \le C$, K is the kernel function, C is a constant of regularization term in Lagrange formulation, and ξ_i, ξ_i^* are slack variables. In this work, RBF kernel is used [35], and parameters are learned by parameter tuning function *tune* using grid search [36].

In summary, the forecasting models with different extracted features (Table III) are notated as following:

- ARIMA: Best model ARIMA(5,0,3) selected by *auto.arima* function using time series data $X_s^{(t)}$.
- LR-T: Linear regression model with (T)emporal features, including *time_of_day*, *day_of_week* and $X_s^{(t-l)}$. Note that interaction effect between *time_of_day* and *day_of_week* is considered throughout the experiments.
- LR-TR: Linear regression with (T)emporal features *time_of_day*, *day_of_week*, $X_s^{(t-l)}$ and (R)ain effect *rain*.
- LR-TRS: Linear regression with (T)emporal features *time_of_day*, *day_of_week*, $X_s^{(t-l)}$, (R)ain effect *rain*, as well as (S)patial features $X_o^{(t-l_o)}$, and K refers to the number of origin stations (Algorithm 1).
- SVR-T: SVR model with (T)emporal features, including *time_of_day*, *day_of_week* and $X_s^{(t-l)}$.
- SVR-TR: SVR with (T)emporal features *time_of_day*, *day_of_week*, $X_s^{(t-l)}$ and (R)ain effect *rain*.
- SVR-TRS: SVR with (T)emporal features *time_of_day*, *day_of_week*, $X_s^{(t-l)}$, (R)ain effect *rain*, as well as (S)patial features $X_o^{(t-l_o)}$, and K refers to the number of origin stations (Algorithm 1).

The metrics for performance evaluation include: (1) RMSE (Root mean squared error), for measuring forecasting errors in the same units; (2) RRSE (Root relative squared error), for comparing forecasting errors in different units; and (3) R-squared score, for measuring the explanatory power of models by calculating the correlation coefficient between observations and predictions, defined as following [37]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{4}$$

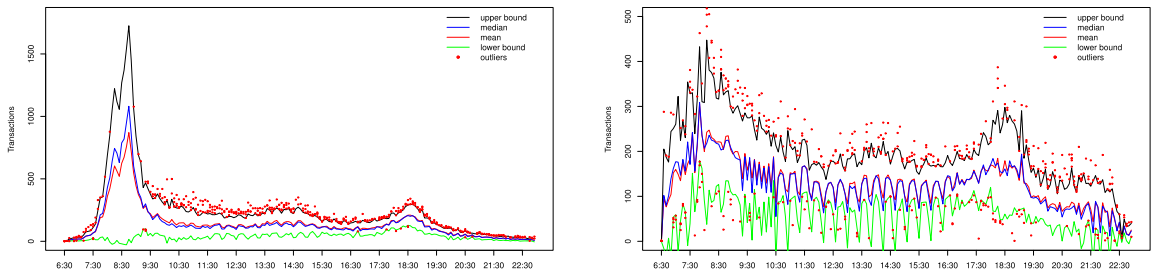
$$RRSE = \sqrt{\sum(\hat{y}_i - y_i)^2 / \sum(\bar{y}_i - y_i)^2} \tag{5}$$

$$R^2 = 1 - \left(\sum(\hat{y}_i - y_i)^2 / \sum(\bar{y}_i - y_i)^2\right) \tag{6}$$

where n is the number of samples, y_i, $\hat{y}_i$ are the observation and predicted value respectively, and $\bar{y}_i$ is the mean of observed values.

B. Forecasting Results

Figure 7 shows the distribution of arrival transactions within each time slot, where each data point is aggregated by 5-min interval per day. The percentile lines are calculated over investigated 47 days (Oct 14-Nov 30), upper and lower bounds are 95% of the area under the normal distribution lies within 1.96 standard deviations of the mean, and red dots are the outlier days with abnormal number of transactions beyond upper and lower bounds. Indeed transaction distribution exhibit different patterns among different stations: as in Figure 7(a), transaction distribution at *huizhanzhongxin* station exhibits obvious temporal pattern, while in Figure 7(b), transaction distribution at *futiankouan* appears to be more random and noisy. Here we take *futiankouan* for case study, since the influence factors of passenger flow might be more complicated besides temporal time series pattern. Throughout the experi-



(a) (b)

Fig. 7. Transactions of arrivals (i.e. tap-out volume) within each time slot, where each data point is aggregated by 5-min over investigated 47 days. (a) *huizhanzhongxin* station. (b) *futiankouan* station.

ments, former 40 days are used for model training, and latter 7 days are used for forecasting evaluation.

Table IV shows the forecasting errors of different models with various extracted features. We use RMSE for absolute error comparison as presented in bars, and RRSE for relative error evaluation as plotted as lines. Generally, we conclude that nonparametric regression model SVR outperforms LR and ARIMA, which is consistent with previous studies as well [1].

First, SVR outperforms ARIMA(5,0,3) significantly (up to 38.0% improvement in RMSE and up to 37.8% improvement in RRSE), and similarly LR performs better than ARIMA(5,0,3). Since SVR and LR leverage multivariate features for predictive learning while ARIMA(5,0,3) uses temporal evidence only, we can conclude that extracted features are validated to be effective for prediction. Moreover, when leveraging the same combination of features, SVR is superior to LR, and SVR-TRS shows the best performance (up to 18.9% improvement in RMSE and up to 16.6% improvement in RRSE), indicating the features are non-linearly correlated and thus nonlinear modeling is preferred based on this data set.

Second, although forecasting errors during rush hours are greater than non-rush hours because passenger volumes experience more variation during rush hours, the improvement of SVR-TRS during rush hours is better than non-rush hours. Specifically, during rush hours, compared to ARIMA(5,0,3), SVR-TRS reduces forecasting error by up to 36.0% in RMSE and 30.8% in RRSE; whereas compared to LR, forecasting error drops by up to 18.9% in RMSE and 11.0% in RRSE. This suggests that leveraging spatial features helps to foresee immediate passenger flow accurately especially during rush hours. In fact, SVR-TRS robustly captures patterns for both rush hours and non-rush hours.

Third, for SVR-TRS model, number of stations K in spatial features is evaluated in Figure 8(a), where the forecasting performs best when $K = 50$ or $K = 60$ approximately. We can observe that when K is too small or too large, it contributes to slight improvement. The possible reasons are as following. When K is too small, only the hottest stations are selected for predicting passenger flow at target station, and the set of hottest stations is usually unchanged whatever target station is. Otherwise, when K is too large, the majority of stations are used for forecasting, which introduces too much noise for model training. It indicates that spatial features extracted from origin stations should be with a moderate size to avoid overfitting or underfitting. Later we use SVR-TRS with $K = 50$ for further evaluation experiments.

TABLE IV
RESULTS OF PERFORMANCE COMPARISON OF DIFFERENT FORECASTING MODELS

Model	Rush hours		Non-rush hours		Overall	
	RMSE	RRSE	RMSE	RRSE	RMSE	RRSE
ARIMA	55.5745	0.6698	39.0188	0.5806	43.5918	0.5986
LR-T	43.8898	0.5086	29.1781	0.4128	30.7459	0.4698
LR-TR	41.1482	0.5127	28.3201	0.3854	29.9222	0.4626
LR-TRS($_{K=100}$)	41.3255	0.5208	27.1484	0.3807	30.1643	0.4423
LR-TRS($_{K=50}$)	40.3937	0.5022	26.9574	0.3771	29.7409	0.4256
LR-TRS($_{K=20}$)	40.8715	0.5097	26.8651	0.3854	30.8318	0.4308
LR-TRS($_{K=10}$)	40.4820	0.5179	27.5344	0.3893	30.1706	0.4366
SVR-T	39.4160	0.5051	26.6803	0.3806	29.2642	0.4507
SVR-TR	38.9421	0.4827	26.6627	0.3705	29.0535	0.4431
SVR-TRS($_{K=100}$)	38.3188	0.4706	26.2847	0.3783	28.2525	0.4177
SVR-TRS($_{K=50}$)	**35.5884**	**0.4636**	**24.2678**	**0.3609**	**27.0269**	**0.3919**
SVR-TRS($_{K=20}$)	36.4851	0.4738	25.1308	0.3819	29.2239	0.4174
SVR-TRS($_{K=10}$)	38.1036	0.4876	25.6074	0.3909	30.0636	0.4249

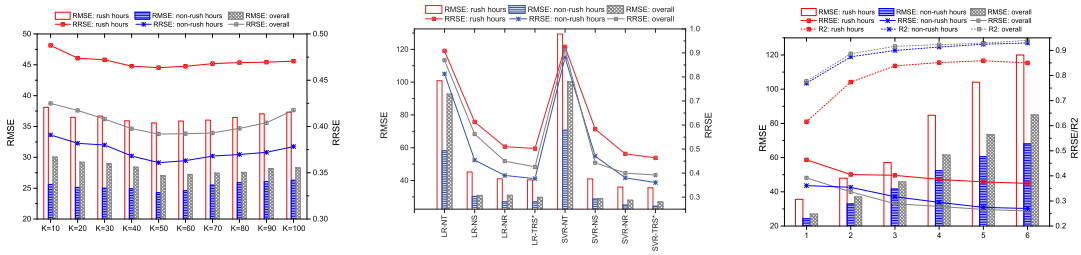
Moreover, we evaluate the effectiveness of each type of feature by eliminating it from -TRS models. Indeed this method has been typically employed in existing feature evaluation efforts [38]. Specifically, we denote -TRS without temporal features as -NT, -TRS without spatial features as -NS, and -TRS without rain effect as -NR. Basically, the better prediction (i.e. the smaller prediction error) we get, the less important the eliminated feature is. As illustrated in Figure 8(b), for both LR and SVR, generally forecasting performance preference is -TRS $\succ$ -NR $\succ$ -NS $\succ$ -NT, where $\succ$ means "better than". That is, the significance of three types of features is T(emporal) $\succ$ S(patial) $\succ$ R(ain).

Note that the rain effect is not significant in this case study. This can be due to the fact that the weather information is aggregated by day instead of by hour. However, we conjecture that rain effect would be improved if more fine-grained weather data is obtained [39], or the competition between metro and other transportation modes is examined [40].

C. Evaluation of Forecasting Step and Forecasting Horizon

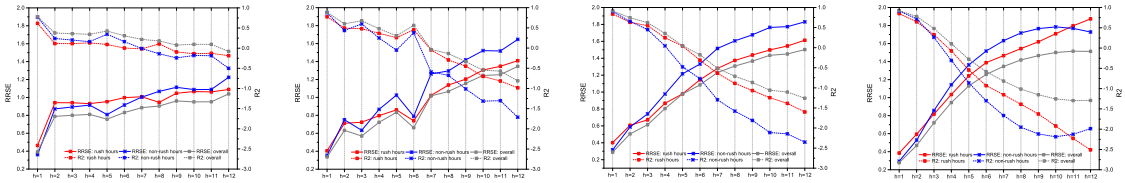
In this section we evaluate the forecasting performance in terms of forecasting step and forecasting horizon.

Forecasting step refers to the granularity of data aggregation, and so far we use 5-min interval. Here we compare the performance with different forecasting steps $\rho = 1, 2, 3, 4, 5, 6$. That is, each $x_s^{(t)}$ is aggregated by time windows as 5-min, 10-min, 15-min, 20-min, 25-min, and 30-min respectively. Since $x_s^{(t)}$ is aggregated by interval ρ, the volume of passengers within each time slot increases when ρ grows, and thus RMSE for different ρ is not comparable because of different measurement units. Accordingly, we also



(a) (b) (c)

Fig. 8. Forecasting results in terms of RMSE (bars) and RRSE (lines) during rush hours (red), non-rush hours (blue) and all together (gray). (a) SVR-TRS with different K. (b) Feature evaluation. (c) SVR-TRS ($K = 50$) with different ρ.



(a) (b) (c) (d)

Fig. 9. Results of SVR-TRS ($K = 50$) with different forecasting steps ρ and forecasting horizons h in terms of RRSE (solid lines) and R2 (dashed lines) during rush hours (red), non-rush hours (blue) and all together (gray). (a) $\rho = 1$. (b) $\rho = 2$. (c) $\rho = 3$. (d) $\rho = 4$.

utilize relative error metric RRSE together with R-squared score R^2 for evaluation.

Figure 8(c) illustrates the forecasting errors of SVR-TRS ($K = 50$) with different ρ. On one hand, absolute metric RMSE increases as ρ grows, because aggregated value $x_s^{(t)}$ becomes larger when the length of t widens. On the other hand, relative metrics RRSE and R2 decrease when ρ grows, and the improvement is not significant after $\rho = 3$ (R2 score is close to 90%). The underlying reason could be that smaller analysis interval increases the fluctuation of traffic flow data stream, while greater interval contributes to more stability.

Forecasting horizon denotes the length of predicted future. In this section we also evaluate the forecasting performance in up to 12-step ahead future, that is, $h = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$, meaning $t + h$ prediction. Figure 9 illustrates the forecasting errors of SVR-TRS ($K = 50$) with different h when $\rho = 1, 2, 3, 4$ respectively. Generally, RRSE ranges from 0 to infinity; the smaller the better, and when RRSE is greater than 1, the forecasting is poor. R2 ranges from negative infinity to 1; the greater the better, and when R2 is negative, the predictive model fits bad. The results of Figure 9 are explained as following:

- In Figure 9(a), when $\rho = 1$ (5-min interval), RRSE remains smaller than 1 and R2 remains positive until $h = 6$ (30-min ahead prediction). Besides, the forecasting performs best when $h = 1$; for $h = 2 \ldots 6$, the forecasting errors increase slightly.
- In Figure 9(b), when $\rho = 2$ (10-min interval), maximum satisfactory forecasting horizon is $h = 4$ (40-min ahead prediction). The performance degrades evidently when ρ grows, but recovers a little bit when $h = 3$ (30-min ahead prediction) and $h = 6$ (60-min ahead prediction).
- In Figure 9(c), when $\rho = 3$ (15-min interval), maximum satisfactory forecasting horizon is $h = 4$ (45-min ahead), and performance degrades linearly when h grows.
- In Figure 9(d), when $\rho = 4$ (20-min interval), maximum satisfactory forecasting horizon is $h = 3$ (60-min ahead), and performance degrades linearly when h grows. However, when $h > 10$, there is a slight increase especially for non-rush hours. The reason behind could be that when ρ and h are large, $X_s^{(t)}$ tends to be stable time series and thus more predictable; and when ρ and h are large enough, the forecasting is no longer short-term.

Therefore, we conclude: (1) short-term passenger flow forecasting could provide fair prediction in immediate future, which is typically up to 60-min ahead; (2) smaller analysis interval contributes to stable and satisfactory forecasting for a longer future; (3) larger analysis interval shows excellent forecasting for next $t + 1$ prediction (i.e. $h = 1$), but performance degrades linearly when forecasting horizon grows.

D. Efficiency Analysis

In fact, our approach consists of an off-line training phase and an online forecasting phase. In the training phase, the major time cost depends on training cost of predictive model used. Taking SVR for example, the computational complexity using SVR with RBF kernel is $\mathcal{O}(n^2 * d)$ [41], where n is the training size and d denotes the number of features. Even though the predictive model could be computationally expensive, it is implemented in the training stage before forecasting. For example, build models with historical data during idle time such as night. To be specific, the average time cost of training SVR-TRS is approximately 4.77 minutes in our data set (40 days * 198 slots training samples).

In the forecasting phase, the computational cost mainly comes from feature extraction. Actually, all features before current time t are extracted during the training phase, and only current t related features are retrieved directly (since it would become previous $t - 1$ when time is sliding) by the time of forecasting. By pushing as many workloads as possible to the

off-line stage, the average time cost is measured in seconds at the time of forecasting. In this way, the forecasting process could be near real time.

V. Discussion and Conclusion

In this work, we have proposed a data-driven framework for short-term passenger flow forecasting, which is composed of three stages: traffic data profiling, feature extraction and predictive modeling. The proposed framework has great potential for solving many other relevant forecasting problems, with specific implementation in feature extraction and model selection. Moreover, by integrating temporal, spatial and weather features together, we have demonstrated that forecasting power could be significantly improved. The empirical conclusions are: (1) nonparametric nonlinear regression model (e.g. SVR) captures volatility characteristics of time series data, and is more flexible for nonlinear and high-dimensional AFC data; (2) other than temporal pattern, OD based spatial features and external weather factor are also effective in improving the accuracy of short-term passenger flow forecasting; (3) smaller analysis interval contributes to stable forecasting for a longer future, whereas larger analysis interval shows excellent forecasting for next $t + 1$ prediction but performance degrades linearly when forecasting horizon grows.

From the perspective of applications, there are three typical scenarios where analyzing and predicting subway passenger flow can provide data support and decision-making insights. (1) Organization and optimization of passenger flow at lines and stations. For example, by accurately forecasting passenger flow dynamics, considering the total number of kilometers and the full load rate of the train, develop optimized operational plans and scheduling strategies to increase the train load rate and reduce operational costs. (2) Help understanding residential travel decisions. Learning future passenger flow from historical traffic flow data at stations by differentiating rush hours and non-rush hours with temporal patterns on a daily or weekly basis, help to understand group traveling behavior of residents. For example, if we could predict extremely large volume at stations during morning peak hours in workdays, more trains should be scheduled. Besides, the surrounding area is likely to be a workplace and the commuting pattern can be further understood. (3) Evacuation design and emergency management in case of congestion. Due to limited space and passenger capacity, when too many passengers flood into the subway station during holidays, commuting peaks, bad weather, sudden incidents or major events and activities, not only the waiting time of passengers would be delayed, but also the safety of passengers might be affected. Thus, forecasting future passenger volume to prepare for congestion and emergency is a significant concern in subway transportation.

Above all, our proposed data-driven framework has presented multi-granularity forecasting capability in terms of aggregated volume in historical observations and predicted length of immediate future. In practical scenarios, proposed framework could provide a multi-angle and multi-granularity view of passenger flow dynamics for roughly or finely prediction. Besides, by integrating various types of features together, the framework exhibits possibility for multi-source heterogeneous data fusion in big data environment. Then, all features are treated as input vectors in machine learning based predictive modeling. In this framework, the exploration and prediction of subway passenger flow are unified, and therefore it provides a comprehensive basis for decision makers.

There are several interesting extensions worth exploring in future works. First of all, we assume propagation effect among transit network, that is, passenger flow at origin contributes to that at destination, and thus we propose to extract spatial features using traffic data profiles from origin stations. The potential limitation behind is that proposed method performs best only when the target station has origin-destination smart card transaction records. In other words, the performance improvement of forecasting passenger flow at transfer stations with incomplete in/out transactions might be restricted, since transfer passenger flow involves passenger flow route assignment issue, which would be our immediate future work.

Additionally, the diversity in individuals travel behavior is not yet considered in this work, since the passenger volume is only concerned with group behavior in an aggregated manner. Indeed, travel behavior analysis and predicting travel purpose, mode and time, should be implemented in individual level instead of aggregated fashion. As an extension of current station-oriented work, future works also include understanding the behavior of individuals from the perspective of passengers.

References

[1] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *J. Transp. Eng.*, vol. 123, no. 4, pp. 261–266, Jul. 1997.

[2] M. Alam, J. Ferreira, and J. Fonseca, "Introduction to intelligent transportation systems," in *Intelligent Transportation Systems*. New York, NY, USA: Springer, 2016, pp. 1–17.

[3] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 557–569, Feb. 2016.

[4] Z. Cao, S. Jiang, J. Zhang, and H. Guo, "A unified framework for vehicle rerouting and traffic light control to reduce traffic congestion," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1958–1973, Jul. 2017.

[5] J. Ma, B. L. Smith, and X. Zhou, "Personalized real-time traffic information provision: Agent-based optimization model and solution framework," *Transp. Res. C, Emerg. Technol.*, vol. 64, pp. 164–182, Mar. 2016.

[6] J. Ma, F. Zhou, and C. Lee, "Providing personalized system optimum traveler information in a congested traffic network with mixed users," *J. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 500–515, 2016.

[7] H. Hashemi and K. F. Abdelghany, "Real-time traffic network state estimation and prediction with decision support capabilities: Application to integrated corridor management," *Transp. Res. C, Emerg. Technol.*, vol. 73, pp. 128–146, Dec. 2016.

[8] S. Faye and C. Chaudet, "Characterizing the topology of an urban wireless sensor network for road traffic management," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5720–5725, Jul. 2016.

[9] W. Leutzbach, *Introduction to the Theory of Traffic Flow*, vol. 47. New York, NY, USA: Springer, 1988.

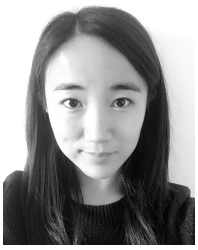
[10] Y. Asakura, T. Kusakabe, L. X. Nguyen, and T. Ushiki, "Incident detection methods using probe vehicles with on-board GPS equipment," *Transp. Res. C, Emerg. Technol.*, vol. 81, pp. 330–341, Aug. 2017.

[11] T. Xu, X. Xu, Y. Hu, and X. Li, "An entropy-based approach for evaluating travel time predictability based on vehicle trajectory data," *Entropy*, vol. 19, no. 4, p. 165, 2017.

[12] A. J. Fernández-Ares, A. M. Mora, S. M. Odeh, P. García-Sánchez, and M. G. Arenas, "Wireless monitoring and tracking system for vehicles: A study case in an urban scenario," *Simul. Model. Pract. Theory*, vol. 73, pp. 22–42, Apr. 2017.

[13] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 3–19, Jun. 2014.

[14] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transp. Rev.*, vol. 24, no. 5, pp. 533–557, Sep. 2004.
[15] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011.
[16] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen, "Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen," in *Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2009, pp. 1–6.
[17] C. Zhong *et al.*, "Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data," *PLoS ONE*, vol. 11, no. 2, p. e0149222, 2016.
[18] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 274–289, Jun. 2017.
[19] X. Ma and Y. Wang, "Development of a data-driven platform for transit performance measures using smart card and GPS data," *J. Transp. Eng.*, vol. 140, no. 12, p. 04014063, 2014.
[20] B. Leng, J. Zeng, Z. Xiong, W. Lv, and Y. Wan, "Probability tree based passenger flow prediction and its application to the Beijing subway system," *Frontiers Comput. Sci.*, vol. 7, no. 2, pp. 195–203, 2013.
[21] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, Oct. 2015.
[22] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1054–1064, Apr. 2018.
[23] Highway Capacity Manual, "Special report 209," 3rd ed., Transp. Res. Board, Washington, DC, USA, Tech. Rep., 1985, p. 985, vol. 1.
[24] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.
[25] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.
[26] Y.-S. Jeong, Y.-J. Byon, M. Mendonca Castro-Neto, and S. M. Easa, "Supervised weighting-online learning algorithm for short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1700–1707, Dec. 2013.
[27] L. Tsirigotis, E. I. Vlahogianni, and M. G. Karlaftis, "Does information on weather affect the performance of short-term traffic forecasting models?" *Int. J. Intell. Transp. Syst. Res.*, vol. 10, no. 1, pp. 1–10, 2012.
[28] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2015.
[29] J. Zhao *et al.*, "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 790–801, Apr. 2017.
[30] J. J. Benedetto and P. J. Ferreira, *Modern Sampling Theory: Mathematics and Applications*. New York, NY, USA: Springer, 2012.
[31] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. New York, NY, USA: Springer, 2009, pp. 1–4.
[32] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *J. Official Statist.*, vol. 6, no. 1, p. 3, 1990.
[33] J. D. Hamilton, *Time Series Analysis*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1994.
[34] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS Tech. Rep.*, vol. 14, no. 1, pp. 5–16, 1998.
[35] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2004.
[36] D. Meyer *et al.*, "Package 'e1071,"' Tech. Rep., 2017.
[37] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2014.
[38] L. Tang, Z. Ni, H. Xiong, and H. Zhu, "Locating targets through mention in Twitter," *World Wide Web*, vol. 18, no. 4, pp. 1019–1049, 2015.
[39] X. Xu, B. Su, X. Zhao, Z. Xu, and Q. Z. Sheng, "Effective traffic flow forecasting using taxi and weather data," in *Proc. 12th Int. Conf. Adv. Data Mining Appl. (ADMA)*, Gold Coast, QLD, Australia, Springer, Dec. 2016, pp. 507–519.
[40] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," *Transp. Res. D, Transport Environ.*, vol. 14, no. 3, pp. 205–221, 2009.
[41] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.



Liyang Tang received the Ph.D. degree in management science and engineering from the Hefei University of Technology in 2015. She is currently a Post-Doctoral Research Fellow at the City University of Hong Kong and an Engineer at the 38th Research Institute, China Electronic Technology Group Corporation. Her research interests include data mining and big data analytics.

Yang Zhao received the bachelor's degree in statistics from the Shandong University of Science and Technology in 2011 and the Ph.D. degree from the City University of Hong Kong in 2015. She is currently the Scientific Officer with the Centre for Systems Informatics Engineering, City University of Hong Kong. Her research interests are in machine learning and statistics, especially their application to real problems.

Javier Cabrera received the Ph.D. degree. He is currently a Professor of Statistics and Biostatistics at Rutgers University. He was the Chief Co-Editor of CSDA and the Director of the Institute of Biostatistics, Rutgers University. He has many publications and books in biostatistics, big data for medical sciences, functional genomics, data mining genomics data, statistical computing and graphics, and computer vision. He is a Fulbright fellow and a Henry Rutgers fellow. He was supported by grants from NSF, the RWJ Foundation, and the Qatar Foundation.

Jian Ma received the B.S. degree in safety engineering from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree in safety technology and engineering from the City University of Hong Kong, Hong Kong, in 2010. From 2011 to 2012, he further performed studies on pedestrian traffic and evacuation dynamics as a Post-Doctoral Fellow with the City University of Hong Kong. He is currently a Professor with the School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China. His research interests include pedestrian traffic, crowd dynamics, emergency evacuation, and intelligent traffic simulation. He is also a Reviewer for international journals, such as the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Kwok Leung Tsui is currently a Chair Professor of industrial engineering with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, and the Founder and the Director of the Center for Systems Informatics Engineering. Prior to joining the City University of Hong Kong, he was a Professor at the School of Industrial and Systems Engineering, Georgia Institute of Technology. His current research interests include data mining, surveillance in healthcare and public health, prognostics and systems health management, calibration and validation of computer models, process control and monitoring, and robust design and Taguchi methods. He is a fellow of the American Statistical Association, the American Society for Quality, and the International Society of Engineering Asset Management; a U.S. representative to the ISO Technical Committee on Statistical Methods. He was a recipient of the National Science Foundation Young Investigator Award. He was the Chair of the INFORMS Section on Quality, Statistics, and Reliability and the Founding Chair of the INFORMS Section on Data Mining.