Transportation Research Part C 77 (2017) 306–328

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks☆

Yang Li [a], Xudong Wang [a], Shuo Sun [b], Xiaolei Ma [c,b,*], Guangquan Lu [c,b]

[a] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
[b] School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China
[c] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China

ARTICLE INFO

*Article history:*
Received 17 February 2016
Received in revised form 18 January 2017
Accepted 6 February 2017
Available online 16 February 2017

*Keywords:*
Matching pursuit orthogonal least squares (MPOLS)
Nonlinear system identification
Radial basis function (RBF) networks
Smart card data
Subway passenger flow prediction
Special events

ABSTRACT

Reliable and accurate short-term subway passenger flow prediction is important for passengers, transit operators, and public agencies. Traditional studies focus on regular demand forecasting and have inherent disadvantages in predicting passenger flows under special events scenarios. These special events may have a disruptive impact on public transportation systems, and should thus be given more attention for proactive management and timely information dissemination. This study proposes a novel multiscale radial basis function (MSRBF) network for forecasting the irregular fluctuation of subway passenger flows. This model is simplified using a matching pursuit orthogonal least squares algorithm through the selection of significant model terms to produce a parsimonious MSRBF model. Combined with transit smart card data, this approach not only exhibits superior predictive performance over prevailing computational intelligence methods for non-regular demand forecasting at least 30 min prior, but also leverages network knowledge to enhance prediction capability and pinpoint vulnerable subway stations for crowd control measures. Three empirical studies with special events in Beijing demonstrate that the proposed algorithm can effectively predict the emergence of passenger flow bursts.



## 1. Introduction

Subway transportation systems are adopted worldwide as effective countermeasures to mitigate the adverse effects of rapid urbanization and traffic congestion (Ma et al., 2015). In the past decades, the construction of urban rail transits in both developed and developing countries has been increasing, and the expanding subway networks have stimulated a sharp rise in ridership (Zhang and Wang, 2014). Taking the Beijing subway system as an example, the average daily ridership in 2012 was 6.74 million passengers compared with the 1.19 million passengers in 2000 (Si et al., 2015). The ridership explosion has triggered a series of issues, such as crowdedness in trains and the insufficient capacity of subway facilities (Zhong et al., 2016). These phenomena became very common in metropolitan cities, including Beijing, Shanghai, and Tokyo (Ma et al., 2012). In particular, passengers in Beijing may not be able to board trains successfully due to overcrowdedness during peak

☆ This article belongs to the Virtual Special Issue on "Smart cards, big data and travel behaviour".

* Corresponding author at: School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China.
*E-mail addresses:* liyang@buaa.edu.cn (Y. Li), wxd_1024@foxmail.com (X. Wang), sunshuo@buaa.edu.cn (S. Sun), xiaolei@buaa.edu.cn (X. Ma), lugq@buaa.edu.cn (G. Lu).

http://dx.doi.org/10.1016/j.trc.2017.02.005
0968-090X/

hours. To alleviate the burden of high passenger demands, passengers are routinely restricted at certain subway stations and have to line up for boarding, or are even prohibited from entering the stations during peak hours. One possible solution to this issue is to provide proactive passenger flow forecasting information for passengers to adjust their travel paths, modes, and departure times. Reliable passenger flow estimation can also benefit transit operators in optimizing service schedules.

Existing literature on subway passenger flow prediction can generally be categorized into long-term and short-term predictions. Long-term passenger flow prediction essentially estimates future travel demands through the four-step transportation planning model or regression techniques (Horowitz, 1984). Regression models are required to establish a relationship between subway ridership and a series of influential factors, such as demographic, economic, and land-use information (Taylor et al., 2009; Chan and Miranda-Moreno, 2013). For short-term passenger flow prediction, statistical and computational intelligence-based models are extensively studied (Karlaftis and Vlahogianni, 2011; Ma et al., 2015; Vlahogianni et al., 2014). Generally, these models can be classified as time series (Ma et al., 2013; Xue et al., 2015), neural network (Tsai et al., 2009), support vector machine (Sun et al., 2015; Wang et al., 2015), and empirical mode decomposition (Jiang et al., 2014). Although these methods can capture the subtle and regular fluctuation of subway passenger flow to some extent, they are not designed to forecast non-regular demands due to special events (Pereira et al., 2015). These special events include concerts, sporting events, parades, and other public events that may attract a large group of people. When these events occur, a significant rise in passenger demand can be observed; such passenger flow irregularities are very difficult to forecast and manage because of poor information availability (Pereira et al., 2015). Determining the spatial and temporal distributions of passenger flows in advance is crucial for emergency response and stampede prevention (Zhong et al., 2015) and is considered a more effective countermeasure than reactive control and evacuation strategies. With reliable passenger flow information, transit operators can adopt several proactive measures to avoid overcrowding, such as bus feeder allocation, warning information dissemination, and temporary closure of stations. However, predicting irregular travel demands under special events is much more difficult than regular demand forecasting, because special events are nonrecurring and leave scant prior information for pattern recognition and machine learning models. Therefore, very limited effort has been made in this domain, with three exceptions. Kuppam et al. (2013) utilized the traditional four-step model to demonstrate the demand for special events. Pereira et al. (2015) developed a neural network algorithm to predict public transit arrivals for special events using social media and smart card data, as well as demonstrated the quality of transport predictions under special events. Ni et al. (2017) developed a hashtag-based event detection algorithm to detect high subway passenger flow based on social media data, and then proposed a parametric and convex optimization-based approach to forecast the passenger flow.

In the context of subway systems, predicting passenger flows under special events deserves more attention, especially in metropolitan cities where a large portion of the inhabitants take the subway for commuting (Zhou et al., 2014). Alternatively, a number of statistical and non-statistical model approaches have been developed, such as autoregressive moving average (ARMA), artificial neural networks (ANNs), support vector regression, linear regression, and radial basis functions (Pereira et al., 2015), to solve the short-term prediction of time series under special events. Among these methods, ANN has particularly gained greater attention due to its outstanding capabilities of capturing the potential nonlinear dynamical properties between input–output data. However, conventional ANNs utilized have issues with overtraining, local optima, and high computational burden (Zhang et al., 2014). The ARMA approach has also attracted substantial attention in forecasting time series to overcome these problems. Recently, some variants of ARMA models were developed to solve the issue of time series prediction, including simplified autoregressive, autoregressive integrated moving average (Khashei et al., 2012), and ARMA (Rojas et al., 2008). However, these models commonly require high model orders to accommodate stochastic variations of forecasting time series. In addition, these data sequence models are essentially linear. Thus, they may not adapt well in revealing the characteristics of the stochastic nature and uncertain nonlinear dynamics of time series data from special events, because most of the real applications are complex and nonlinear in nature (De Gooijer and Hyndman, 2006).

To tackle these problems, we proposed an identification approach for detecting and forecasting subway passenger flows under special events. This approach does not require any prior information regarding the events and only relies on limited observational subway passenger flow data. To achieve this objective, we proposed a novel class of *multiscale radial basis function* (MSRBF) networks to represent the underlying dynamics of the abnormal subway network. Generally, a conventional single scale or kernel width RBF network means that all basis functions have only a common single scale, or each basis function has a single individual scale. By contrast, the new RBF network in this work is employed as a number of multiscale basis functions, where each basis function contains multiple scale kernel widths. The detailed procedure of constructing an MSRBF network is as follows. First, in the MSRBF network, some unsupervised clustering algorithms, such as the fast fuzzy *c*-means clustering approach, are initially adopted to pre-cluster and determine the positions or centers of the basis functions. Second, for each selected center, the associated kernel widths (scales) are determined heuristically, and the selected centers and scales are then restricted to a fixed grid. Finally, an MSRBF network is converted into a model with linear parameters. An effective matching pursuit orthogonal least squares (MPOLS) algorithm (Billings and Wei, 2005), together with a generalized cross-validation (GCV) criterion (Li et al., 2011b, 2011a), is then applied to train the MSRBF network. A parsimonious model that only includes a relatively small number of regressor terms is likewise achieved and further used to predict the abnormal subway passenger flow at least 30 min ahead. Different from the majority of existing literature on passenger flow prediction, one of the advantages of the proposed algorithm is that it is capable of predicting station-level passenger flows by considering network knowledge and of ranking the relative contributions of other subway stations on prediction. Before the sudden rise of passenger flows at a certain station can be observed, there may be several other stations with high fluctuations of

passenger flows. The irregularities of passenger flows for these stations are highly related to the occurrence of passenger flow burst at the targeted station. Thus, they can be used as inputs for prediction. In this case, both spatial and temporal features of passenger flows are considered in our model to enhance the predictive capability under special events scenarios.

The remainder of this paper is organized as follows. Section 2 describes the datasets used in this study. Section 3 presents the proposed approach and the acquired methodology. Section 4 provides the methodology applied for the experimental comparison with other prevailing prediction algorithms. Finally, Section 5 discusses the conclusion and future works.

## 2. Data source

The Beijing subway system comprises 18 lines and 319 stations with 527 km of subway network (Si et al., 2015). Fig. 1 presents the layout of the Beijing subway network. Smart card data are used to generate subway passenger flows. In 2006, the entire subway system in Beijing began to accept the smart card. The fare was initially fixed, but it then switched to distance-based in 2014. In 2014, cardholders who spent more than 100 RMB on subway fare receive credit rebates (up to 27% discount) for the following month. This policy attracts a great number of passengers taking the subway, with daily passenger flows of 9.2786 million. Passengers are required to tap their cards at the gate before entering and exiting the subway station (Long and Thill, 2015), and both boarding and alighting information (e.g., times and stops) are recorded (Ma et al., 2013). For each station, the boarding and alighting passenger flows can be estimated and aggregated every 15 min.

Passenger flow data from July 1, 2012 to October 31, 2012 were used as the dataset in this study. A total of 19 stations were selected as the candidate stations for prediction and are displayed in Fig. 1. The abbreviation of each station is listed in Table 1. These stations are either important transfer hubs or stations with very high passenger demands. Their average daily boarding and alighting passenger flows are calculated in Table 1. Among these stations, WKS, DZM, and HDHZ stations are the three target stations for prediction. The WKS station is adjacent to the MasterCard Center, which is an indoor stadium with a capacity of 18,000 and frequently hosts important sports games and vocal concerts. There was one pop-star concert on July 21, 2012 (Saturday, 7:00 PM). Before the performances began, the crowd flowed out of the WKS station and increased

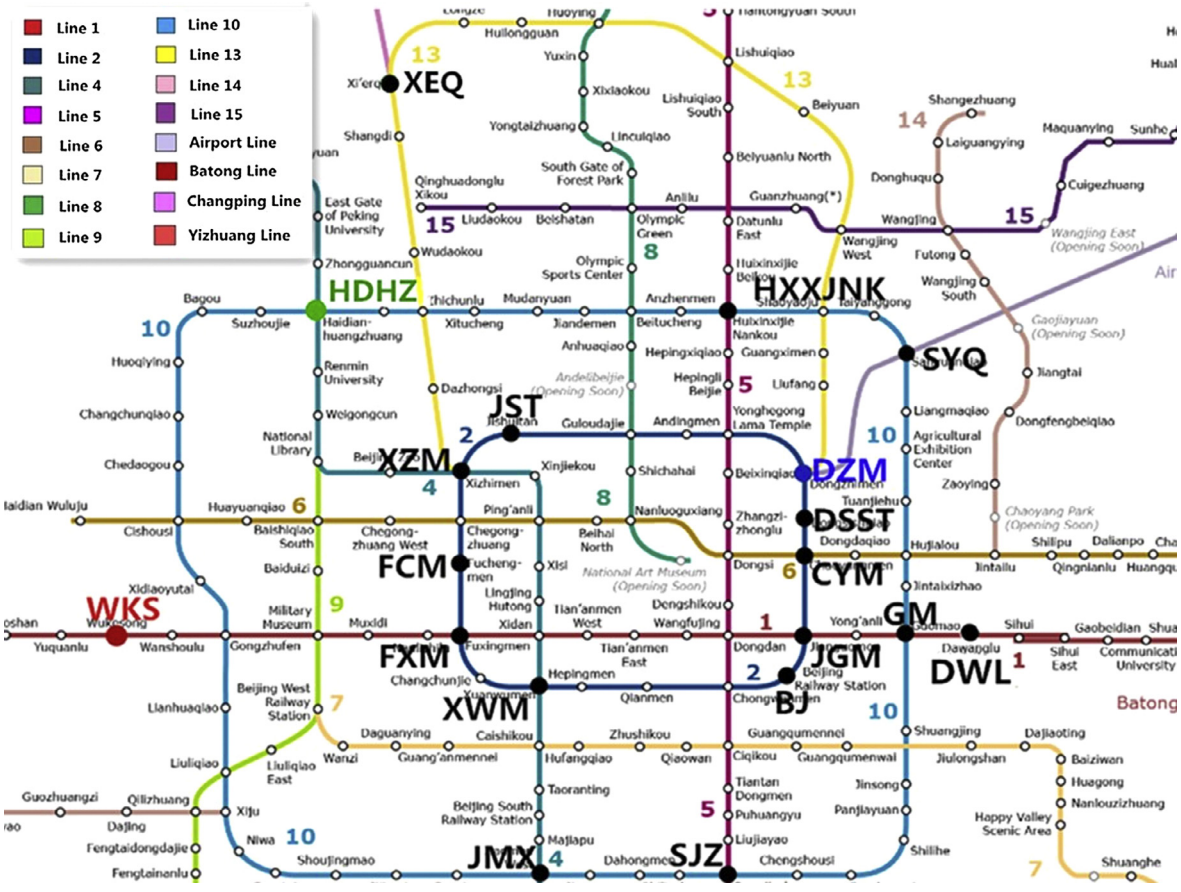


**Fig. 1.** Beijing subway network (The WKS station is marked in red, the DZM station is marked in blue, the HDHZ station is marked in green, and the remaining stations are marked in black.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Average daily boarding and alighting passenger demands for 19 subway stations.

| Station name | Boarding | Alighting |
|---|---|---|
| Beijing Station (BS) | 34,261 | 36,906 |
| Chaoyangmen (CYM) | 34,374 | 35,529 |
| Dawanglu (DWL) | 61,608 | 61,888 |
| Dongsishitiao (DSST) | 28,337 | 29,699 |
| **Dongzhimen (DZM)** | **68,115** | **76,757** |
| Fuchengmen (FCM) | 40,719 | 41,256 |
| Fuxingmen (FXM) | 25,738 | 25,965 |
| Guomao (GM) | 37,648 | 42,513 |
| **Haidianhuangzhuang (HDHZ)** | **16,530** | **24,348** |
| Huixingxijienankou (HXXJNK) | 15,917 | 15,931 |
| Jishuitan (JST) | 42,887 | 41,686 |
| Jianguomen (JGM) | 19,172 | 20,078 |
| Jiaomenxi (JMX) | 15,654 | 15,373 |
| Sanyuanqiao (SYQ) | 21,642 | 26,334 |
| **Wukesong (WKS)** | **22,802** | **19,444** |
| Songjiazhuang (SJZ) | 36,461 | 34,595 |
| Xierqi (XEQ) | 47,962 | 47,204 |
| Xizhimen (XZM) | 61,966 | 59,315 |
| Xuanwumen (XWM) | 22,397 | 21,278 |

the alighting passenger demands as shown in Fig. 2(a). Unlike the WKS station, the DZM station connects an important long-distance bus hub where passengers are transferred to long-distance coaches to other cities in China. This trend became obvious during traditional Chinese holidays, such as Spring Festival and Mid-Autumn Festival. On September 29, 2012 (the day before the Mid-Autumn Festival), a vast number of passengers returned to their hometowns for reunions and were transferred at the DZM station for other surface transportation modes. Compared with other normal Saturdays, the return flow caused a sudden rise in alighting passenger demands for the entire day. Fig. 2(b) reveals that passenger demands increased by more than 40%. Different from the previous two scenarios (i.e., special events held adjacent to the WKS station and DZM station), a subway service disruption caused by a train signal failure occurred at approximately 7:00 PM on August 23, 2012 at the HDHZ station. The HDHZ station is an interchange station connecting lines 4 and 10 and is located in the mixed land-use area surrounded by universities, academic institutes, residence communities, and shopping malls. The disruption stranded a large number of passengers and forced these passengers to change their routes by using buses and taxis, leading to a sudden rise of egressing passenger flows. Given that no prior warning information was released for the emergency evacuation, this can be considered an unplanned event. The events that happened at the WKS and DZM stations were known in advance. Thus, the sudden increase of passenger flows was expected to some extent. However, detecting the exact time when the irregular passenger flows arise and pinpointing the possible stations that generate the most incoming passengers for the target stations are difficult. Compared with planned events, predicting an unplanned surge of passenger flow because of special events is more challenging and useful because these events do not receive sufficient treatment or attention before they actually occur. Moreover, no external information, such as event location, time, and type, is available as prior knowledge to estimate the potential impact.

To forecast the non-regular passenger demands on these three stations, the boarding passenger flows from the remaining 18 stations are considered as possible contributors to the sudden passenger flow increase and incorporated into the prediction model as inputs. The previous time series of alighting passenger flows at the same stations also serve as inputs. The rationale of choosing 18 stations is that the incoming passenger flow at each station ranks in top 18 among 321 subway stations, but these stations contribute approximately 27% of total boarding demands on a daily basis. Incorporating additional stations will significantly increase the complexity of the proposed algorithm. Therefore, a tradeoff between accuracy and efficiency is made. We aim to determine the dynamics of passenger flows between the remaining stations and the target stations and pinpoint the stations that are most vulnerable for special attention because these stations contribute more passenger flows to the WKS, DZM, and HDHZ stations.

## 3. Methodology

### 3.1. The NARX model

To forecast the passenger flow of a target subway station, a multi-input and single-output (MISO) nonlinear system identification and parametric modeling problem is considered. A general form of a discrete-time Non-linear AutoRegressive Moving Average with eXogenous inputs (NARMAX) model is widely applied to construct the MISO system, which can be described by the following nonlinear recursive difference formula (Leontaritis and Billings, 1985; Wei et al., 2004a):

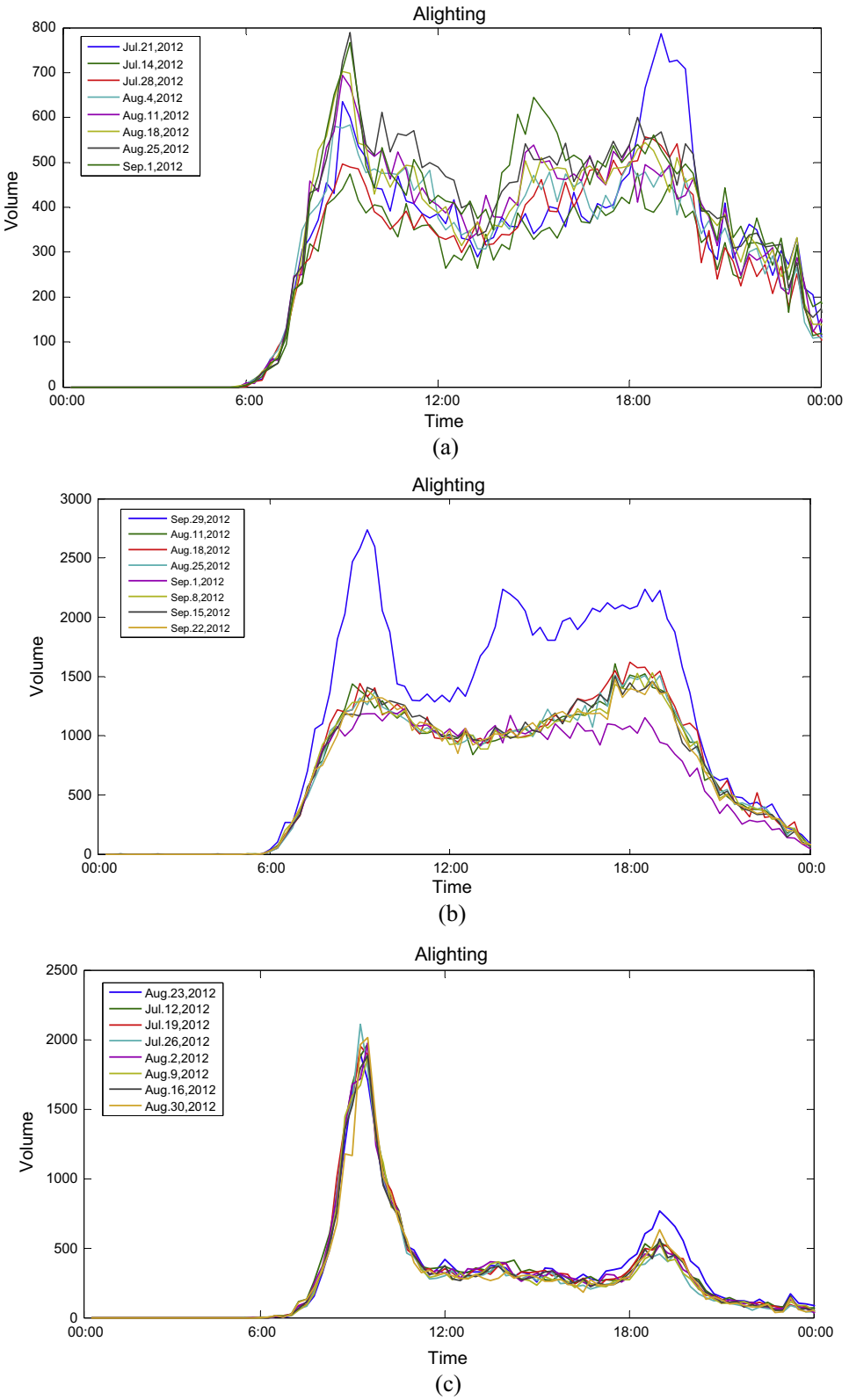


**Fig. 2.** Alighting passenger demand fluctuation at WKS, DZM, and HDHZ subway stations: (a) a concert was held at 7:30 PM on July 21, 2012 (Planned event); (b) passengers returned home before the Mid-Autumn Festival on September 29, 2012 (Planned event); (c) subway service disruption happened on August 23, 2012 (Unplanned event)

$$y(t) = f(y(t-1), \ldots, y(t-n_y), u_i(t-1), \ldots, u_i(t-n_u), \varepsilon(t-1), \ldots, \varepsilon(t-n_\varepsilon)) + \varepsilon(t) \tag{1}$$

where $y(t)$ refers to the alighting passenger flow at the target station at time step $t$, $u_i(t)$ refers to the boarding passenger flows at station $i$ at time step $t$, $\varepsilon(t)$ represents the model error at time step $t$, and $f$ indicates an unknown linear or nonlinear mapping function and usually links the system output to the system inputs between the available inputs and outputs. $n_y$, $n_u$, and $n_\varepsilon$ indicate the maximum time lags in the previous outputs, inputs, and the model error, respectively. The unit time lag in this study is 15 min with the same sampling rate of subway passenger flows. The comprehensive description of the NARMAX methods of dynamical systems is derived from Wei and Billings (Wei et al., 2004a).

For many practical problems, the nonlinear function $f(\cdot)$ in Equation (1) is generally unknown, and the most commonly employed approach to approximate the unknown function $f(\cdot)$ is to adopt a polynomial NARMAX model. To simplify the problem, the NARX model, a subset of the NARMAX model without considering the moving average noise terms in Eq. (1), is applied in our cases. The NARX model can be described in Eq. (2):

$$y(t) = f(y(t-1), \ldots, y(t-n_y), u_i(t-1), \ldots, u_i(t-n_u)) + \varepsilon(t) \tag{2}$$

Define $q = n_y + n_u$ and $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_q(t)]$ with

$$x_k(t) = \begin{cases} y(t-k) & 1 \leqslant k \leqslant n_y \\ u(t-(k-n_y)) & n_y + 1 \leqslant k \leqslant n_y + n_u \end{cases} \tag{3}$$

Thus, the NARX model can be defined in a linear parameter form as follows:

$$y(t) = \hat{f}(\mathbf{x}(t)) + e(t) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) + e(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta} + e(t) \tag{4}$$

where $M$ represents the total number of regressor terms, $\theta_m$ represent the model coefficients, and $\phi_m(\mathbf{x}(t))$ $(m = 1, 2, \ldots, M)$ are the model regressors. $\boldsymbol{\varphi}(t) = [\phi_1(\mathbf{x}(t), \ldots, \phi_M(\mathbf{x}(t)))]^T$ and $\boldsymbol{\theta}$ indicate the regressor matrix and the parameter matrix respectively.

In this study, we will adopt a novel MSRBF network model with Gaussian kernels to approximate nonlinear function $\hat{f}$. The related content will be provided in the next section.

### 3.2. MSRBF network

Generally, the MSRBF network includes both small and large scales in the network in a hierarchical multiscale manner. Thus, it can accommodate both the local and global properties of the basis functions. In the multiscale modeling scheme, multiple kernel widths or different scale parameters are assigned to each basis function. A block diagram of an MSRBF network is shown in Fig. 3. How to construct the network structure and determine the centers and scales in the multiscale modeling framework are discussed in the following section.

#### 3.2.1. Constructing the MSRBF network structure

The MSRBF network of a nonlinear dynamical system is represented as

$$y(t) = \sum_{k=1}^{q} \theta_k^{(linear)} x_k(t) + \sum_{i=0}^{I} \sum_{j=0}^{J} \sum_{m=1}^{Nc} \theta_{i,j,m}^{(RBF)} \varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)}) \tag{5}$$

where $\theta_k^{(linear)}$ and $\theta_{i,j,m}^{(RBF)}$ are the unknown parameters to be estimated, and $\varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)})$ is the $m$th Gaussian basis function that is defined by

$$\varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)}) = \exp\left[ -\sum_{k=1}^{n_y} \left( \frac{x_k(t) - c_{m,k}}{s_{y,m}^{(i)}} \right)^2 - \sum_{k=n_y+1}^{n_y+n_u} \left( \frac{x_k(t) - c_{m,k}}{s_{u,m}^{(j)}} \right)^2 \right] \tag{6}$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_q(t)]$ given in (3) is the network input vector, $\mathbf{c}_m(t) = [c_{m,1}, \ldots, c_{m,q}]$ is the center vector of the $m$th basis function, and $\mathbf{s}_m^{(i,j)}$ is the scale of the $m$th basis function in the network. $\mathbf{s}_m^{(i,j)}$ is represented as

$$\mathbf{s}_m^{(i,j)} = \left[ \overbrace{s_{y,m}^{(i)}, \ldots, s_{y,m}^{(i)}}^{1:n_y}, \overbrace{s_{u,m}^{(j)}, \ldots, s_{u,m}^{(j)}}^{1:n_u} \right] \tag{7}$$

Note that $N_c$ is the number of centers or basis functions in the MSRBF network, and $(I + 1)$ and $(J + 1)$ denote the number of scales for the output and input variables in the $m$th center respectively. Therefore, for a single-input and single-output system, a total of $M = (I + 1)(J + 1)N_c$ basis functions are included in the network. Particularly, for a MISO system, the number of basis functions for the input variables is $(J + 1)^n$, where $n$ is the number of input variables. Thus, the network contains a total of $M = (I + 1)(J + 1)^n N_c$ basis functions.

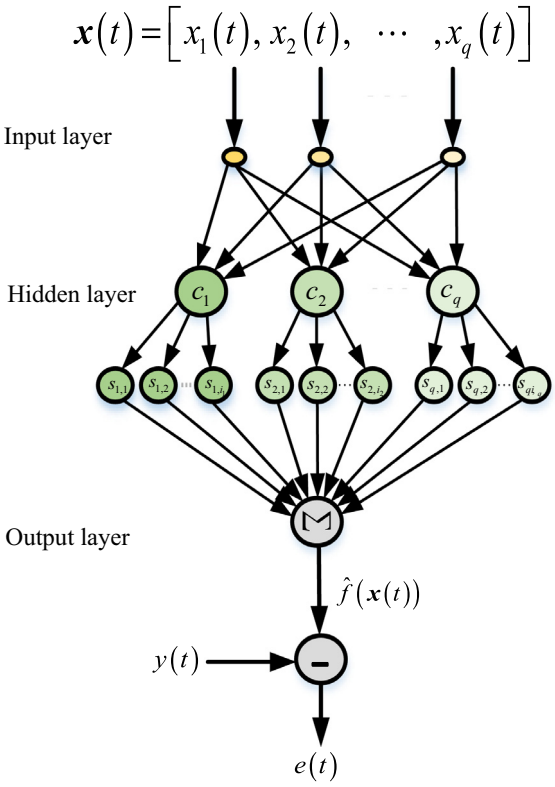


**Fig. 3.** Schematic of the MSRBF network.

*3.2.2. Determination of the center*

All measured observations can be regarded as candidate kernel centers $\mathbf{c}_m$ when the length of the measured dataset is not very long. In comparison, if a long dataset is considered and all the observations are still regarded as candidate kernel centers, then a great number of model terms or regressors will be included in the initial MSRBF network. Thus, the training of the network will be time-consuming. Recently, to resolve this problem, a sample of a learning algorithm based on the orthogonal least squares method was proposed to choose optimal centers one by one until an adequate network is constructed and each selected center maximizes the increment to the energy of the desired output (Chen et al., 1991). This method provided a simple but efficient way for choosing centers and avoiding numerical ill-conditioning. Shi et al. (2005) adopted sensitivity analysis to construct RBF networks, where the number of hidden neurons and the centers in the network were determined by the minimal number of hidden neurons with the maximal sensitivity and the maximization of the output's sensitivity to the training data respectively.

Generally, the well-known *k*-means clustering algorithm, which assumed that each data vector belongs exclusively to a single cluster, was applied to select the centers in the network (Moody and Darken, 1989). By contrast, in practical application, data vectors commonly belong to more than one cluster and are associated with each element as a set of membership level (Li et al., 2016a). Thus, the fast fuzzy c-means (FFCM) clustering algorithm was used to significantly reduce the number of candidate centers of the basis function in the network (Biniaz and Abbasi, 2013). The FFCM algorithm, coupled with the partition index criterion proposed by Bensaid et al. (1996a), is briefly introduced below.

In the FFCM clustering algorithm, the objective function with respect to fuzzy membership $\mu_{m,j}$ and cluster centers $\mathbf{c}_m$ is defined by

$$J_\gamma = \sum_{m=1}^{N_k}\sum_{j=1}^{N}\mu_{m,j}^{\gamma}\|\mathbf{x}_j - \mathbf{c}_m\|^2 \tag{8}$$

where $\mathbf{x}_j (j = 1, 2, \ldots, N)$ is the $j$th input vector, $N$ is the length of the input vector, $N_k$ is the number of clusters, $\gamma > 1$ is the fuzziness index, and $\|\cdot\|$ is the Euclidean norm.

The procedure of updating centers in the FFCM algorithm is briefly described as follows:

**Step1:** For $q$-dimensional input vector, rearrange $\mu_{mj}$ to $N_k \times N$ dimension matrix.

**Step2:** Let new fuzzy membership be $\mu^*_{mj}$ and label matrix $L = \{L^1, L^2, \ldots, L^{N_k}\}$, where $L^k$ is the label vector of the $k$th cluster in the current iteration.

**Step3:** Denote all data points corresponding to $L^k$ label matrix be $I^k$.

**Step4:** Define $I^k = I_1^k, I_2^k, \ldots, I_{c_{n_k}}^k$ for $k$th cluster, where $c_{n_k}$ is the number of data points in the $k$th cluster.

**Step5:** Update the center of the $k$th cluster using the following formula:

$$c_k^* = \frac{\sum_{i=1}^{c_{n_k}} I_i^k}{c_{n_k}} \tag{9}$$

Similar to the $k$-means clustering algorithm, the number of centers should also be given first in the FFCM algorithm. Some criteria, including partition coefficient, classification entropy, and partition index, can be applied to choose the optimal clusters. The partition coefficient and classification entropy are not directly connected to a geometrical property (Xie and Beni, 1991). Hence, the partition index criterion, which is a sum of the individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster, is adopted to determine the optimal clusters. The optimal value of $N_k$ is the value that minimizes the partition index (SC) as follows (Bensaid et al., 1996b):

$$SC(N_k) = \sum_{m=1}^{N_k} \frac{\sum_{j=1}^{N} \mu_{mj}^{\gamma} \|\mathbf{x}_j - \mathbf{c}_m\|^2}{\varpi_m \sum_{k=1}^{N_k} \|\mathbf{c}_m - \mathbf{c}_k\|^2} \tag{10}$$

where $\varpi_m$ is the fuzzy cardinality of cluster $m$ defined as $\varpi_m = \sum_{j=1}^{N} \mu_{mj}$.

The above FFCM clustering algorithm with partition index can be applied to select the number of centers in the MSRBF network. For a given training dataset of length $N$, let $N_c = \arg\min_{N_k}\{SC(N_k)\}$. Thus, the $N_c$ candidate centers (commonly $N_c \ll N$) will be at least included in the MSRBF network, which can be determined by a fast fuzzy $c$-mean clustering algorithm. Generally, the conventional $k$-means clustering algorithm can only obtain a locally optimal solution, which mainly depends on the initial locations in the cluster centers. To overcome this problem, some improved clustering algorithms including the FFCM is employed in this study (Biniaz and Abbasi, 2013).

#### 3.2.3. Determination of the scales

For the given $N_c$ pairs of input–output measurements $\{u(t), y(t)\}_{t=1}^{N}$, let $\sigma_u$ and $\sigma_y$ be the standard derivation of the input vector $\{u(t)\}_{t=1}^{N}$ and output vector $\{y(t)\}_{t=1}^{N}$ respectively. The scale vector in (7) can be defined by (Billings et al., 2007)

$$s_{y,m}^{(i)} = \beta\alpha^{-i}\sigma_y, i = 0, 1, \ldots, I \tag{11}$$

$$s_{u,m}^{(j)} = \beta\alpha^{-j}\sigma_u, j = 0, 1, \ldots, J \tag{12}$$

where $m = 1, 2, \ldots, N_c$ and $\alpha > 1$ are constants. A substantial number of experiments have shown that a good selection for the constants $\alpha$ and $\beta$ is to assign $\alpha = 2$ and $1 \leqslant \beta \leqslant 3$ (Wei et al., 2004a).

Let $\wp_3 = \{\varphi_{i,j,m}(\cdot; \delta_m^{(ij)}, \mathbf{c}_m)\}$ with $i = 0, \ldots, I$; $j = 0, \ldots, J$; $m = 1, \ldots, N_c$. The triple-indexed set $\wp_3$ is indicated as the dictionary associated with the new MSRBF networks. To simplify the descriptions, redefine the elements of $\wp_3$ so that the triple index $(i, j, m)$ can be represented by a single index $m = 1, 2, \ldots, M$, where $M$ is the number of basis functions included in the network, to form a single-index dictionary $\wp_1 = \{\phi_m(\cdot) : \phi_m \in \wp_3, m = 1, 2, \ldots, M\}$.In this study, two types of dictionaries $\wp_1$ and $\wp_3$ will not be distinguished, and both dictionaries will be used to indicate a uniform symbol $\wp$. Thus, the MSRBF network in (5) can be represented by

$$\hat{f}(\mathbf{x}(t)) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) \tag{13}$$

The derivation discussed in this section can easily be described as a multiple-input and single-output case, which will be discussed in Section 4.

### 3.3. Model structure selection and the MPOLS algorithm

In the nonlinear system identification and modeling procedure, model structure selection or model subset detection is usually a key step (Li et al., 2016c). It commonly consists of selecting first the significant model regressor terms from a redundant candidate model term, and then setting the significant regressor terms to construct a parsimonious final model. As shown in the later examples, numerous candidate model terms or basis functions may be involved in the MSRBF network (13) when the parameters $I$, $J$, and $N_c$ are large or the number of input variable $n$ is large. However, many of these candidate model terms may be redundant, and only a subset of these model terms is significantly included in the model (Li et al., 2016b, 2012). In particular, these redundant candidate model terms may lead to a large number of parameters for estimation in the model. Consequently, the redundant model may become oversensitive to the training dataset and may exhibit poor generalization properties (Wei et al., 2004b). Therefore, detecting which candidate terms should be contained in the MSRBF model is crucial. In the present study, an efficient MPOLS algorithm (Billings and Wei, 2005), which is a simple and relatively

computationally efficient regressor selection algorithm, is employed to solve the model structure detection problem for the MSRBF network models. The detailed description of the MPOLS algorithm can be found in Appendix A.

Once the significant model terms are chosen, we can easily verify that the relationship between the selected original basis $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$ and the associated orthogonal bases $w_1, w_2, \ldots, w_n$ are represented as

$$\mathbf{A}_n = \mathbf{W}_n \mathbf{U}_n \tag{14}$$

where $\mathbf{A}_n = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$, $\mathbf{W}_n$ is an $N \times n$ orthogonal matrix with columns $w_1, w_2 \ldots, w_n$, and $\mathbf{U}_n$ is an $n \times n$ unit upper triangular matrix with entries that are calculated from the orthogonalization procedure. The unknown parameter vector, which is denoted by $\boldsymbol{\theta}_n = [\theta_1, \ldots, \theta_n]^T$ for the model from the original bases, can then be obtained from the triangular equation $\mathbf{U}_n \boldsymbol{\theta}_n = \mathbf{g}_n$ with $\mathbf{g}_n = [g_1, g_2, \ldots, g_n]^T$ and $g_k = (\mathbf{y}^T w_k)/(w_k^T w_k)$.

### 3.4. Multi-step-ahead forecasting

Multi-step ahead forecasting is a challenging task in time-series prediction because of the increasing number of uncertainties from various sources, such as the lack of information and accumulated errors (Sorjamaa et al., 2007). The common goal of multi-step ahead time-series forecasting task is to predict the values of $y(t+s)$ with $s \geqslant 1$ using a historical time series $\{y(t)\}_{t=1}^{N}$ included $N$ observed sequences. To achieve this goal, a model or a predictor is commonly constructed from the available data. To obtain the multi-step-ahead predictors of nonlinear time series, the iterative approach is adopted (Wei and Billings, 2006). Considering the $s$-step-ahead forecasting problem, we first forecast the first step by adopting the model. The forecasted value is applied as a part of the input variables for the next step prediction. Thus, the entire horizon can be repeatedly predicted step by step.

The task for $s$-step-ahead forecasts is constructing a model that can predict the value of $y(t+s)$ by adopting a set of selected variables $\{y(t), y(t-1), \ldots, y(t-d+1)\}$. Let the trained one-step-ahead model be $f^{(1)} = f(y(t), y(t-1), \ldots, y(t-d+1))$; thus, the forecasts are defined by (Ben Taieb et al., 2012)

$$y(t+s) = \begin{cases} f^{(s)}(\hat{y}(t+1), \ldots, \hat{y}(t+s-1), y(N-d+s), \ldots, y(t)), s \in \{2, \ldots, d\} \\ f^{(s)}(\hat{y}(t+s-d), \ldots, \hat{y}(t+s-1)), s \in \{d+1, \ldots, s\} \end{cases} \tag{15}$$

where $f^{(s)}$ with $s \geqslant 1$ are nonlinear functions, and $d$ is the model order or maximum lag. In practical applications, the nonlinear function $f^{(s)}$ is generally unknown and complex. Therefore, a class of flexible models, which exhibit excellent approximation capabilities and can represent a broad class of highly complex systems, is required to achieve multi-step-ahead predictions. The model that employs MSRBF as the basis function to approximate the $s$-step predictor $f^{(s)}$ will be investigated in this study as a new approach to obtain accurate $s$-step predictions.

## 4. Experimental result

In the experiments, three scenarios with extreme high passenger flows from three subway stations in a Beijing subway network are employed to illustrate the applicability and effectiveness of the newly proposed MSRBF network for nonlinear system modeling compared with that of state-of-the-art prediction algorithms, including SVM, boosted regression tree (BRT), and single-scale RBF (SSRBF) networks. The mean absolute percentage error (MAPE), variance of absolute percentage error (VAPE), and root mean square error (RMSE) will be used to evaluate the performance of the associated models. The prediction performance is evaluated by adopting the defined MAPE, VAPE, and RMSE indices (Balestrassi et al., 2009; Boto-Giralda et al., 2010; Sun et al., 2015)

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\bar{y}} \times 100\% \tag{16}$$

$$\text{VAPE} = \frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^2 \times 100\% \tag{17}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{18}$$

where $N$ is the number of observational samples, $y = \{y_1, y_2, \ldots, y_N\}$, $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$, and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ represent the observations of the system output, the model predictions, and the average of the observations, respectively. $e_i = |(y_i - \hat{y}_i)|/y_i$ and $\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i$ indicate the absolute error and the mean of absolute error respectively.

### 4.1. Example 1: WKS subway station

In this subsection, WKS station, which is the target station for prediction, was considered a structure-unknown (black-box) dynamical system. The objective was to construct a mathematical model that can be applied to forecast the alighting passenger flow of WKS station on July 21, 2012. The detailed data description is presented in Section 2. The subway system was considered an 18-input and single-output system, whereas the alighting passenger flow of WKS station (circled in red) was the system output. The boarding passenger flows of 18 transfer stations (circled in other colors) in Fig. 1 were the system inputs. The smart card data were recorded from 4:45 am to 23:15 pm with a sampling rate of 15 min. Thus, 76 data points were obtained on a daily basis, resulting in a total of 15 day observations and $15 \times 76 = 1140$ data points, as indicated in Fig. 4. The first 1064 data points were used for model identification, and the remaining 76 data points were employed for model prediction. The objective was to conduct iterative multistep predictions using the proposed MSRBF modeling approach.

For the convenience of description, let $y_W(t)$ represent the alighting passenger flows of WKS target station, and $u_1(t), \ldots, u_{18}(t)$ indicate the boarding passenger flows from 18 original transfer stations in Beijing subway network, respectively.

The next step is to determine the maximum time lags of inputs (i.e., $n_y$ and $n_u$) shown in Equation (2). The time series of alighting passengers at WKS station was considered, and $y_W(t-1)$ was selected as the primary contributor to the current alighting passenger flow, indicating $n_y = 1$. The selection of $n_u$ is based on the shortest commuting times from the 18 transfer stations to the targeted WKS station. The passengers are assumed to choose these routes with the least number of transfers, and the commuting time between any two stations will be estimated by counting both in-train travel time and transfer time, as presented in Fig. 5. For example, transferring from FXM station to WKS station requires 30 min, which is equivalent to two time lags considering the sample rate of 15 min. That is to say, the boarding passenger flows at FXM station 30 min ago are unlikely to contribute to the alighting passenger flows at WKS station, and should not be included in the prediction model. However, the passengers boarding from the FXM station may take longer than 30 min to arrive at the WKS station because each passenger's actual travel time between the two stations may exceed the minimum commuting time (i.e., 30 min). To incorporate the late arrivals, a buffer of 30 min is added to the shortest commuting time, which is equivalent to two time lags. Therefore, the maximum time lag of boarding passenger flows from FXM station can be computed as the shortest commuting time (30 min) plus the buffer time (30 min) divided by the sampling rate (15 min), thus yielding four in this scenario. This result implies the boarding passenger flows at time steps $t-2, t-3, t-4$ are incorporated in the model, but the passenger flow at time step $t-1$ cannot influence the alighting passenger flow at WKS station because passengers still transfer from FXM station to WKS station. Based on a similar logic, the maximum time lags for the remaining 17 stations can be calculated. The results demonstrate that each station will contribute their boarding passenger flows with three time lags to the alighting passenger flow at WKS station. Therefore, a total of 55 candidate variables were chosen. The variables contain the alighting passenger flow at WKS station at time step $t-1$ and the boarding passenger flows at the remaining 18 stations, including three time lags according to the example calculation process. The variables are considered potential contributors for constructing the MSRBF network by using the MPOLS algorithm.

The contribution of the 55 candidate variables will unlikely be equal in the identified model. Some of the variables may be insignificant, and the associated model terms will be automatically removed from the model. The ERR criterion is a very useful index to indicate the significance of model terms defined in Appendix A. The model eventually includes only several significant terms, where most of the terms slightly contribute to the system output and can thus be disregarded. The GCV criteria, which is an effective model validation criteria for selecting the number of model terms, together with ERR criteria,

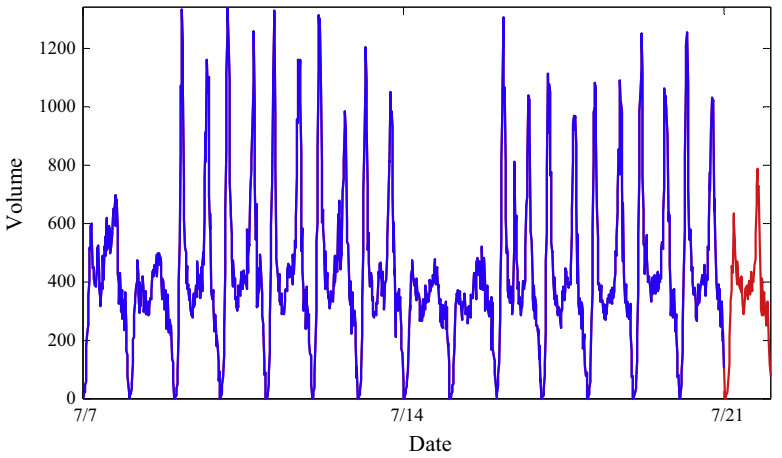


**Fig. 4.** Original transfer passenger flow time series at WKS station from July 7 to July 21, 2012. (Blue line is training data, Red line is predicted data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

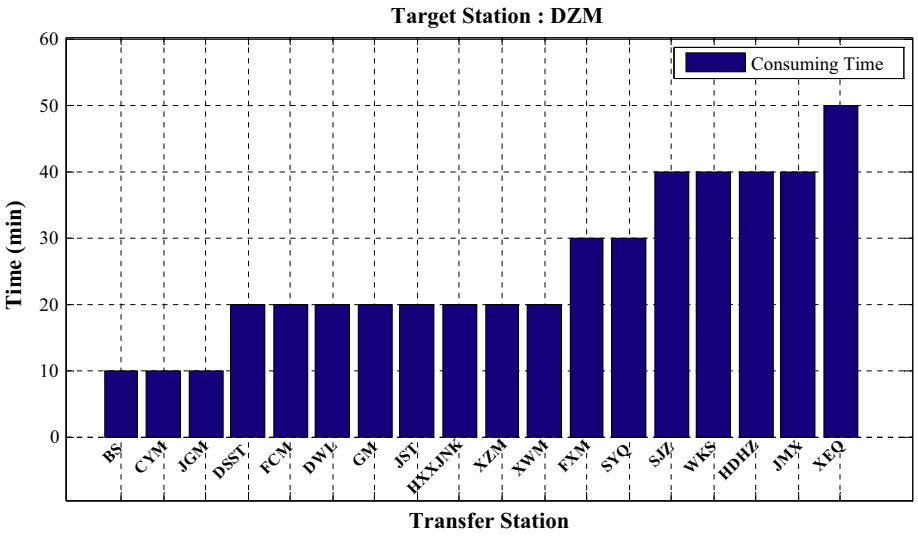


**Fig. 5.** Consumption time from the 18 transfer stations to the WKS target station.

are used to select the significant terms in the initial MSRBF network. Therefore, 11 significant variables were chosen from a total of 55 candidate variables (Wei et al., 2004b). The input vector for the MSRBF network model was chosen to be $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_{11}(t)] = [y_W(t-1), u_1(t-4), u_4(t-2), u_6(t-2), u_6(t-4), u_8(t-2), u_8(t-4), u_{11}(t-5), u_{14}(t-3), u_{16}(t-4), u_{17}(t-4)]$. The selected model variables with estimated parameters and ERR values are listed in Table 2.

The fast fuzzy $c$-means clustering algorithm was applied in the training dataset, and the SC index given in Eq. (10) suggests that the optimal number of clusters $N_c$ for this dataset was 15. The 1064 observational data samples were partitioned into 15 groups, and the centers of the 15 groups were chosen as the candidate centers for constructing the MSRBF network. The basis functions in the MSRBF network model (6) were given by

$$\boldsymbol{\varphi}_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)}) = \exp\left[-\left(\frac{x_1(t) - c_{m,1}}{s_{y_w,m}^{(i)}}\right)^2 - \sum_{k=2}^{11}\left(\frac{x_k(t) - c_{m,k}}{s_{u,m}^{(j)}}\right)^2\right] \tag{19}$$

where $m = 1, 2, \ldots, 15$ and the $m$th scale vector $\mathbf{s}_m^{(i,j)}$ is represented by

$$\mathbf{s}_m^{(i,j)} = \left[s_{y_w,m}^{(i)}, s_{u_1,m}^{(j)}, s_{u_4,m}^{(j)}, \overbrace{s_{u_6,m}^{(j)}, s_{u_6,m}^{(j)}}^{1:2}, \overbrace{s_{u_8,m}^{(j)}, s_{u_8,m}^{(j)}}^{1:2} s_{u_{11},m}^{(j)}, s_{u_{14},m}^{(j)}, s_{u_{16},m}^{(j)}, s_{u_{17},m}^{(j)}\right] \tag{20}$$

According to (11) and (12), parameters $\alpha$ and $\beta, I, J$ in the scale $\mathbf{s}_m^{(i,j)}$ from the MSRBF network were chosen as 2, 2, 1, and 1, respectively.

**Table 2**
Selected variables, estimated parameters, and the corresponding ERR values in Example 1.

| Search steps | Terms | Estimates of parameters | ERRs (×100%) |
|---|---|---|---|
| 1 | $y_w(t-1)$ | 0.5196 | 9.631E+001 |
| 2 | $u_{16}(t-4)$ | 0.0878 | 3.885E−001 |
| 3 | $u_4(t-2)$ | 0.1631 | 6.870E−001 |
| 4 | $u_8(t-2)$ | 0.1500 | 8.450E−002 |
| 5 | $u_1(t-4)$ | −0.1037 | 2.605E−001 |
| 6 | $u_{17}(t-4)$ | 1.0811 | 9.199E−002 |
| 7 | $u_6(t-2)$ | 0.4681 | 1.213E−001 |
| 8 | $u_6(t-4)$ | −0.5132 | 2.365E−001 |
| 9 | $u_{14}(t-3)$ | 0.1073 | 5.580E−002 |
| 10 | $u_8(t-4)$ | −0.1346 | 7.849E−002 |
| 11 | $u_{11}(t-5)$ | 0.0537 | 2.356E−002 |

*Note:* $y_w$, $u_4$, $u_{16}$, $u_{14}$, $u_6$, $u_{12}$, $u_7$, and $u_{18}$ indicate the passenger flow of the WKS target station, and JST, HXXJNK, DZM, JMX and DWL, XZM, and XEQ are the respective transfer stations.

Thus, the initial network model involves a total of $11 + 2^{11} \times 15 = 30731$ candidate model regressor terms. The MPOLS algorithm was adopted to the 30,731 candidate regressor terms over the 1064 training data points. According to the value of GCV, 67 significant regressor terms were selected and adopted to build the final MSRBF network for forecasting the time series of passenger flow.

The validation dataset was applied to test the performance of the identified model, which consists of 76 data points of alighting passenger flow at WKS station on July 21, 2012. When a concert was held at 19:00 near the WKS target station, the passenger flow suddenly increased. Multi-step ahead prediction was discussed in Section 3. The one-step ahead prediction is equivalent to 15-min-ahead prediction because the interval time is 15 min. To allow a more proactive action to mitigate the adverse effect of a sudden increase of passenger flow, two-step ahead prediction is conducted. This process will notify both transit operators and travelers 30 min ahead of an extreme event occurrence in a timely fashion. Both one-step ahead and two-step ahead predictions are calculated from the identified 67-term MSRBF model and other conventional methods, such as SVM, BRT, and SSRBF network. All the comparative methods adopted the same 55 candidate variables as input matrix. The results of SVM were acquired by using the LIBSVM package, where the optimal parameters were obtained by particle swarm optimization algorithm. The BRT algorithm was implemented by following the work of (Hara and Chellappa, 2013) with the optimal number of trees. The scale of SSRBF was also optimally determined. The performance comparisons of both one-step-ahead and two-step-ahead (TSA) prediction results with different models are presented in Figs. 6 and 7.

The results show that the proposed MSRBF network exhibited a superior performance to predict non-passenger flow. A demand peak was observed at around 18:30 PM in WKS station because of the concert. The sudden increase of passenger flow could be successfully identified by the proposed algorithm. Fig. 6 indicates that the MSRBF network outperforms other methods with a closer estimate of passenger flows with special events. Even for a more rigorous TSA prediction test, the proposed MSRBF network still exhibits the best prediction performance compared with other algorithms, as indicated in Fig. 7.

Fig. 8 presents the histograms of the alighting passenger flow error measures obtained by the proposed MSRBF network and conventional methods for one-step-ahead and TSA predictions. The proposed MSRBF network generates the lowest forecasting errors (MAPE, VAPE, and RMSE) compared with the other approaches, thus demonstrating the effectiveness of the proposed method.

The statistical results in Table 3 indicate that the proposed MSRBF network consistently outperforms other approaches with the lowest MAPE, VAPE, and RMSE values for both one-step-ahead and TSA predictions. For example, MAPE, VAPE, and RMSE values from the proposed MSRBF network are 11.8038, 1.5910, and 62.7863 for one-step-ahead predictions, and 15.1206, 3.2279, and 85.3464 for TSA predictions, respectively. These metrics are extremely lower than those of other methods, such as SVM, BRT, and SSRBF. Therefore, these results quantitatively indicate that the proposed MSRBF network effectively predicts a sudden increase in subway passenger flow.

One advantage of the proposed MSRBF network is that the structure of the MSRBF network involves both linear and nonlinear parts, which can effectively describe the stochastics and dynamic relationship for nonlinear systems. Although the nonlinear part may increase computational complexity, the MPOLS algorithm is applied to eliminate most redundant terms for less computational burdens. SVM also performs well in predicting small-scale datasets. However, the proposed MSRBF network functions well for both small-scale and large-scale datasets. Unlike the traditional SSRBF network, the MSRBF network accommodates both local and global properties of the basis functions to include both small and large scales in

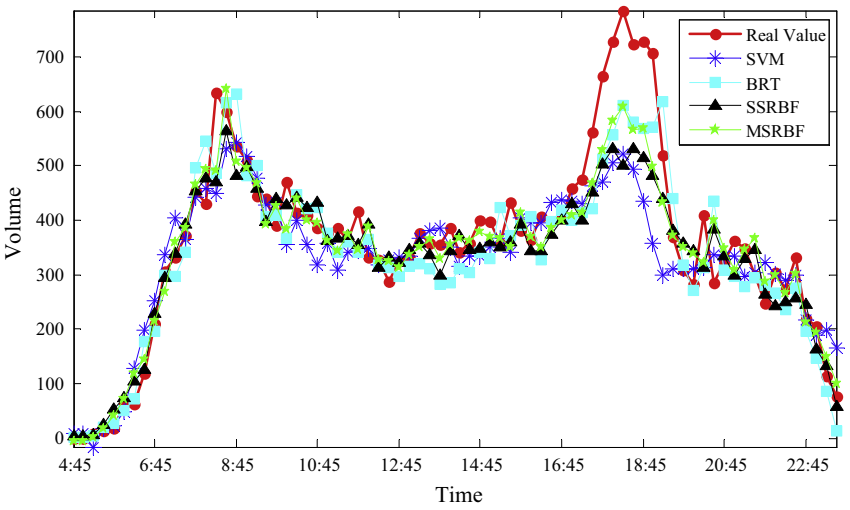


**Fig. 6.** One-step-ahead prediction results of alighting passenger flow at WKS station on July 21, 2012.

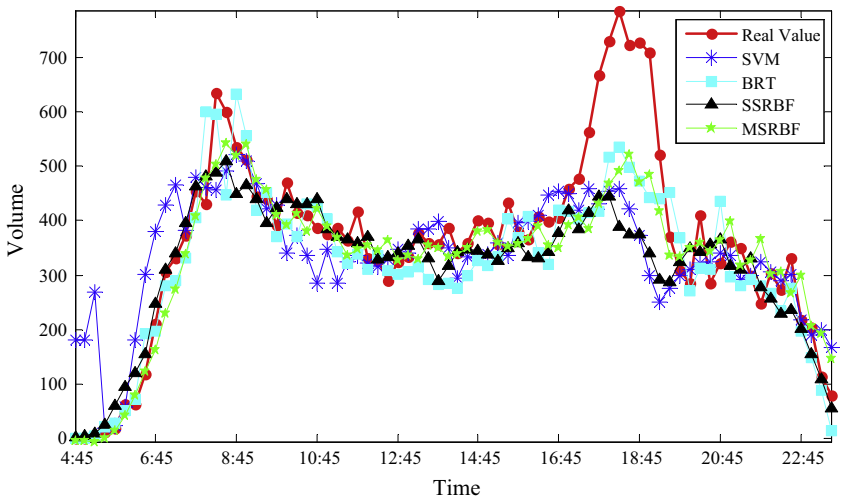


**Fig. 7.** Two-step-ahead prediction results of alighting passenger flow at WKS station on July 21, 2012.

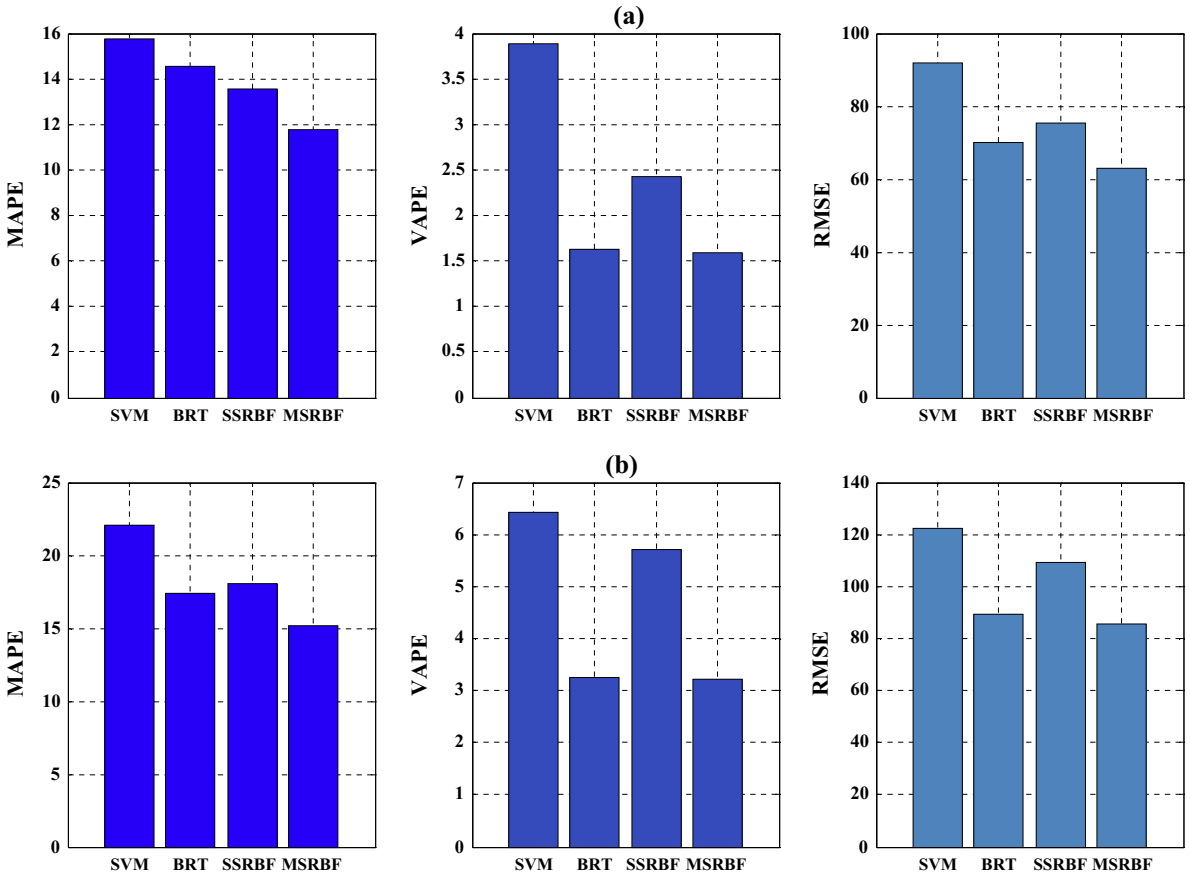


**Fig. 8.** Three-measure values from four forecasting methods at WKS target station: (a) one-step-ahead predictions, (b) two-step-ahead predictions.

the network, and thus yield better prediction performance. The proposed MSRBF network generates the optimal forecasting performance, particularly on predicting unusual passenger flow, which likely implies the occurrence of a special event, such as a concert or sport game.

**Table 3**
Predictions of passenger flow at WKS target station on July 21, 2012.

| Methods | One-step-ahead | | | Two-step-ahead | | |
|---|---|---|---|---|---|---|
| | MAPE | VAPE | RMSE | MAPE | VAPE | RMSE |
| **SVM** | 15.7608 | 3.8876 | 91.7416 | 22.0388 | 6.4300 | 122.1048 |
| **BRT** | 14.5227 | 1.6327 | 70.1059 | 17.3865 | 3.2573 | 89.3661 |
| **SSRBF** | 13.5816 | 2.4339 | 75.1777 | 18.0762 | 5.7126 | 108.9132 |
| **MSRBF** | **11.8038** | **1.5910** | **62.7863** | **15.1206** | **3.2279** | **85.3464** |

*Note:* The bold values indicate the minimum error rate from Example 1.

### 4.2. Example 2: DZM subway station

Unlike the WKS station, the alighting passenger volume at DZM station on September 29, 2012 significantly increased for the entire day compared with the usual weekends, as indicated in Fig. 2(b). This result occurred because most passengers returned home before the Mid-autumn festival via the long-distance bus hub adjacent to the station. Similar to Example 1, the subway system was also considered an 18-input and single-output system. Let system output variable $y_D(t)$ be the alighting passenger flow at DZM station, and the remaining 18 main transfer stations $u'_1(t),\ldots,u'_{18}(t)$ are the boarding passenger flows as system input variables. Therefore, the 1140 data points from September 15 to September 29 can be visualized in Fig. 9.

The shortest commuting time between DZM subway station and other 18 transfer stations is calculated in Fig. 10. Similar to example 1, 55 candidate variables, $y_D(t-j_D)$ with $j_D=1$, $u'_k(t-j_1)$ with $k=1,2,3$ and $j_1=1,2,3$, $u'_k(t-j_2)$ with $k=4,\ldots,12$ and $j_2=2,3,4$, $u'_k(t-j_3)$ with $k=13,\ldots,17$ and $j_3=3,4,5$, $u'_k(t-j_4)$ with $k=18$ and $j_4=4,5,6$, were selected. Both the GCV and ERR criteria were applied to tune up the 12 most significant terms among the initial 55 variables into the prediction model. Therefore, the input vector for the MSRBF network model can be represented as $\mathbf{x}(t)=[x_1(t),x_2(t),\ldots,x_{12}(t)]=[y_D(t-1),u'_1(t-1),u'_2(t-3),u'_3(t-1),u'_8(t-2),u'_8(t-4),u'_9(t-4),u'_{13}(t-3),u'_{13}(t-4),u'_{15}(t-3),u'_{17}(t-3),u'_{17}(t-5)]$. The selected model variables with the estimated parameters and ERR values are shown in Table 4.

Applying the fast fuzzy c-means clustering algorithm to cluster the centers leads to the optimal number of clusters $N_c$, which is five. The training dataset is partitioned into five groups, and the basis functions in the MSRBF network model (6) are given by

$$\varphi_{i,j,m}(\mathbf{x}(t);\mathbf{c}_m,\mathbf{s}_m^{(i,j)})=\exp\left[-\left(\frac{x_1(t)-c_{m,1}}{s_{y_D,m}^{(i)}}\right)^2-\sum_{k=2}^{12}\left(\frac{x_k(t)-c_{m,k}}{s_{u,m}^{(j)}}\right)^2\right] \tag{21}$$

where $m=1,2,\ldots,5$, and the $m$th scale vector $\mathbf{s}_m^{(i,j)}$ is defined by

$$\mathbf{s}_m^{(i,j)}=\left[s_{y_D,m}^{(i)},s_{u'_1,m}^{(j)},s_{u'_2,m}^{(j)},s_{u'_3,m}^{(j)},\overbrace{s_{u'_8,m}^{(j)},s_{u'_8,m}^{(j)}}^{1:2},s_{u'_9,m}^{(j)},\overbrace{s_{u'_{13},m}^{(j)},s_{u'_{13},m}^{(j)}}^{1:2},s_{u'_{15},m}^{(j)},\overbrace{s_{u'_{17},m}^{(j)},s_{u'_{17},m}^{(j)}}^{1:2}\right] \tag{22}$$

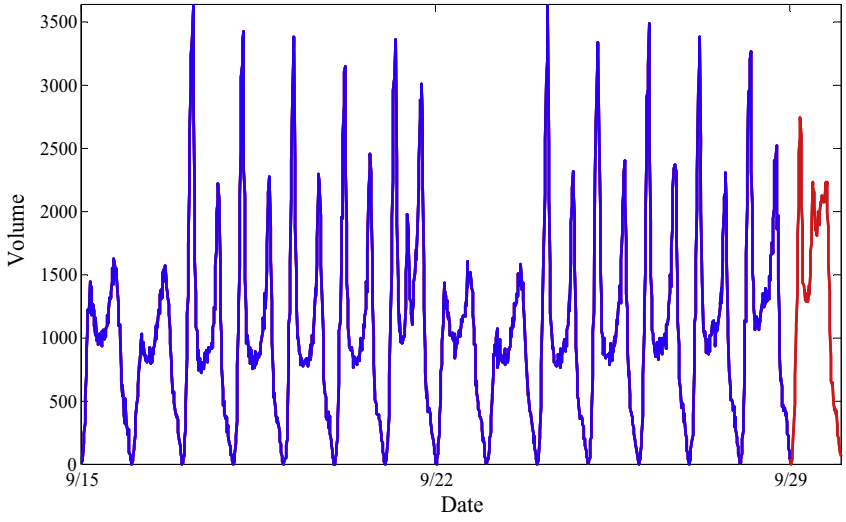


**Fig. 9.** Original transfer passenger flow time series at DZM station from September 15–29, 2012. (Blue line is training data, and red line is predicted data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

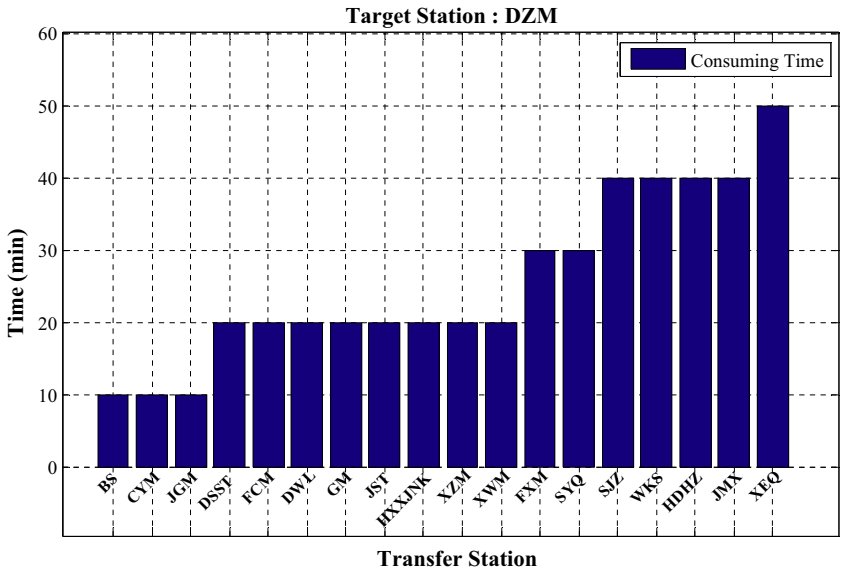


**Fig. 10.** Real consumption time between DZM subway station and other 18 subway stations.

**Table 4**
Selected model terms, estimated parameters, and ERR values in Example 2.

| Search steps | Term | Estimates of parameters | ERRs (×100%) |
|---|---|---|---|
| 1 | $y_D(t-1)$ | 0.7541 | 9.754e+001 |
| 2 | $u'_8(t-4)$ | −0.1193 | 7.735e−001 |
| 3 | $u'_8(t-2)$ | 0.3999 | 5.166e−001 |
| 4 | $u'_{13}(t-3)$ | 2.6696 | 9.999e−002 |
| 5 | $u'_{15}(t-3)$ | 0.1299 | 1.172e−001 |
| 6 | $u'_9(t-4)$ | −0.3400 | 4.845e−002 |
| 7 | $u'_{17}(t-3)$ | 0.7057 | 4.825e−002 |
| 8 | $u'_3(t-1)$ | 0.3398 | 4.023e−002 |
| 9 | $u'_2(t-3)$ | −0.1945 | 2.793e−002 |
| 10 | $u'_{17}(t-5)$ | −0.5443 | 3.405e−002 |
| 11 | $u'_1(t-1)$ | 0.0811 | 2.859e−002 |
| 12 | $u'_{13}(t-4)$ | −1.2956 | 1.346e−002 |

*Note:* $y_D(t), u'_8(t), u'_{13}(t), u'_{15}(t), u'_9(t), u'_{17}(t), u'_3(t), u'_2(t), u'_1(t)$ represent the passenger flows of DZM, JST, SYQ, WKS, HXXJNK, JMX, JGM, CYM, and BS stations, respectively.

According to (11) and (12), the scales of the MSRBF network $\mathbf{s}_m^{(i,j)}$ can be calculated, which is similar to example 1, where $\alpha, \beta, I, J$ are also selected to be 2, 1, 1, and 1.

The initial network model involves 12 linear terms and $5 \times 2^{12} = 20(,)480$ nonlinear terms in total. The MPOLS algorithm is applied to select the most important terms over 20,492 terms. A total of 36 significant terms are selected according to the value of the GCV criteria. The 36 terms are adopted to form the final MSRBF network to forecast the passenger flow series. A total of 76 samples on September 29, 2012 are used for the one-step ahead and two-step ahead predictions. The comparison results from four methods are shown in Figs. 11 and 12.

Fig. 11 shows that the MSRBF network yields more accurate results than the SSRBF network and the BRT model and significantly outperforms SVM. The trend of abnormal subway passenger flows can be accurately captured by the proposed MSRBF network. To further demonstrate the effectiveness of the proposed algorithm, multi-step-ahead prediction is undertaken to prevent the model from achieving good performance for one-step-ahead prediction, although the model is inferior. The comparison results from different forecasting methods are shown in Fig. 12. Similar to the outcome for one-step-ahead prediction, the proposed MSRBF network still exhibited the best performance among all the tested methods, whereas the prediction accuracy the SSRBF and BRT decreased. These results are further confirmed by the quantitative measures in Fig. 13 and Table 5, which indicate that the MAPE, VAPE, and RMSE values from the proposed MSRBF network are significantly lower than those of the other conventional prediction methods. The proposed algorithm can accurately detect the occurrence of a special event at least 30 min in advance, thus saving a considerable amount of time for information dissemination and crowd control.

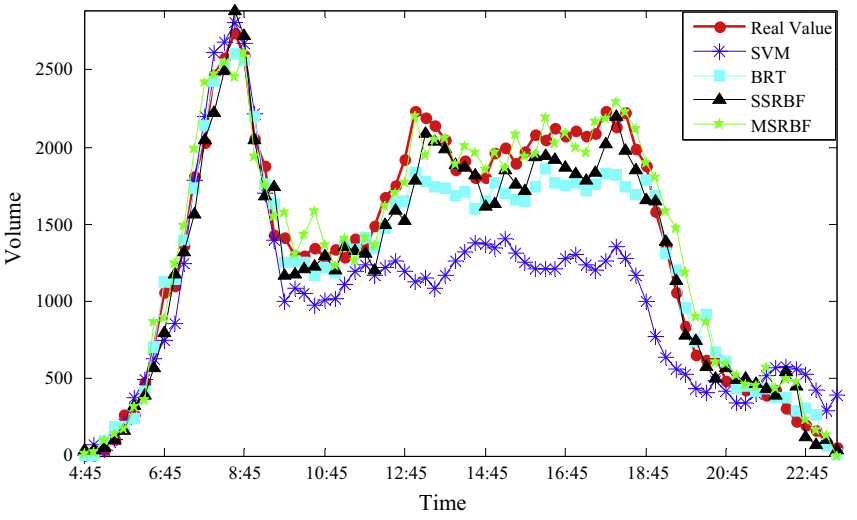


**Fig. 11.** One-step-ahead prediction results of alighting passenger flow at DZM station on Sept. 29, 2012.

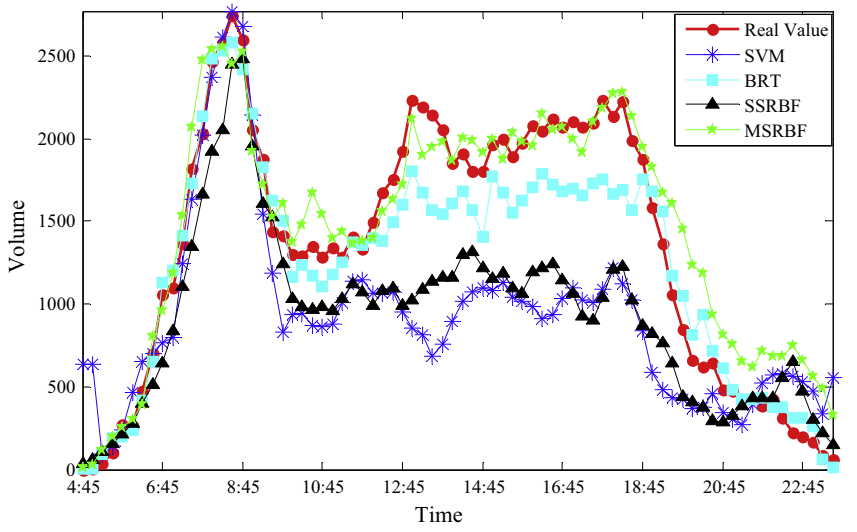


**Fig. 12.** Two-step-ahead prediction results of alighting passenger flow at DZM station on Sept. 29, 2012.

### *4.3. Example 3: HDHZ subway station*

Concerts events or returning home before the festival are categorized as planned events. Example 3 is set as an unplanned event to demonstrate the superiority of the proposed method further. The alighting passenger volume at HDHZ station on August 23, 2012 increased in the afternoon unlike that on usual weekdays, as presented in Fig. 2(c). This event occurred because the train signal failed for a while and the passengers remained at the HDHZ station. Similar to the previous examples, the HDHZ subway system was an 18-input and single-output system. Let $y_H(t)$ be the alighting passenger flow at HDHZ station and $u_1^*(t),\ldots,u_{18}^*(t)$ be the boarding passenger flows from the remaining 18 transfer stations. The 1140 data points from August 9–23 are presented in Fig. 14.

Fig. 15 shows the shortest commuting time between HDHZ subway station and other 18 transfer stations. The 55 candidate variables of the system, $y_H(t-p_H)$ with $p_H=1$, $u_k^*(t-p_1)$ with $k=1,2,\ldots,10$ and $p_1=2,3,4$, $u_k^*(t-p_2)$ with $k=11,\ldots,16$ and $p_2=3,4,5$, $u_k^*(t-p_3)$ with $k=17,18$. The initial 10 significant terms were determined by the GCV and ERR criteria, and the input vector for the MSRBF network model can be represented as $\mathbf{x}(t)=[x_1(t),x_2(t),\ldots,x_{10}(t)]=[y_H(t-1),u_1^*(t-2),u_1^*(t-4),u_8^*(t-3),u_9^*(t-3),u_9^*(t-4),u_{16}^*(t-3),u_{16}^*(t-4),u_{18}^*(t-5),u_{18}^*(t-6)]$. The selected model variables with estimated parameters and ERR values are presented in Table 6.

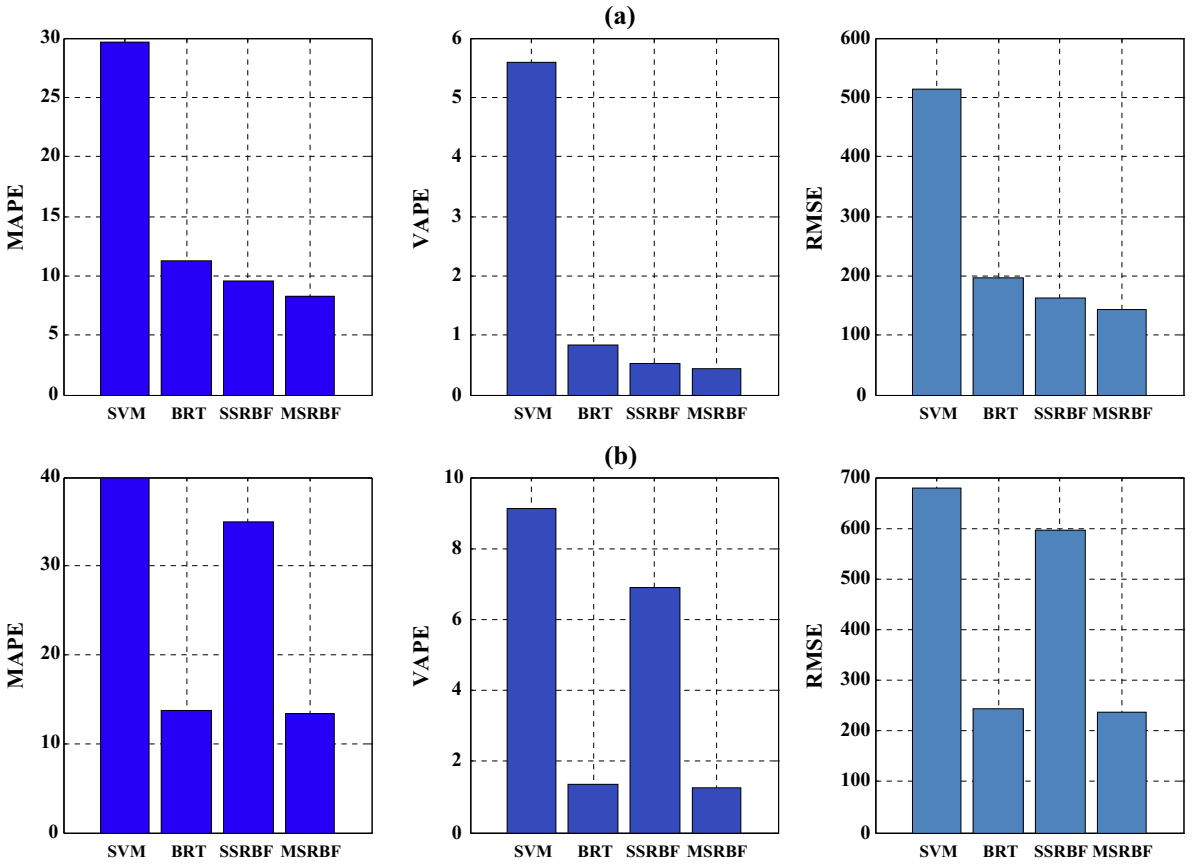


**Fig. 13.** Three measures from four methods at DZM target station: (a) one-step-ahead predictions, and (b) two-step-ahead predictions.

**Table 5**
Predictions of passenger flow at DZM station on Sept. 29, 2012.

| Methods | One-step-ahead | | | Two-step-ahead | | |
|---|---|---|---|---|---|---|
| | MAPE | VAPE | RMSE | MAPE | VAPE | RMSE |
| **SVM** | 29.6094 | 5.5889 | 514.0743 | 39.9336 | 9.1452 | 679.6365 |
| **BRT** | 11.2654 | 0.8311 | 196.5157 | 13.7130 | 1.3375 | 243.2507 |
| **SSRBF** | 9.5299 | 0.5431 | 162.9400 | 35.0508 | 6.9167 | 594.5454 |
| **MSRBF** | **8.2634** | **0.4411** | **143.8376** | **13.3725** | **1.2359** | **235.9457** |

*Note:* The bold values indicate the minimum error rate from Example 2.

The optimal number of clusters $N_c$ is 29, which was obtained by the fast fuzzy $c$-means clustering algorithm. Therefore, the training data set is partitioned into 29 groups. The basis functions in the MSRBF network model (6) were given by

$$\varphi_{i,j,m}(\mathbf{x}(t);\mathbf{c}_m,\mathbf{s}_m^{(i,j)}) = \exp\left[-\left(\frac{x_1(t)-c_{m,1}}{s_{y_D,m}^{(i)}}\right)^2 - \sum_{k=2}^{10}\left(\frac{x_k(t)-c_{m,k}}{s_{u,m}^{(j)}}\right)^2\right] \tag{23}$$

where $m = 1, 2, \ldots, 29$ and the $m$th scale vector $\mathbf{s}_m^{(i,j)}$

$$\mathbf{s}_m^{(i,j)} = \left[s_{y_H,m}^{(i)}, \overbrace{s_{u_1^*,m}^{(j)}, s_{u_1^*,m}^{(j)}}^{1:2}, s_{u_8^*,m}^{(j)}, \overbrace{s_{u_9^*,m}^{(j)}, s_{u_9^*,m}^{(j)}}^{1:2}, \overbrace{s_{u_{16}^*,m}^{(j)}, s_{u_{16}^*,m}^{(j)}}^{1:2}, \overbrace{s_{u_{18}^*,m}^{(j)}, s_{u_{18}^*,m}^{(j)}}^{1:2}\right] \tag{24}$$

The scales of the MSRBF network $\mathbf{s}_m^{(i,j)}$ can be calculated by Eqs. (11) and (12), and parameters $\alpha, \beta, I, J$ are selected to be 2, 2, 1, and 1.

The initial network model of HDHZ subway station involves a total of 10 linear terms and $29 \times 2^{10} = 29,696$ nonlinear terms. Similar to the previous examples, the MPOLS algorithm is utilized to eliminate the redundant terms of over 29,706

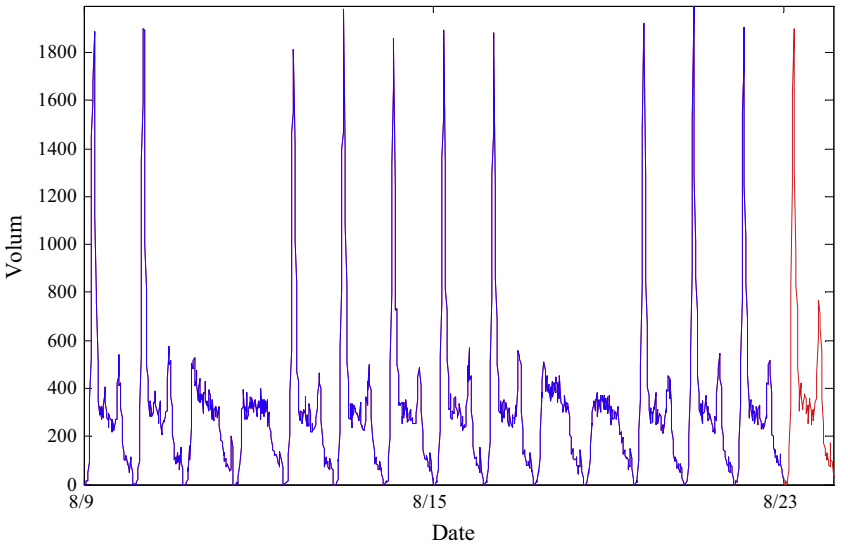


**Fig. 14.** Original transfer passenger flow time series at HDHZ station from August 9–23, 2012. (Blue line is training data, and red line is predicted data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

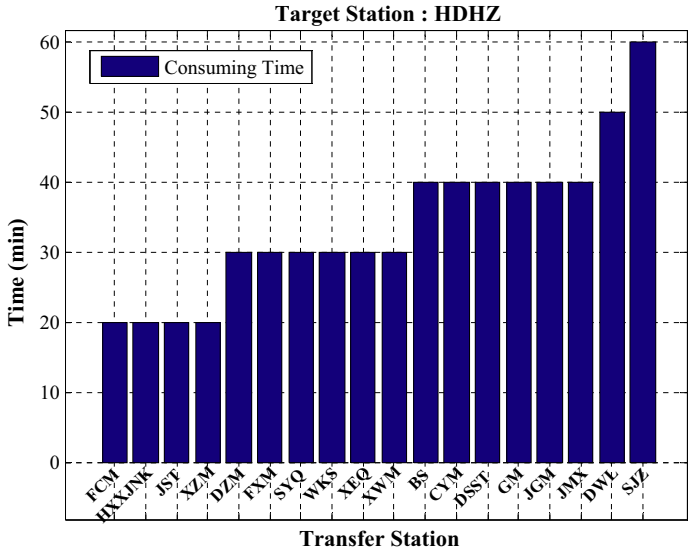


**Fig. 15.** Real consumption time between HDHZ subway station and other 18 subway stations.

**Table 6**
Selected model terms, estimated parameters, and ERR values in Example 3.

| Search steps | Term | Estimates of parameters | ERRs($\times$100%) |
|---|---|---|---|
| 1 | $u^*_{16}(t-4)$ | 0.3434 | 9.572e+001 |
| 2 | $u^*_{8}(t-3)$ | 0.0307 | 1.9115e+000 |
| 3 | $u^*_{18}(t-6)$ | −0.2457 | 2.789e−001 |
| 4 | $y_H(t-1)$ | 0.3580 | 4.896e−001 |
| 5 | $u^*_{16}(t-3)$ | 0.6018 | 1.588e−001 |
| 6 | $u^*_{9}(t-4)$ | −0.0216 | 5.048e−002 |
| 7 | $u^*_{1}(t-4)$ | 0.1019 | 9.345e−002 |
| 8 | $u^*_{18}(t-5)$ | 0.1628 | 2.280e−002 |
| 9 | $u^*_{9}(t-3)$ | −0.0571 | 2.012e−002 |
| 10 | $u^*_{1}(t-2)$ | 0.0410 | 1.690e−002 |

*Note:* $u^*_{16}(t), u^*_8(t), u^*_{18}(t), y_H(t), u^*_9(t), u^*_1(t)$ represent the passenger flows of the JMX, WKS, SJZ, HDHZ, XEQ, and FCM stations, respectively.

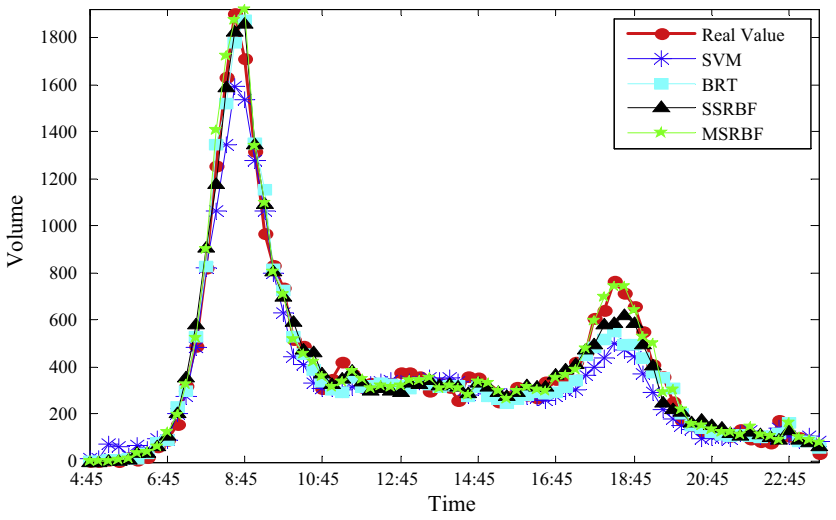


**Fig. 16.** One-step-ahead prediction results of alighting passenger flow at HDHZ station on Aug. 23, 2012.

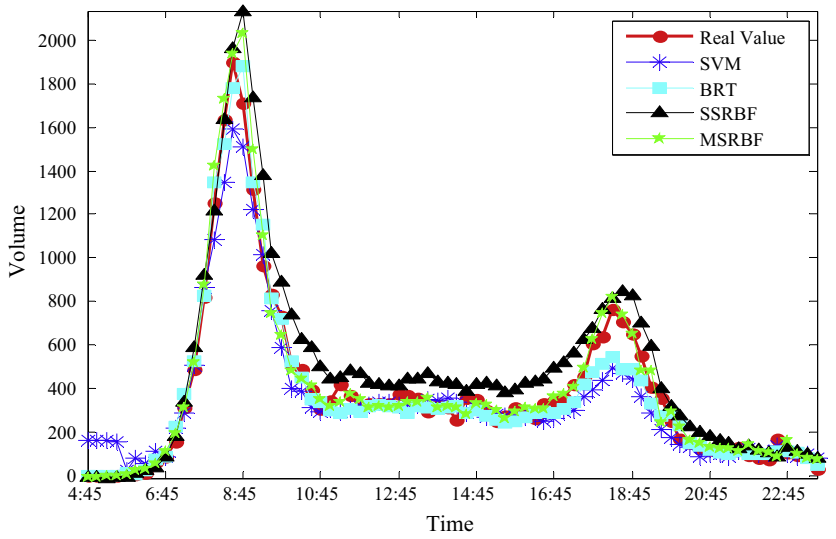


**Fig. 17.** Two-step-ahead prediction results of alighting passenger flow at HDHZ station on Aug. 23, 2012.

terms, and 19 significant terms are selected according to the value of the GCV criteria. The 19 terms are adopted to form the final MSRBF network to forecast the passenger flow series on August 23, 2012. The 76 samples on that day are used for the one-step ahead and two-step ahead predictions. The comparison results from four methods are shown in Figs. 16 and 17 respectively.

The one-step-ahead prediction and TSA prediction are presented in Figs. 16 and 17 respectively. The figures show that the MSRBF network yields the result of unplanned event more exactly than the tested methods BRT, SVM, and SSRBF network in

**Table 7**
Predictions of passenger flow at HDHZ station on Aug. 23, 2012.

| Methods | One-step-ahead | | | Two-step-ahead | | |
|---|---|---|---|---|---|---|
| | MAPE | VAPE | RMSE | MAPE | VAPE | RMSE |
| **SVM** | 16.8591 | 3.3995 | 97.5344 | 19.7495 | 3.5522 | 106.5762 |
| **BRT** | 11.2488 | 1.7529 | 67.8242 | 12.1775 | 1.7550 | 70.2490 |
| **SSRBF** | 10.6231 | 0.8299 | 54.6337 | 25.6854 | 5.0596 | 133.2901 |
| **MSRBF** | **9.0696** | **0.8280** | **50.1554** | **11.0555** | **1.5470** | **64.9659** |

*Note:* The bold values indicate the minimum error rate from Example 3.

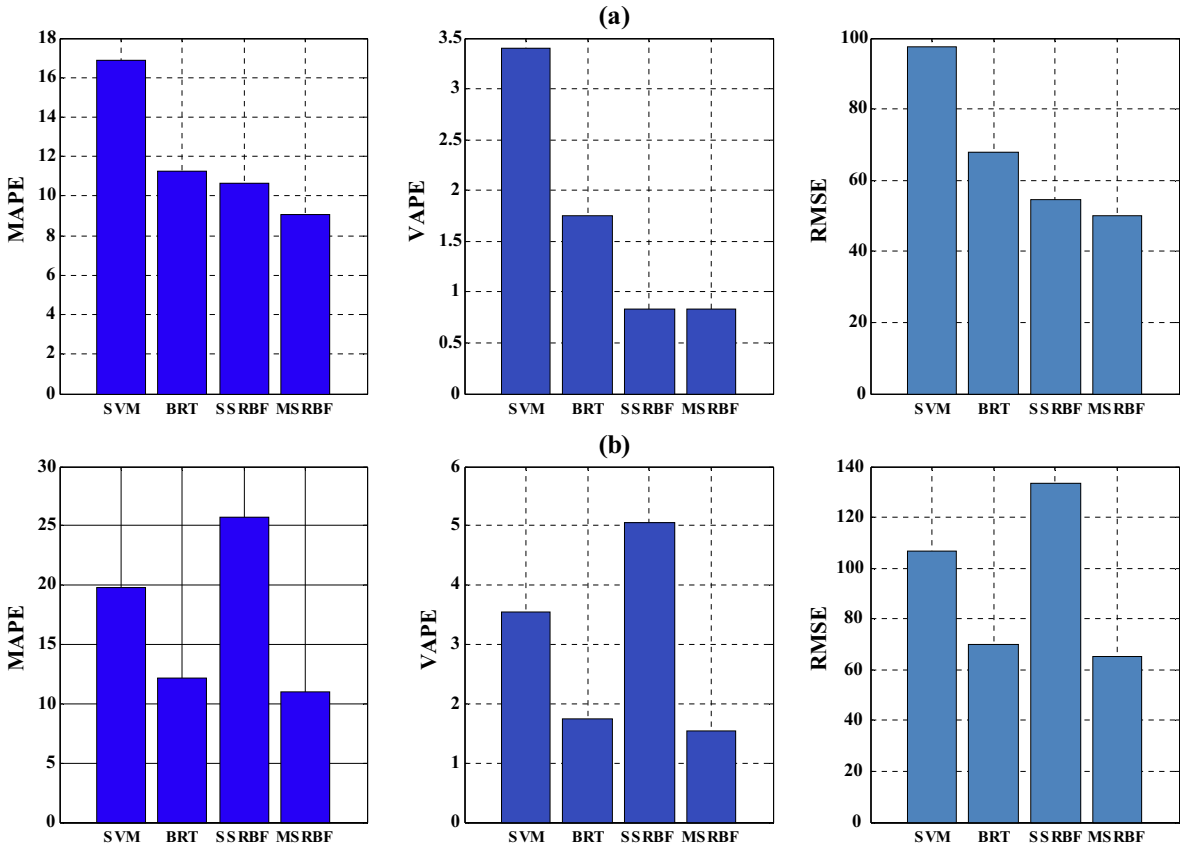


**Fig. 18.** Three measures from four methods at HDHZ target station: (a) one-step-ahead predictions, and (b) two-step-ahead predictions.

the one-step-ahead prediction and TSA prediction. This result demonstrates that the proposed MSRBF network could determine and reflect the complex relationship between the subway stations and exhibit superiority in forecasting multi-step-ahead prediction.

To illustrate the superiority of the proposed method, the results are further confirmed by the quantitative measures in Table 7 and Fig. 18. The MAPE, VAPE, and RMSE values from the proposed MSRBF network are lower than those of the other conventional prediction methods, which are significantly lower than those of SVM and SSRBF. Compared to the BRT method, which both shows a stable performance on one-step-ahead prediction and TSA prediction, the results of proposed MSRBF method are more accurate.

The proposed MSRBF network exhibited the best forecasting performance among the four methods, particularly in predicting the unusual passenger flow. The occurrence of planned events, such as concert or sport game, and the unplanned events, such as train signal failure, could both be forecasted 30 min earlier.

## 5. Conclusions

Predicting subway disruptions caused by special events is a well-known challenge in transportation planning and operation, because non-regular passenger demands are very difficult to determine using traditional statistical methods or machine learning approaches. A timely and reliable detection method of sudden demand change is crucial for transit operators to disseminate evaluation information to the crowd and allocate resources (e.g., feeder and shuttle buses) to transport heavy passenger flows in subway systems. In addition, understanding the relationship between incoming passenger flows at the network level and the demand surge at a particular station can assist governmental agencies in policy making. For example, subway operators may strategically close certain subway stations to limit passenger crowding before special events, and the government can adopt a more flexible subway fare to reduce passenger demands.

Most existing literature focuses on regular passenger demand forecasting in short- or long-prediction horizons with very limited studies on subway passenger flow prediction under special events. In this paper, a novel MSRBF network approach is proposed for forecasting the short-term passenger flow (30 min ahead). The results obtained with the proposed methodology using multi different scales are compared with conventional techniques, such as BRT, SVM, and SMRBF. The proposed approach for predicting short-term passenger flow outperforms these techniques, particularly when an emergent event

occurs. The proposed model involves two parts. The linear submodel can be used to track the linear relationship between the output and the inputs, and the MSRBF submodel can be applied to capture the effect of the nonlinear dynamics, which may not be completely explained by the linear submodel. This measure effectively enhances the capability of predicting short-term passenger flow.

The main disadvantage of the proposed model approach is that a large number of candidate terms may be involved in the initial model when the number of system input vectors is large. To address this problem, an effective MPOLS algorithm is employed to select the most significant candidate terms to construct a parsimonious linear parameters model. Thus, the problem is resolved successfully. The results of this study provide a new way to predict the short-term passenger flow, which is based on the parameter model. The proposed MSRBF network modeling framework, together with the MPOLS algorithm, can be improved further to achieve better prediction effect.

The methodology proposed in this study was developed for multi-step ahead forecasting results obtained only from the passenger flow of a Beijing subway network. The degree of generality of the algorithm cannot be assessed at present. Future studies can investigate the performance of the proposed method on different benchmark datasets and some practical applications on real time-series forecasts using the promising MSRBF network approach. Another interesting future direction is to incorporate the transfer passenger demands from adjacent bus stops. This treatment is expected to enhance prediction accuracy from the perspective of multimodal transportation systems.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China [61671042, 61403016, 51408019, U1564212, 71402011], the Science and Technology Research Foundation for Transportation [2015318221020], the Specialized Research Fund for the Doctoral Program of Higher Education [20131102120008], the Beihang University Innovation and Practice Fund [YCSJ-02-2015-12], and the Beijing Nova Program [z151100000315048].

## Appendix A

*The MPOLS algorithm*

Let $\mathbf{y} = [y(1), \ldots, y(N)]^T$ be a vector of observational outputs at $N$ time instants and $\boldsymbol{\varphi}_m = [\phi_m(1), \ldots, \phi_m(N)]^T$ be a vector formed by the $m$th candidate model term with $m = 1, 2, \ldots, M$. Meanwhile, let $\wp = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M\}$ be a dictionary with the $M$ candidate bases. A large number of modeling experiences and practical modeling identifications showed that finite dimensional set $\wp$ is usually redundant. The model term selection problem is equivalent to finding, from $M$ candidate model terms, a full dimensional subset $\wp_n = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n\} = \{\boldsymbol{\varphi}_{i_1}, \ldots, \boldsymbol{\varphi}_{i_n}\}$ of $n$ bases where $n \leqslant M$. Thus, $\mathbf{y}$ can be approximated by using a linear parameter combination of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \ldots, \boldsymbol{\alpha}_n$ as below

$$\mathbf{y} = \theta_1\boldsymbol{\alpha}_1 + \ldots + \theta_n\boldsymbol{\alpha}_n + \mathbf{e} \tag{25}$$

or in a compact matrix expression

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \tag{26}$$

where the matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ is a full column rank, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]$ is a coefficient vector, and $\mathbf{e}$ is the approximation error.

The procedure of model structure selection starts in (25) to find vector $\boldsymbol{\varphi}_{\ell_1}$ from the candidate regressor family $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M\}$. Thus, $\boldsymbol{\varphi}_{\ell_1}$ is the "best" matching regressor to $\mathbf{y}$, namely, $\boldsymbol{\varphi}_{\ell_1}$ with the following linear regression:

$$y(t) = a_m\boldsymbol{\varphi}_m(t) + \xi_m(t) \tag{27}$$

The minimum of the mean squared error can be achieved as presented in (28).

$$\frac{1}{N}\sum_{t=1}^{N}\xi_{\ell_1}^2(t) = \frac{1}{N}\sum_{t=1}^{N}(y(t) - a_{\ell_1}\boldsymbol{\varphi}_{\ell_1}(t))^2 = \min_m\left\{\frac{1}{N}\sum_{t=1}^{N}[y(t) - a_m\boldsymbol{\varphi}_m(t)]^2\right\} \tag{28}$$

The "best" matching regressor term $\boldsymbol{\varphi}_{\ell_1}$ can be found by using an orthogonal projection method that is defined by

$$\cos\chi = \frac{\mathbf{y}^T}{\sqrt{\mathbf{y}^T\mathbf{y}}}\frac{\boldsymbol{\varphi}_m}{\sqrt{\boldsymbol{\varphi}_m^T\boldsymbol{\varphi}_m}} \tag{29}$$

$$\|\boldsymbol{\varphi}_m^1\| = \|\mathbf{y}\|\cos\chi = \frac{\mathbf{y}^T\boldsymbol{\varphi}_m}{\sqrt{\boldsymbol{\varphi}_m^T\boldsymbol{\varphi}_m}} \tag{30}$$

such that

$$\sum_{t=1}^{N}\xi_m^2(t) = \|\xi_m\|^2 = \|\mathbf{y}\|^2 - \|\boldsymbol{\varphi}_m^1\|^2 = \mathbf{y}^T\mathbf{y} - \frac{(\mathbf{y}^T\boldsymbol{\varphi}_m)^2}{\boldsymbol{\varphi}_m^T\boldsymbol{\varphi}_m} \tag{31}$$

Thus,

$$\ell_1 = \arg\max_{1\leqslant m\leqslant M}\left\{\frac{(\mathbf{y}^T\boldsymbol{\varphi}_m)^2}{\boldsymbol{\varphi}_m^T\boldsymbol{\varphi}_m}\right\} \tag{32}$$

The first significant basis can be selected as $\boldsymbol{\alpha}_1 = \boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal basis can be selected as $w_1(t) = \boldsymbol{\varphi}_{\ell_1}(t)$. Set $g_1 = (\mathbf{y}^T w_1)/(w_1^T w_1)$, $ERR_1 = g_1^2\frac{(w_1^T w_1)}{(\mathbf{y}^T\mathbf{y})}$, and $\varepsilon_1(t) = y(t) - g_1 w_1(t)$.Similarly, in the second step, find a vector $\boldsymbol{\varphi}_{\ell_2}$ from the candidate regressor term family $\{\boldsymbol{\varphi}_m : 1 \leqslant m \leqslant M, m\neq\ell_1\}$. Thus, $\boldsymbol{\varphi}_{\ell_2}$ is the "best" matching regressor term to $\varepsilon_1$. According to (27) and (28), $\ell_2$ can be obtained as follows:

$$\ell_2 = \arg\max_{1\leqslant m\leqslant M, m\neq\ell_1}\left\{\frac{(\varepsilon_1^T\boldsymbol{\varphi}_m)^2}{\boldsymbol{\varphi}_m^T\boldsymbol{\varphi}_m}\right\} \tag{33}$$

Thus, the second significant basis can be chosen as $\boldsymbol{\alpha}_2(t) = \boldsymbol{\varphi}_{\ell_2}(t)$, and the second associated orthogonal basis can be selected by

$$w_2 = \boldsymbol{\alpha}_2 - \frac{w_1^T\boldsymbol{\alpha}_2}{w_1^T w_1}w_1 \tag{34}$$

and set $g_2 = (\mathbf{y}^T w_2)/(w_2^T w_2)$, $ERR_2 = g_2^2(w_2^T w_2)/(\mathbf{y}^T\mathbf{y})$, $\varepsilon_2(t) = \varepsilon_1(t) - g_2 w_2(t)$.

The $k$th significant model term can be selected as

$$\ell_k = \arg\max_{\substack{1\leqslant m\leqslant M,\\ m\neq\ell_1,\ldots}}\left\{\frac{(\varepsilon_{k-1}^T\boldsymbol{\varphi}_k)^2}{\boldsymbol{\varphi}_k^T\boldsymbol{\varphi}_k}\right\} \tag{35}$$

Set $\boldsymbol{\alpha}_k(t) = \boldsymbol{\varphi}_{\ell_k}(t)$, and orthogonalize $\boldsymbol{\alpha}_k$ with $w_1, w_2 \ldots, w_{k-1}$ as below

$$w_k = \boldsymbol{\alpha}_k - \frac{w_1^T\boldsymbol{\alpha}_k}{w_1^T w_1}w_1 - \frac{w_2^T\boldsymbol{\alpha}_k}{w_2^T w_2}w_2 - \cdots - \frac{w_{k-1}^T\boldsymbol{\alpha}_k}{w_{k-1}^T w_{k-1}}w_{k-1} \tag{36}$$

$$ERR_k = g_k^2\frac{(w_k^T w_k)}{(\mathbf{y}^T\mathbf{y})} \tag{37}$$

and calculate $g_k = (\mathbf{y}^T w_k)/(w_k^T w_k)$, and $\varepsilon_k(t) = \varepsilon_{k-1}(t) - g_k w_k(t)$. Note that the $\varepsilon_k$ can be used to form a criteria for model selection, and the $ERR_k$ is just equal to the error reduction ratio, which provides a simple but effective means for obtaining a subset of significant regressors. According to the value of $ERR_k$, the significant terms can be selected in a forward-regression manner. The detailed procedure was discussed by Billings and Wei (2005).

The procedure of model term selection can be terminated when some desired termination conditions are met. In this study, the generalized cross-validation (GCV) criterion (Billings et al., 2007) is adopted to determine the model size

$$GCV(n) = \left(\frac{N}{N-\lambda n}\right)^2 MSE(n) \tag{38}$$

where $\lambda = \max\{1, \rho N\}$ and $0 \leqslant \rho \leqslant 0.01$. The mean-square-error (MSE) in (38) is represented by $MSE = \sum_{t=1}^{N}[y(t) - \hat{y}(t)]^2/N$, where $\hat{y}(t)$ is the one-step ahead prediction produced from the associated $n$ term model. The model selection procedure will be terminated at the step when the index function GCV($n$) is minimized.

## References


Balestrassi, P.P., Popova, E., Paiva, A.P., Lima, J.W.M., 2009. Design of experiments on neural network's training for nonlinear time series forecasting. Neurocomputing 72, 1160–1178.

Ben Taieb, S., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst. Appl. 39, 7067–7083.

Bensaid, A., Hall, L., Bezdek, J., 1996a. Validity guided reclustering with applications to image segmentation. IEEE Trans. Fuzzy Syst. 4, 112–122.

Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P., Silbiger, M.L., Arrington, J.A., Murtagh, R.F., 1996b. Validity-guided (re)clustering with applications to image segmentation. IEEE Trans. Fuzzy Syst. 4.

Billings, S.A., Wei, H.L., 2005. The wavelet-NARMAX representation: a hybrid model structure combining polynomial models with multiresolution wavelet decompositions. Int. J. Syst. Sci. 36, 137–152.

Billings, S.A., Wei, H.L., Balikhin, M.A., 2007. Generalized multiscale radial basis function networks. Neural Networks 20, 1081–1094.

Biniaz, A., Abbasi, A., 2013. Fast FCM with spatial neighborhood information for brain Mr image segmentation. J. Artif. Intell. Soft Comput. Res. 3, 15–26.

Boto-Giralda, D., Diaz-Pernas, F.J., Gonzalez-Ortega, D., Diez-Higuera, J.F., Anton-Rodriguez, M., Martinez-Zarzuela, M., Torre-Diez, I., 2010. Wavelet-based denoising for traffic volume time series forecasting with self-organizing neural networks. Comput Aided Civ. Infrastruct. Eng. 25, 530–545.

Chan, S., Miranda-Moreno, L., 2013. A station-level ridership model for the metro network in Montreal, Quebec. Can. J. Civ. Eng. 40, 254–262.

Chen, S., Cowan, C.N., Grant, P.M., 1991. Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans. Nerual Networks 2. 302-209.

de Gooijer, J.G., Hyndman, R.J., 2006. 25 years of time series forecasting. Int. J. Forecast. 22, 443–473.

Hara, K., Chellappa, R., 2013. Computationally efficient regression on a dependency graph for human pose estimation. Comput. Vis. Pattern Recog. 9, 3390–3397.
Horowitz, A.J., 1984. Simplifications for single-route transit-ridership forecasting models. Transportation 12, 261–275.
Jiang, X., Zhang, L., Chen, X.M., 2014. Short-term forecasting of high-speed rail demand: a hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China. Transport. Res. Part C Emerg. Technol. 44, 110–127.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transport. Res. Part C-Emerg. Technol. 19, 387–399.
Khashei, M., Bijari, M., Hejazi, S.R., 2012. Combining seasonal ARIMA models with computational intelligence techniques for time series forecasting. Soft. Comput. 16, 1091–1105.
Kuppam, A., Copperman, R., Lemp, J., Rossi, T., Livshits, V., Vallabhaneni, L., Jeon, K., Brown, E., 2013. Special events travel surveys and model development. Transport. Lett. 5, 67–82.
Leontaritis, I.J., Billings, S.A., 1985. Input-output parametric model for non-linear systems. Part I: Deterministic non-linear systems. Int. J. Control 41, 303–328.
Li, Y., Liu, Q., Tan, S.R., Chan, R.H.M., 2016a. High-resolution time-frequency analysis of EEG signals using multiscale radial basis functions. Neurocomputing 195, 96–103.
Li, Y., Luo, M.L., Li, K., 2016b. A multiwavelet-based time-varying model identification approach for time-frequency analysis of EEG signals. Neurocomputing 193, 106–114.
Li, Y., Wei, H.L., Billings, S.A., 2011a. Identification of time-varying systems using multi-wavelet basis functions. IEEE Trans. Control Syst. Technol. 19, 656–663.
Li, Y., Wei, H.L., Billings, S.A., Liao, X.F., 2012. Time-varying linear and nonlinear parametric model for Granger causality analysis. Phys. Rev. E 85.
Li, Y., Wei, H.L., Billings, S.A., Sarrigiannis, P.G., 2011b. Time-varying model identification for time-frequency feature extraction from EEG data. J. Neurosci. Methods 196, 151–158.
Li, Y., Wei, H.L., Billings, S.A., Sarrigiannis, P.G., 2016c. Identification of nonlinear time-varying systems using an online sliding-window and common model structure selection (CMSS) approach with applications to EEG. Int. J. Syst. Sci. 47, 2671–2681.
Long, Y., Thill, J., 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Comput. Environ. Urban Syst. 53, 19–35.
Ma, X.L., Tao, Z.M., Wang, Y.H., Yu, H.Y., Wang, Y.P., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transport. Res. Part C-Emerg. Technol. 54, 187–197.
Ma, X.L., Wang, Y.H., Chen, F., Liu, J.F., 2012. Transit smart card data mining for passenger origin information extraction. J. Zhejiang Univ. Sci. C-Comput. Electr. 13, 750–760.
Ma, X.L., Wu, Y.J., Wang, Y.H., Chen, F., Liu, J.F., 2013. Mining smart card data for transit riders' travel patterns. Transport. Res. Part C Emerg. Technol. 36, 1–12.
Moody, J., Darken, C.J., 1989. Fast learning in networks of locally-tunned processing units. Neural Comput. 1, 281–294.
Ni, M., He, Q., Gao, J., 2017. Forecasting the subway passenger flow under event occurrences with social media. IEEE Trans. Intell. Transport. Syst. (in press).
Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2015. Using data from the web to predict public transport arrivals under special events scenarios. J. Intell. Transport. Syst. 19, 273–288.
Rojas, I., Valenzuela, O., Rojas, F., Guillen, A., Herrera, L.J., Pomares, H., Marquez, L., Pasadas, M., 2008. Soft-computing techniques and ARMA model for time series prediction. Neurocomputing 71, 519–537.
Shi, D., Yeung, D.S., Gao, J., 2005. Sensitivity analysis applied to the construction of radial basis function networks. Neural Networks 18, 951–957.
Si, B., Fu, L., Liu, J., Shiravi, S., Gao, Z., 2015. A multi-class transit assignment model for estimating transit passenger flows—a case study of Beijing subway network. J. Adv. Transport.
Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y.N., Lendasse, A., 2007. Methodology for long-term prediction of time series. Neurocomputing 70, 2861–2869.
Sun, Y.X., Leng, B., Guan, W., 2015. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. Neurocomputing 166, 109–121.
Taylor, B.D., Miller, D., Iseki, H., Fink, C., 2009. Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. Transport. Res. Part A Policy Pract. 43, 60–77.
Tsai, T.H., Lee, C.K., Wei, C.H., 2009. Neural network based temporal feature models for short-term railway passenger demand forecasting. Expert Syst. Appl. 36, 3728–3736.
Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where we're going. Transport. Res. Part C Emerg. Technol. 43, 3–19.
Wang, X., An, K., Tang, L., Chen, X., 2015. Short term prediction of freeway exiting volume based on SVM and KNN. Int. J. Transport. Sci. Technol. 4, 337–352.
Wei, H.L., Billings, S.A., 2006. An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon. Nonlinear Proc. Geophys. 13, 577–584.
Wei, H.L., Billings, S.A., Balikhin, M., 2004a. Prediction of the DST index using multiresolution wavelet models. J. Geophys. Res. Space Phys. 109.
Wei, H.L., Billings, S.A., Liu, J., 2004b. Term and variable selection for non-linear system identification. Int. J. Control 77, 86–110.
Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 13, 841–847.
Xue, R., Sun, D., Chen, S.K., 2015. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. Discr. Dynam. Nat. Soc.
Zhang, D.P., Wang, X.K., 2014. Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC. J. Transp. Geogr. 41, 107–115.
Zhang, G.Y., Wu, Y.G., Liu, Y.Q., 2014. An advanced wind speed multi-step ahead forecasting approach with characteristic component analysis. J. Renew. Sustain. Energy 6.
Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G., 2016. Variability in regularity: mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. PLoS One 11, e0149222.
Zhong, C., Manley, E., Arisona, S.M., Batty, M., Schmitt, G., 2015. Measuring variability of mobility patterns from multiday smart-card data. J. Comput. Sci. 9, 125–130.
Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the Beijing metropolitan area: an exploration combining smartcard and travel survey data. J. Transp. Geogr. 41, 175–183.