Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2014, Article ID 397154, 8 pages
http://dx.doi.org/10.1155/2014/397154

Research Article

Passenger Flow Prediction of Subway Transfer Stations Based on Nonparametric Regression Model

Yujuan Sun,[1] Guanghou Zhang,[2] and Huanhuan Yin[3]

[1] *College of Mechanical Engineering and Applied Electronics Technology, Beijing University of Technology, Beijing 100124, China*
[2] *Institute of Comprehensive Transportation of NDRC, Beijing 100038, China*
[3] *Research Institute of Highway Ministry of Transport, Beijing 100088, China*

Correspondence should be addressed to Yujuan Sun; yjsun@bjut.edu.cn

Received 10 November 2013; Revised 30 March 2014; Accepted 2 April 2014; Published 24 April 2014

Academic Editor: Huimin Niu

Passenger flow is increasing dramatically with accomplishment of subway network system in big cities of China. As convergence nodes of subway lines, transfer stations need to assume more passengers due to amount transfer demand among different lines. Then, transfer facilities have to face great pressure such as pedestrian congestion or other abnormal situations. In order to avoid pedestrian congestion or warn the management before it occurs, it is very necessary to predict the transfer passenger flow to forecast pedestrian congestions. Thus, based on nonparametric regression theory, a transfer passenger flow prediction model was proposed. In order to test and illustrate the prediction model, data of transfer passenger flow for one month in XIDAN transfer station were used to calibrate and validate the model. By comparing with Kalman filter model and support vector machine regression model, the results show that the nonparametric regression model has the advantages of high accuracy and strong transplant ability and could predict transfer passenger flow accurately for different intervals.

1. Introduction

Most cities in China are facing serious traffic problems, such as traffic congestion, pollution, and accidents. It is agreed that subway system is one of the efficient countermeasures to solve traffic problems. However, passenger flow is increasing dramatically with accomplishment of subway network system in big cities. As convergence nodes of subway lines, transfer stations need to assume more passengers due to amount transfer demand among different lines. Transfer facilities have to face great traffic pressure because passengers always arrive in a very short time. Consequently, pedestrian congestion or other abnormal situations will occur more easily. So, in order to avoid pedestrian congestion or warn the management before it occurs, it is very necessary to predict the transfer passenger flow to forecast pedestrian congestions.

Nonparametric regression was selected as the prediction method to forecast the passenger flow due to the fact that the authors have demonstrated the advantages of nonparametric regression over other approaches, such as Kalman filtering [1, 2] and neural networks [3, 4] in previous research efforts, based on sufficient history data.

Nonparametric regression is suitable for uncertain and nonlinear dynamic system. It is founded on chaotic system theory. Earlier work by Smith [5] found that a simple implementation of the nearest neighbor forecasting approach provided reasonably accurate traffic condition forecasts. In 1987, Yakowitz [6] suggested the using of K-nearest neighbor method in time series forecasting. The basic approach of nonparametric regression is heavily influenced by its roots in pattern recognition [7]. In essence, the approach locates the state of the system (defined by the independent variables) in a neighborhood of past, similar states. Once this neighborhood has been established, the past cases in the neighborhood are used to estimate the value of the dependent variable.

Nonparametric regression model is quite suitable for deterministic and nonlinear prediction. And it could be used in the situation without transcendental knowledge and enough historical data. It can try to find the nearest neighbor between historical data and current data, and with the nearest neighbor, it tries to predict the flow in the next interval.

The algorithm assumes that the intrinsic links of all factors are all contained in the historical data. So, the information can be obtained directly from the historical data instead of establishing an approximate model for it. In other words, the nonparametric modeling does not smooth the historical data. Therefore, the predicted effect is more precise than the parameters modeling, especially in the special events. As a free parameter, portable, and high prediction accuracy algorithm, the error of nonparametric regression is relatively small. What is more, this model is quite suitable for computer programming and can be applied to the complex environment.

The basic idea of nonparametric regression is to form a typical historical database, which is on the basis of comprehensive analysis of a large number of historical data. The historical database contains variety of traffic state trends as well as the typical rules. Each type of data in the sample library represents a traffic evolution trend. The latest traffic data collected in real-time are matched with historical data to find the nearest K-group data. The prediction of coming traffic state is determined by the nearest neighbor trends of the K-group data. Accordingly, the whole algorithm has no fixed parameters and coefficients. It can predict the next period traffic state totally based on the sample database evolution trend and the value of real-time data. Historical data series are the typical mode of traffic evolution, which play an important role in the short-term prediction. Figure 1 shows the principle of nonparametric regression theory.

Due to well prediction ability, kinds of nonparametric regression models were used to forecast traffic states gradually. In 1991, Davis and Nihan [8] used the nonparametric regression in traffic forecasting. In 1997, Smith and Demetsky [9] used the last 5 months' data to forecast the traffic flow. The definition of state vector included historical average flows; the results were better than historical average and neural network methods. Oswald et al. [10] researched how to speed up the runtime of nonparametric regression, but the accuracy was degraded. Qi and Smith [11] developed a distance metric that can be effectively used with categorical data which commonly make up traffic event data. The metric was based on the influence of variable values on a measurable objective to the purpose of selecting the nearest neighbors. When this method was incorporated in a nonparametric regression forecasting model, it was demonstrated to outperform parametric forecasting models significantly.

Tang and Gao [12] enhanced the automatic incident detection ability for forecasting traffic flows based on improved nonparametric regression algorithms and standard deviation algorithms. Turochy [13] coupled nonparametric regression with a condition monitoring method which characterized the extent to which the current traffic conditions deviate from those that may be expected based on historical data. The mean absolute percentage errors for two of the four nearest neighbor forecasting procedures were reduced. Kindzerske and Ni [14] introduced a composite approach based on nonparametric regression which was used to predict traffic conditions. The composite approach performed the nearest neighbor search for each loop detector station only using the data which are in proximity to the detector's position on the roadway. This method accommodated every detector station individually to minimize the forecast error on the entire roadway. And the composite approach can predict the onset and propagation of traffic shock waves.

Liu et al. [15] proposed a recursive nonparametric regression model and implemented it to forecast traffic flows and queue evolution in a congested actuated intersection. The model can be used to substitute traditional simulation software in the lower level of a real-time traffic control system to search the optimal control variables and then utilize the found solutions as the inputs in the simulation software in the upper level of that control system to attain the system performances. Shi and Ren [16] proposed a new method called MW model to improve the accuracy and computing speed of the nonparametric regression model when the database was too large and hard to search in short-term traffic flow forecasting. Zhang et al. [17] proposed a rule-based K-nearest neighbor nonparametric regression model to forecast large scale traffic flow of urban road networks. Rules were extracted from the historical data using Rough Set Theory, which assisted in finding the near neighbors.

Sun and Zhang [18] also proposed a selective random subspace predictor (SRSP) which was very similar to nonparametric regression model. The SRSP built selective input space based on Pearson correlation coefficients and then generated random input subspace to forecast. The method which the SRSP used to select relative variable could be used in nonparametric regression model.

From the previous literature review, it can be found that kinds of nonparametric regression models were widely used to predict traffic condition of motor vehicles. However, there were few research works related with pedestrian traffic. So, in order to test and verify the applicability of nonparametric regression in pedestrian traffic condition prediction, the K-nearest neighbor nonparametric regression model was used to forecast the transfer passenger flow of subway stations. The nonparametric regression's advantages of high accuracy and strong transplant ability are showed while being compared with Kalman filter model and support vector machine regression model.

2. Procedure of Nonparametric Regression Prediction

The application of nonparametric regression prediction contains five key steps: choosing clustering methods of historical database, the definition of state vector, the determining of similar mechanism, the choosing of the nearest neighbor mechanism, and the choosing of prediction function.

2.1. Choosing Clustering Methods of Historical Database. The first and critical step in nonparametric regression is historical data preparation, whose quality directly determines the prediction effect of nonparametric regression. What is more, the prediction effect of nonparametric regression is closely related to the choosing of clustering methods and computational time. Therefore, firstly, in order to search enough nearest neighbors, the historical database which was

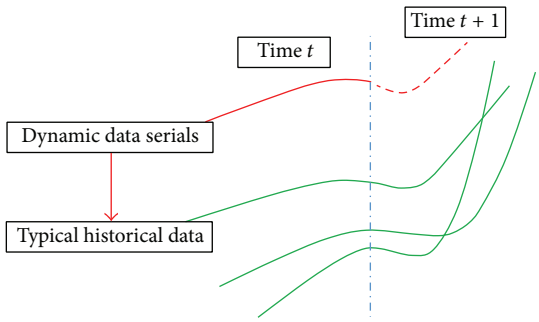

Figure 1: The schematic illustration of nonparametric regression theory.

built by clustering method must cover all state of the system. Secondly, clustering method should be able to meet the requirements in the dynamic data real-time classification and to meet the requirements of real-time, online programming. But now, traditional clustering methods take the average state vector or a single historical value as the clustering objects; it is difficult to reflect the data changing trends characteristics. Thus, the paper will focus on discussing the improvement of clustering methods and the model computational speed.

2.2. Definition of State Vector. State vector is composed of the minimum number of state variables, which are associated with the predictor variables. Because maybe there are a lot of state variables associated with predictor variables, it is necessary to properly select the number of state vectors to achieve the best balance between accuracy and computational speed.

2.3. Similar Mechanism. It is an important concept in the nonparametric regression, which means how to evaluate the similarity of the current point and the historical database. The most commonly used metric method is the Euclidean distance or weighted Euclidean distance.

2.4. Choosing the Nearest Neighbor Mechanism. As a core concept of nonparametric regression, the nearest neighbor mechanism refers to the point in the history database and how to become a close neighbor of the current point. There are two mechanisms: minimum K-nearest neighbor method and nuclear nearest neighbor method, respectively. The minimum K-nearest neighbor method means K points, whose similarity is the biggest in historical database. The nuclear nearest neighbor method refers to taking the current point as the core; all points within the radius of R become the nearest neighbor of the current point.

2.5. Prediction Function Selection. After finding the nearest neighbor points, a function needs to be used to take advantage of these points to predict the next period value. Commonly used methods are average, weighted average, and so on.

3. Improvement of Typical Model

3.1. Improvement of Historical Data Clustering. The basic procedure of nonparametric regression prediction is to compare the recent data status with the historical data and figure out the most similar data serials which would be used to predict the future data status. So, in order to provide the most similar data serial, the historical database should include enough historical information. And, in order to reflect as many trends of data serial as possible, all the historical data were stored in the database without any processing. So, the organization method of data serial in historical database determines the calculation efficiency of the prediction model. The historical database is the foundation of transfer passenger flow prediction. The core concept of the nonparametric regression is to match recent data with the historical database. From all the matches, either the K nearest matches or all the matches below a given distance threshold are located. According to the data storing system of computer science, an improved historical data organization method is proposed. This method quantifies the trend of historical data serial and sets different value for different trend which is used to cluster the historical data serials.

If the length of the data serials is n, then the historical data serial is $S_h(t) = \{S_h(t-n+1), S_h(t-n+2), \ldots, S_h(t)\}$, and the recent data serial is $S(t) = \{S(t-n+1), S(t-n+2), \ldots, S(t)\}$. Thus, the next data serial of historical database and recent status are $S_h(t+1)$ and $S(t+1)$, respectively.

If $d = \{0, 1, 2\}$ is the trend description serial of data serial, then the value of the trend description serial is

$$D_{\text{label}}(i) = \begin{cases} 0 & s(t-n+i) = s(t-n+i+1) \\ 1 & s(t-n+i) < s(t-n+i+1) \\ 2 & s(t-n+i) > s(t-n+i+1), \end{cases} \quad i = 1, 2, \ldots, n-1. \tag{1}$$

The number of clustering types of historical database is

$$C_{\text{no}} = 3^{n-1}. \tag{2}$$

For one data serial, the clustering label is

$$C_{\text{label}} = D_{\text{label}}(1) \times 3^{n-2} + \cdots + D_{\text{label}}(n-2) \times 3^1 + D_{\text{label}}(n-1) \times 3^0. \tag{3}$$

Figure 2 is the trend of one data serial with length of 4. Based on (2), the number of clustering types in historical database is

$$C_{\text{no}} = 3^{4-1} = 27. \tag{4}$$

And the clustering label is

$$C_{\text{label}} = 1 \times 3^2 + 1 \times 3^1 + 2 \times 3^0 = 14. \tag{5}$$

3.2. The Selection of Data Serial. Based on the experimental analysis, the neighbor data are chosen as the state vector. The vector contains four current transfer passenger flow trend

Table 1: Precision of nonparametric regression forecasting model.

Performance	Time					
	7:00–9:00			17:00–19:00		
	1 minute	3 minutes	5 minutes	1 minute	3 minutes	5 minutes
Average relative error	12.20%	8.10%	6.30%	11.80%	6.00%	4.00%
Maximum relative error	42.00%	35.00%	23.00%	31.00%	24.00%	13.00%
Equalization coefficient	0.91	0.96	0.96	0.93	0.96	0.98

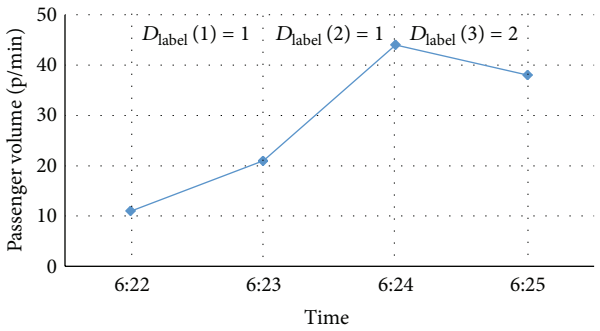

Figure 2: Illustration of trend label of state vector for nonparametric regression.

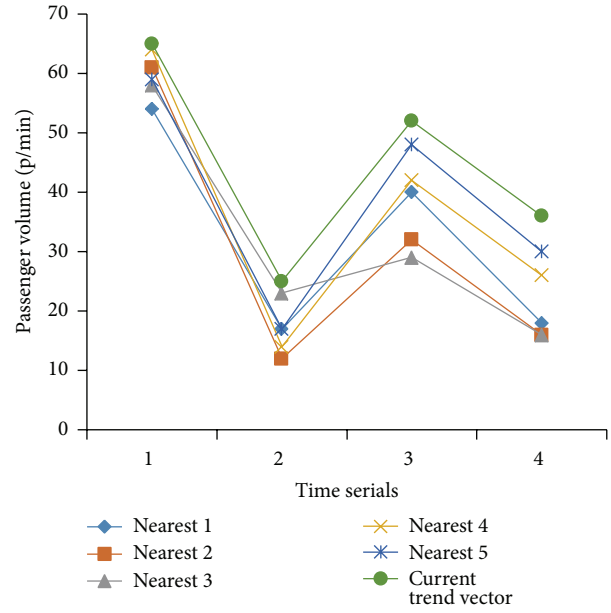

Figure 3: Comparison of state vector of prediction and similar neighborhood.

data and five historical transfer passenger flow trend data. Four neighbor data are selected as data serial. The prediction model calculates the clustering label based on the trend of the four neighbor data and searches for the most similar data serials from history database. Then, the future data status is predicted according to the next trend of the most similar data serials.

3.3. The Similar Mechanism. The Euclidean distance is used to calculate the similar level between the recent data serial and the historical data serials. The equation is

$$d_i = \left(\frac{1}{4}\left(S(t-3) - S_h(t-3)\right)^2 + \frac{1}{4}\left(S(t-2) - S_h(t-2)\right)^2 + \frac{1}{4}\left(S(t-1) - S_h(t-1)\right)^2 + \frac{1}{4}\left(S(t) - S_h(t)\right)^2\right)^{1/2}. \quad (6)$$

Except for the Euclidean distance, the weights of the most similar historical data serials are also used in the prediction model. As shown in (7), β_i is the weight of the most similar historical data serial i. The bigger the β_i is, the more remarkable the influence level on the prediction result of data serial i is:

$$\beta_i = \frac{d_i}{\sum_{j=1}^{k} d_j}, \quad (7)$$

where k is the number of the most similar data serials.

3.4. The Selection of Neighbor Mechanism. K-neighbor mechanism is used to select the nearest neighbors. K represents for the numbers of nearest neighbors which are selected from historical database, and has close relation to the database's character. Based on the previous research results [13, 14, 19], the K is 5.

3.5. The Improvement of Selection Model. The weighted average method based on the reciprocal of the matching distance is chosen as the prediction function. The shorter distance point is the more similar point. Then, the weighing is bigger. For most nonparametric regression prediction models, the next value of the most similar historical serial is used as the prediction value of recent data serial. The next value and weighted coefficient based on the historical data are used to predict the transfer passenger flow in the prediction algorithm. In the state vector of the prediction model, the historical data of the current time and the nearest time are used to identify different prediction coefficient, and the historical data of the next trend are used to calculate the prediction data directly.

However, due to reasons such as the lack of historical data or abnormal flow, the next value of recent data serial may change dramatically, taking Figure 3 as an example. So, in order to improve the prediction accuracy, the amending

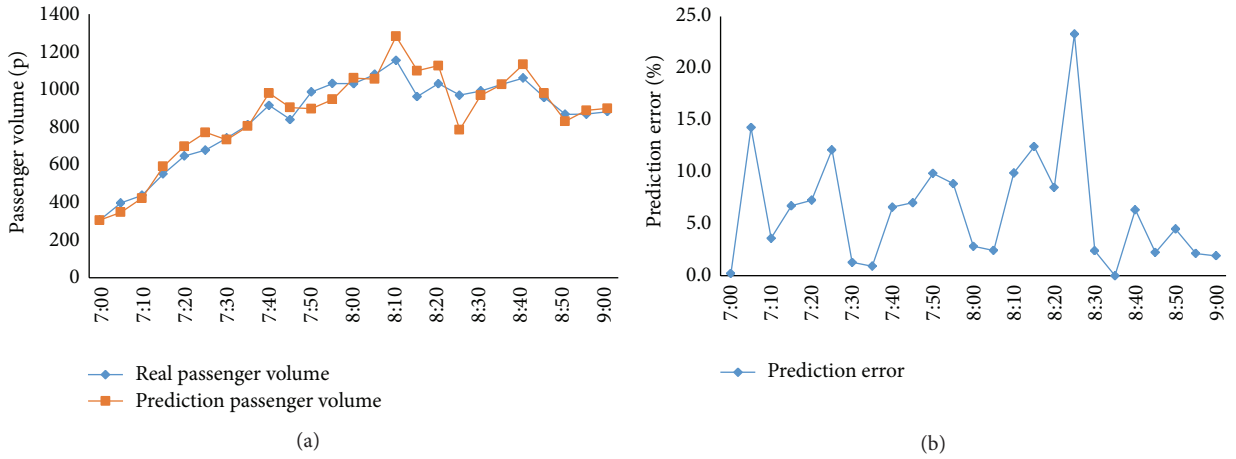

FIGURE 4: Forecasting result for each 5 minutes in morning peak hour using nonparametric regression.

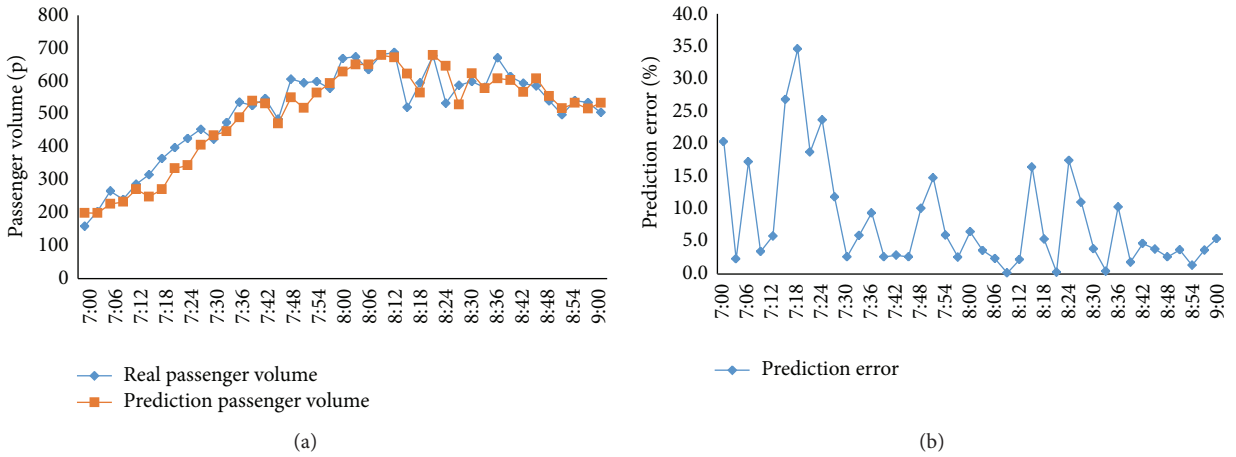

FIGURE 5: Forecasting result for each 3 minutes in morning peak hour using nonparametric regression.

coefficient with average value of recent data serial is proposed. The improved model is

$$s(t+1) = \sum_{i=1}^{K} \frac{\beta_i s_{hi}(t+1)\,\overline{x}}{\overline{x}_{hi}}, \tag{8}$$

where $\overline{x}$ is the average value of recent data serial given as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} s(t-i+1) \tag{9}$$

and $\overline{x}_{hi}$ is the average value of the neighbor data serial i given as

$$\overline{x}_{hi} = \frac{1}{n}\sum_{j=1}^{n} s_{hi}(t-j+1). \tag{10}$$

4. Application

In order to test the accuracy of the prediction model, the transfer passenger flow of XIDAN station was used to calibrate the model. The historical database was built with the transfer passenger flow from July 26 to August 25, 2011. The prediction data were the passenger flow of August 25, 2011. The prediction results are illustrated in Figure 4 to Figure 9.

4.1. Forecasting Results of Peak Hours from 7:00 to 9:00. See Figures 4, 5, and 6.

4.2. Forecasting Results of Peak Hours from 17:00 to 19:00. See Figures 7, 8, and 9.

The prediction performance for different time and intervals is shown in Table 1. It is obvious that the improved nonparametric regression model has very high prediction accuracy. The maximum average relative error is less than 15%. To compare the applicability of different prediction model, the Kalman filter model and support vector machine regression model are chosen to predict the transfer passenger flow. Figures 10, 11, and 12 show the comparison of prediction capability of three different models. Compared with the Kalman filter model and support vector machine regression model, the accuracy of the predicted transfer passenger flow

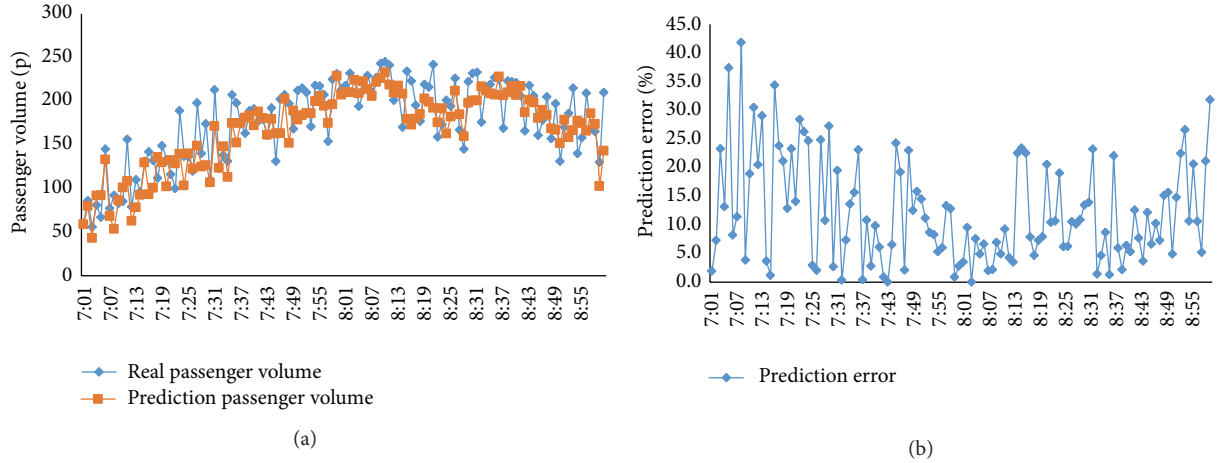

Figure 6: Forecasting result for each 1 minute in morning peak hour using nonparametric regression.

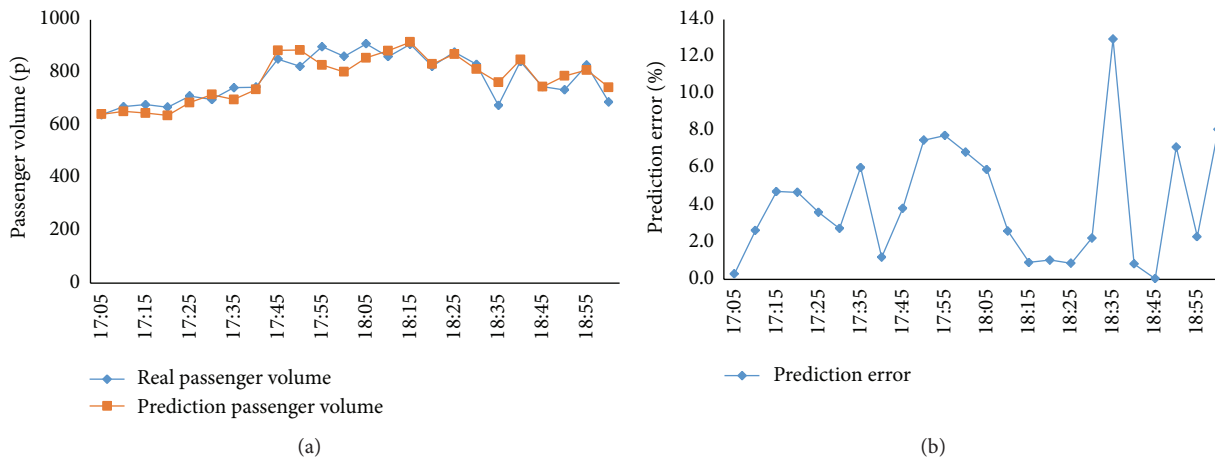

Figure 7: Forecasting the result for each 5 minutes in evening peak hour using nonparametric regression.

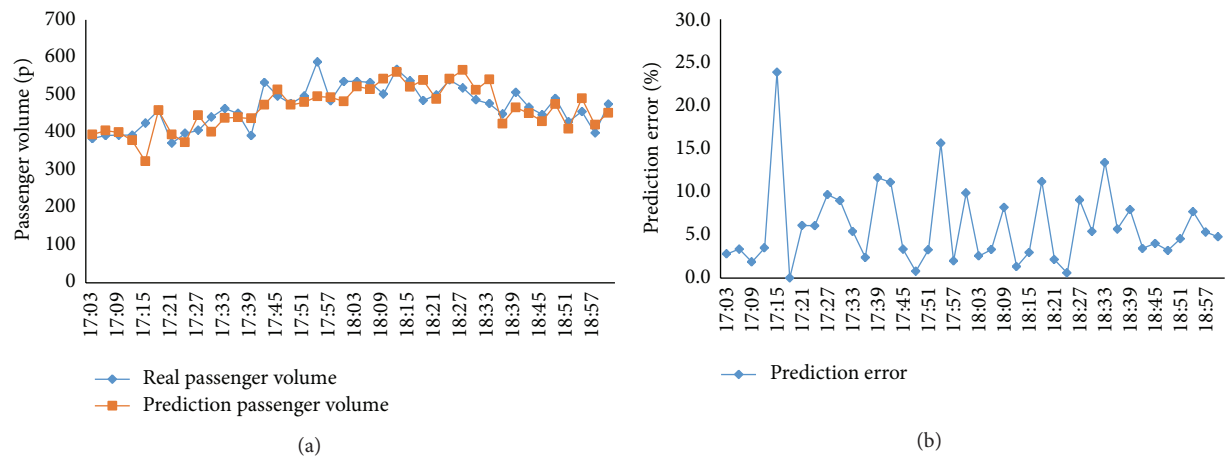

Figure 8: Forecasting the result for each 3 minutes in evening peak hour using nonparametric regression.

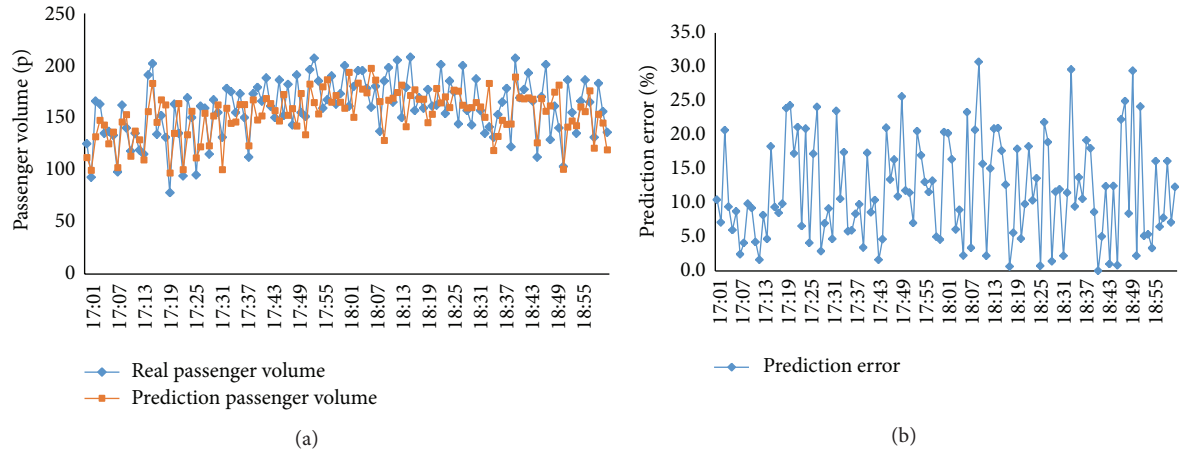

Figure 9: Forecasting the result for each 1 minute in evening peak hour using nonparametric regression.

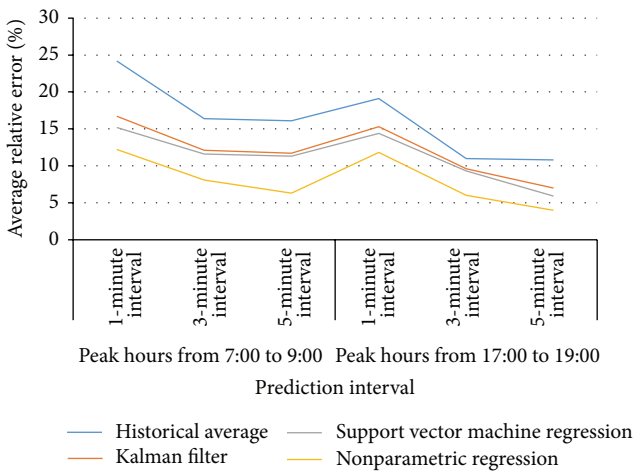

Figure 10: Comparison of average relative error for different forecasting models.

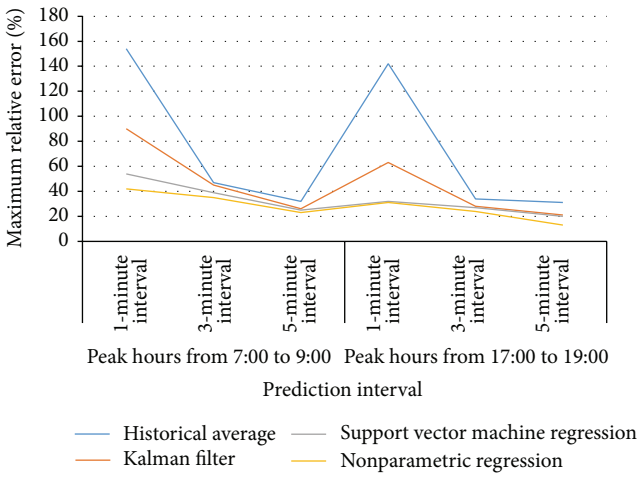

Figure 11: Comparison of maximum relative error for different forecasting models.

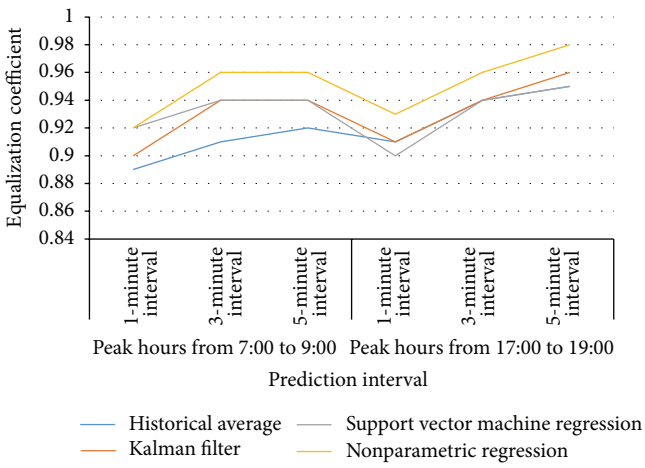

Figure 12: Comparison of equalization coefficient for different forecasting models.

and the stability of the error for the improved nonparametric regression model have been improved significantly. So, the improved nonparametric regression prediction model can be used in real application.

5. Conclusions

As a convergence node of subway lines, transfer stations need to assume more passengers due to amount transfer demand among different lines. So, it is really very necessary to predict the transfer passenger flow to avoid pedestrian congestion or warn the management before it occurs.

Based on nonparametric regression theory, a transfer passenger flow prediction model was proposed. And data of transfer passenger flow for one month in XIDAN transfer station were used to calibrate and validate the model. The results show that the model could predict transfer passenger flow accurately for different intervals. What is more, the prediction accuracy is also much better than Kalman filter model and support vector machine regression model. The

bigger the interval is, the more accurate the prediction result is. The maximum average relative error is 12.20%, which means that the prediction model can be used in real application.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B*, vol. 18, no. 1, pp. 1–11, 1984.

[2] J. Whittaker, S. Garside, and K. Lindveld, "Tracking and predicting a network traffic process," *International Journal of Forecasting*, vol. 13, no. 1, pp. 51–61, 1997.

[3] S. C. Chang, S. J. Kim, and B. H. Ahn, "Traffic-flow forecasting using time series analysis and artificial neural network: the application of judgmental adjustment," in *Proceeding of the 3rd IEEE International Conference on Intelligent Transportation Systems*, Dearborn, Mich, USA, October 2000.

[4] L. R. Rilett and D. Park, "Direct forecasting of freeway corridor travel times using spectral basis neural networks," *Transportation Research Record*, no. 1752, pp. 140–147, 2001.

[5] B. L. Smith, *Forecasting freeway traffic flow for intelligent transportation system applications [Doctoral dissertation]*, Department of Civil Engineering, University of Virginia, Charlottesville, Va, USA, 1995.

[6] S. Yakowitz, "Nearest-neighbour methods for time series analysis," *Journal of Time Series Analysis*, vol. 8, no. 2, pp. 235–247, 1987.

[7] M. Karlsson and S. Yakowitz, "Rainfall-runoff forecasting methods, old and new," *Stochastic Hydrology and Hydraulics*, vol. 1, no. 4, pp. 303–318, 1987.

[8] G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," *Journal of Transportation Engineering*, vol. 117, no. 2, pp. 178–188, 1991.

[9] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: comparison of modeling approaches," *Journal of Transportation Engineering*, vol. 123, no. 4, pp. 261–266, 1997.

[10] R. K. Oswald, W. T. Scherer, and B. L. Smith, "Traffic flow forecasting using approximate nearest neighbor nonparametric regression," Research Report Uvacts-15-13-7, Center for transportation studies at the University of Virginia, 2001.

[11] Y. Qi and B. L. Smith, "Identifying nearest neighbors in a large-scale incident data archive," *Transportation Research Record*, no. 1879, pp. 89–98, 2004.

[12] S. Tang and H. Gao, "Traffic-incident detection-algorithm based on nonparametric regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 38–42, 2005.

[13] R. E. Turochy, "Enhancing short-term traffic forecasting with traffic condition information," *Journal of Transportation Engineering*, vol. 132, no. 6, pp. 469–474, 2006.

[14] M. D. Kindzerske and D. Ni, "Composite nearest neighbor nonparametric regression to improve traffic prediction," *Transportation Research Record*, no. 1993, pp. 30–35, 2007.

[15] K. Liu, R. Ghaman, and F. Xiang, "An adaptive procedure for prediction of traffic conditions at signalized intersection," in *Proceedings of the 14th World Congress on Intelligent Transport Systems (ITS '07)*, Beijing, China, October 2007.

[16] X. Shi and Q. Ren, "The advanced nonparametric model for short-Term traffic volume forecasting," in *Proceedings of the 11th International Conference of Chinese Transportation Professionals (ICCTP '11)*, pp. 1442–1453, Nanjing, China, August 2011.

[17] L. Zhang, Z. Peng, D. Sun, and X. Liu, "A novel rule-based traffic state forecasting approach for large-scale road networks," *Transportation Research Record*, vol. 2279, pp. 3–11, 2012.

[18] S. Sun and C. Zhang, "The selective random subspace predictor for traffic flow forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 367–373, 2007.

[19] S. Clark, "Traffic prediction using multivariate nonparametric regression," *Journal of Transportation Engineering*, vol. 129, no. 2, pp. 161–168, 2003.