IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Research on forecast method of railway passenger flow demand in pre-sale period

To cite this article: Zhengzheng Wei and Xinghua Shan 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **563** 052080

View the article online for updates and enhancements.

This content was downloaded from IP address 115.219.78.107 on 12/10/2019 at 21:15

IOP Conf. Series: Materials Science and Engineering **563** (2019) 052080 doi:10.1088/1757-899X/563/5/052080

Research on forecast method of railway passenger flow demand in pre-sale period

WEI Zhengzheng[1], SHAN Xinghua[2]

[1]Railway Technology Research College, China Academy of Railway Science Corporation Limited, Beijing

[2]Railway Technology Research College, China Academy of Railway Science Corporation Limited, Beijing

[1]ninna.weina@163.com, [2]shanxinghua@vip.sina.com

Abstract. The railway demand scientific forecasting is a rapid response mechanism for the railway sector, rational layout train plan for achieving definitive dynamic adjustment, and the basis for carrying out revenue management. In the railway industry, with changes of the seasons, holidays, emergencies and strategic competitor (such as airline routes for a certain period of reduced discount), passenger flow demand changes. Based on actual demand, this paper proposes an adaptive and flexible method of the railway passenger flow forecasting, which can be adapted to the above factors.

1. Introduction

With the accelerating pace of China's high-speed railways' "going out" strategy, the scientific design and management of high-speed railway passenger transport products need to be improved. Among them, short-term passenger flow forecasting work is crucial in the temporary adjustment of train operation plan, dynamic adjustment of ticket amount, and is an important basis for revenue management. Especially after entering the pre-sale period, whether it is the automatic pre-scoring of the ticket amount or the revenue management, in order to maximize the income, it is necessary to continuously redistribute the votes of different sections. An important basis for a reasonable allocation is to accurately predict changes in passenger flow during the pre-sale period.

For the forecast of reservation demand during the pre-sale period, the commonly used method is mainly the incremental method. The traditional pick up prediction is a forecasting method for incremental bookings. The idea is to first predict the incremental demand in a short period of time, and then accumulate these increments and the total forecast demand at a certain point in the future will be derived at last.

2. Traditional pick up booking demand forecasting

Pick up is a predictive method by studying how much the future demand increases. The classical Pick Up is a cumulative data matrix that predicts future aggregate demand by adding the total demand at the current observation point to the future average demand. For the high-speed railway train OD passenger flow forecast, its idea is to predict the daily booking request during the pre-sale period, and then add these predetermined requests to get the total passenger flow forecast value of the boarding date. As shown in Equations 1 and 2:

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

$$F_T(i+1) = X_C(i+1) + \frac{I_C(1)+\cdots+I_C(i)}{i} \quad (1)$$

$$R_C(i+1) = \frac{X_{C-1}(i+1)-X_C(i+1)}{X_c(i+1)} \times 100\% \quad (2)$$

Among them

$F_T(i+1)$: When the high-speed train is on sale, the forecast of the total demand for the $i+1$ corresponding date;

$X_T(i)$: When the high-speed train is on sale, the actual total demand for i corresponds to the date;

$X_C(i)$: The current observation point, that is, the total actual reservation amount of the i corresponding date on the C day of the pre-sale period;

$I_C(i) = X_T(i) - X_C(i)$: The demand increase of the high-speed train at the current time point C to the time of the sale.

In order to better explain the idea of the pick up method, a further description with a numerical example is presented here. The training set of this paper selects the G11 train from Beijing South station to Shanghai Hongqiao station. The boarding date is from January 1, 2017 to March 19, 2017. The booking data for the pre-sale period during the 19th month is forecasted for the booking data for the pre-sale period from March 20, 2017 to March 25, 2017 in the test set. As shown in Table 1: The table shows the booking amount of the G11 train from Beijing South to Shanghai Hongqiao during the pre-sale period between January 1, 2017 and March 25, 2017. The sequence data 1, 2, 3, ... 10 in the table respectively indicate the reservation time of 10 days in advance of the reservation. Since 2017, the pre-sale of the railway has been changed to 60 days. Only the booking data for the 10 days prior to the pre-sale period is listed in this table. The number in the table represents the amount of reservations that have occurred, and "?" indicates that there are no subscription requests to be predicted that have yet to occur. The traditional incremental model selects the average value of the reservation amount corresponding to the boarding date of the prediction target as the predicted value of the target booking amount, that is, the average value of the known data of each column in the table is used as the prediction of the non-recurring demand. For example, when you want to predict the booking amount on the date of March 20, 2017, the booking data will be 1, 1, 8, and 8 on the date of the arrival date from January 1, 2017 to March 19, 2017. The average value obtained by the addition is used as the predicted value of its daily booking demand. Another approach is to improve the traditional incremental model, which also uses the known booking data on the booking matrix as input to the forecasting demand forecast, but instead of simply using the average as the forecast for the corresponding booking demand, an index is used. A smooth method calculates the prediction of the corresponding subscription requirement. In practice, this method is proved to be significantly better than the traditional incremental model.

Table 1. Demand booking during the pre-sale period

G11 Beijing Nan-Shanghai Hongqiao pre-sale period booking amount										boarding Date
10	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	
9	19	32	28	20	9	6	4	5	1	2017-01-01
3	3	5	3	6	4	7	33	39	9	2017-01-02
2	1	2	4	8	3	7	11	26	8	2017-01-03
9	1	13	18	21	30	10	35	76	12	2017-01-04
…	…	…	…	…	…	…	…	…	…	……
?	?	?	?	?	?	?	?	?	?	2017-03-20
?	?	?	?	?	?	?	?	?	?	……
?	?	?	?	?	?	?	?	?	?	2017-03-25

3. Exponential smoothing based on clustering increments

3.1. Model description

When the high-speed train tickets enters the pre-sale period, the demand for continuous pre-sorting or revenue management needs to be continuously optimized, and it is necessary to continuously predict the future reservation demand during the pre-sale period based on the existing reservation amount. This paper firstly analyzes the historical reservation time series curve of a specific demand, divides the similar reservation curves into one class, and then divides the time series of reservation demand to be predicted into a certain class according to the KNN algorithm. Finally, for future booking needs, all the booking samples in the class to which this booking time series belongs are used as training samples, and an exponential smoothing is used to predict future booking needs. The framework is shown in Figure 1:

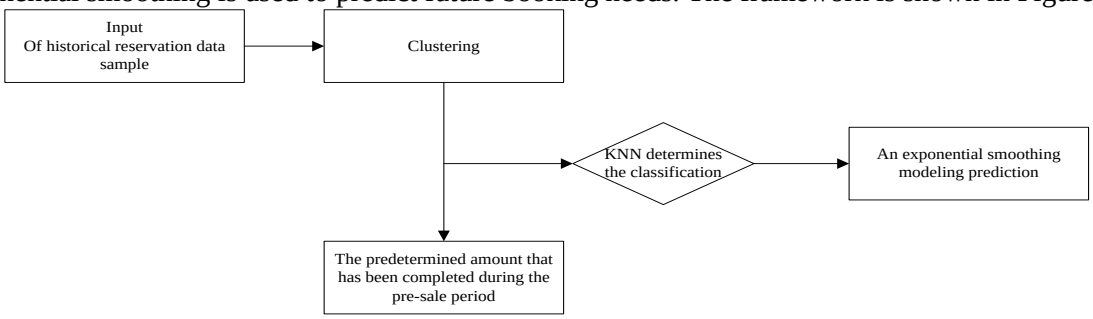

Figure 1. The framework of the model

3.2. Booking time series clustering

Calculating the similarity measure of time series is a hot issue in time series data mining research. The problem of similarity metrics in the study of time series is usually focused on the study of Euclidean distance improvement algorithms. Although Euclidean distance is continuously improved and extended by many researchers, Euclidean distance can support time series amplitude translation and scaling, but it is for time series. Time axis stretching and bending seem to be powerless. To this end, Berndt and Clifford [1] introduced the Dynamic Time Warping (DTW) distance widely used in speech recognition into time series similarity metrics. Firstly, the historical pre-sale time series of the target predicted train OD are clustered. The literature [2, 3] discusses the cluster analysis algorithm in machine learning in detail. At present, there are many methods for clustering. According to different basic ideas, they can be roughly divided into partitioning method, layering method, density-based method, grid-based method and model-based method. Take the G11 train from Beijing South to Shanghai Hongqiao from January 1 to March 19, 2017 as an example. Select the time-to-day booking time for the boarding date in this time period, use the DTW to calculate the distance method, and use the condensed The hierarchical clustering method clusters 78 time series curves as shown in Figure 2.

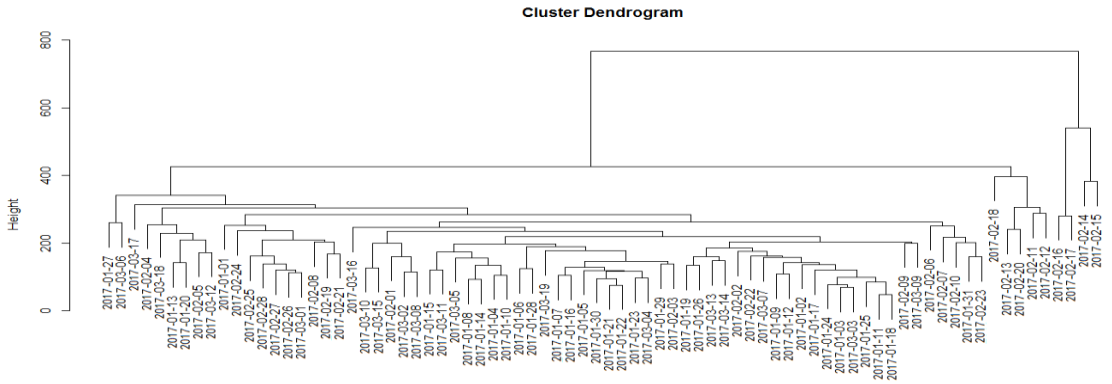

Figure 2. Example of time series curve clustering

The closer the leaf nodes are to the cluster nodes, the more similar the points are, for example, the arrival date of 2017-01-13, 2017-01-20, 2017-02-05, 2017-03-12. If we want to predict 2017-03-12 as the booking value of the boarding date within the pre-sale period, we can choose the booking value of the booking time series with the arrival date of 2017-01-13 and 2017-01-20 as the booking demand. Compared with the traditional incremental prediction method, the algorithm of this paper mainly classifies the training samples by clustering the scheduled time series, and provides support for the training set of the target demand reservation prediction in the next section 3.3.

3.3. Classification of reservation requirements

After the high-speed railway train tickets enters the pre-sale period, it is necessary to judge the classification of the reservation time series by the amount of reservations sold for each day, and then use the subscription amount of the classification as a training sample to predict the reservation amount of future demand. With the development of machine learning and data mining in recent years, the classification of time series has been widely concerned. Many time series classification algorithms have been proposed, including the decision tree proposed by Rodriguez & Alonso [4], the neural network proposed by Nanopoulos & Manolopoulos [5], Bayesian classification, SVM classification and so on. Practice has proved that the 1-Nearest-Neighbor algorithm based on DTW distance is the best method in the time series classification algorithm.

The pre-sale time series of the G11 train from January 1, 2017 to March 19, 2017 for the G11 train from Beijing South to Shanghai Hongqiao will be used as the training set for the cluster. It is now judged that the classification of test date scheduled time series on the 20th and the 21st of March, when the forecast target enters the pre-sale period, because the pre-sale period has just been adjusted to 60 days, and the existing pre-sale fare is implemented as a single static fare, usually When the boarding date is non-holiday, the booking amount for the 30 days before the pre-sale period is generally zero. Therefore, the order booking amount will be obtained from the 31st day from the pre-sale period, and two booking time series that have occurred will be calculated separately. The DTW value of the subscription time series corresponding to the training set is shown in Figure 3:

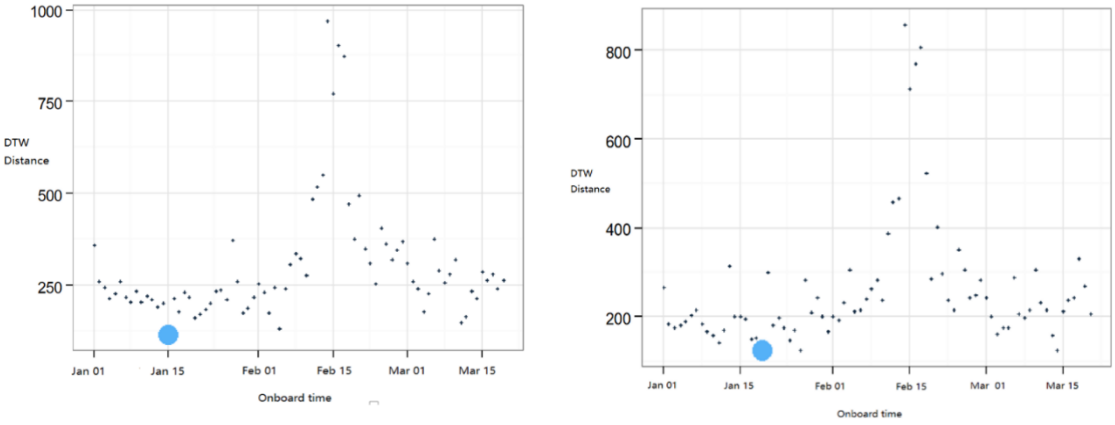

Figure 3. Example of determining classification

Obviously, the boarding dates corresponding to the points where the boarding date is the smallest in the booking time series DTW on March 20th and March 21st are January 15th and January 19th, respectively. According to the time series clustering example proposed in Section 3.2, the classification corresponding to the above two boarding dates is searched, and the training set for predicting the target presale reservation data in each category is selected according to the experimental result, and it will be used in Section 3.4.

3.4. Booking demand forecast

The running route of high-speed railway is more stable than the existing trains, and rarely adjusts frequently while the interval is short. Therefore, for the reservation demand forecast of high-speed

railway trains, the future reservation demand can be predicted according to the high-speed train history reservation demand. First, according to Sections 3.2 and 3.3, the target booking time series to be predicted are classified into similar subscription time series classes, and then each type of historical booking amount is used as training data, and an exponential smoothing method is used to calculate the subscription requirements that have not occurred to make the predictions. Through the discriminant classification method of subscribing time series in Section 3.3, the target reservation time series to be predicted is firstly classified into the corresponding classification obtained in Section 3.2, and then the reservation amount of the boarding date corresponding to the classification reservation time series is used as the target demand reservation prediction. The training data is such that the obtained training data is similar to the data to be predicted. The data with such characteristics is generally more suitable for solving the target prediction problem by using an exponential smoothing method, and the exponential smoothing prediction for the pre-sale period subscription amount is as follows. As shown in (3):

$$F_T(i+1) = \alpha X_T(i) + (1-\alpha)F_T(i) \quad (3)$$

Among it， $i \in S$， S denotes a clustering reservation time series classified as a class in Section 4.1;

$F_T(i+1)$：When the high-speed railway is available for sale, the first $i+1$ booking time series corresponds to the pre-sale of the T day reservation.

$X_T(i)$: When the high-speed railway is available for sale, the order of i for the T reservations corresponds to the actual booking amount of the pre-sale day;

$F_T(i)$: When the high-speed railway is available for sale, the booking schedule of the i is corresponding to the forecast of the pre-sale T days；

α: The smoothing coefficient, also known as the weighting factor, ranges from $0 < \alpha < 1$.

Use the algorithm proposed in Section 3.2 to cluster the daily booking curves from January 1, 2017 to March 19, 2017, and use the discriminant algorithm used in Section 3.3 to discriminate the subscription time series that have occurred in the test set. In the classification of the training set, the classification of each test date in the test set is as shown in Table 2, and the time series training set in the classification is used as the feature input of the test set reservation quantity prediction algorithm, and the test is performed by using an exponential smoothing method. The daily booking data is concentrated for prediction, and the prediction results are compared with the traditional incremental model and the incremental model based on exponential smoothing improvement.

Table 2. Classification of test sets

Test Set	Mar 20	Mar 21	Mar 22	Mar 23	Mar 24	Mar 25
Affiliation	Jan 15	Jan 26	Jan 6	Mar 5	Jan 6	Jan 22
Training set	Mar 5	Mar 19	Mar 28	Mar 13	Mar 18	Mar 5
classification	Mar 11	Mar 14	Mar 18	Mar 14	Mar 22	Mar 12

If it is necessary to predict the booking data for the pre-sale period on March 20, the pre-sales data for January 15, March 5, and March 11 will be used as the input training set for the predictive model. To predict the booking data on the day of March 20th, the bookings of 10, 10, and 2 on January 15th, March 5th, and March 11th will be used as input to an exponential smoothing model to obtain prediction results. Similarly, the predicted results of other pre-sale advance dates can be obtained.

The proposed algorithm and the improved incremental-based exponential smoothing method and the traditional incremental method are used to predict the daily booking amount during the pre-sale period on March 20. The comparison between the predicted value and the actual booking amount is shown in Figure 4:

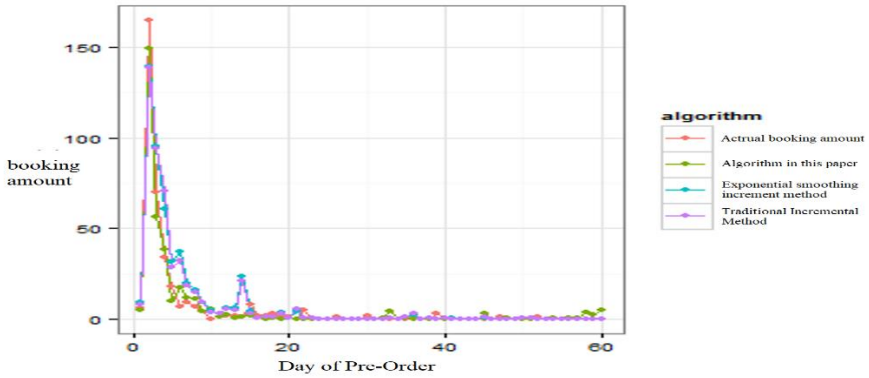

Figure 4. Comparison of the predicted and actual values

When the pre-sale period is just entered, the algorithm has a smaller deviation from the actual reservation value than other algorithms, and the booking amount increases with the approaching date of the boarding. This algorithm shows obvious advantages compared with other algorithms. Basically matches the actual booking value. Calculate the RMSE values of the algorithms for the pre-sale for 60 days from March 20 to March 25 in the test set, as shown in Table 3:

Table 3. RMSE values of the various algorithms of the test set

RMSE	Mar 20	Mar 21	Mar 22	Mar 23	Mar 24	Mar 25
Algorithm	3.7	5.7	8	11	4.5	5.4
Exponential method	8.1	9.2	8.7	14	8.3	9.8
Traditional method	8.2	12	10	16	12	11

The corresponding column contrast chart is shown in Figure 5:

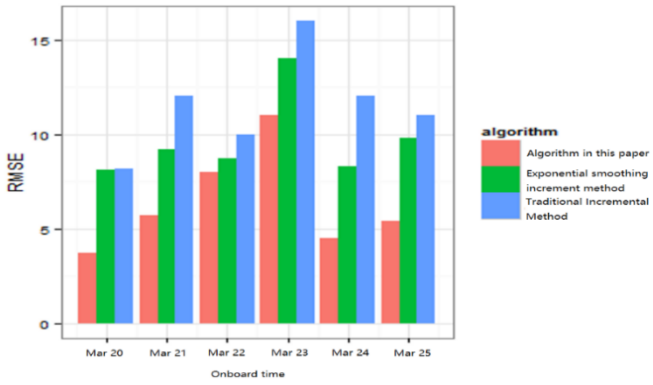

Figure 5. Comparison of the RMSE of the algorithm and each algorithm in this paper

Compared with the RMSE values of the days in the test set, the proposed algorithm is superior to the results predicted by the incremental method based on exponential smoothing and the traditional incremental method. Among them, the forecast results of the test set on March 20 were the most ideal, and the forecast results on March 23 were poor. The incremental method based on exponential smoothing improvement is superior to the traditional incremental method with a smaller advantage.

4. Summary

According to the characteristics of high-speed railway ticket sales entering the pre-sale period booking amount, firstly, the historical reservation time series of the target predicted vehicle number are clustered by DTW distance, and then the 1-Nearest-Neighbor algorithm based on DTW distance is used to select

the target reservation time series to be predicted. Finally, the corresponding subscription data in the classified category is used as the training value of the exponential smoothing algorithm to obtain the predicted value of the reserved amount. The experiment proves that the prediction model proposed in this paper for the pre-sale period of each day is significantly better than the traditional model method.

Acknowledgment
Fund Project: Advanced Rail Transit Project of National Key R&D Program (2018YFB1201404)
Fund Project: Science and Technology Research and Development Program of China Railway Corporation (2017X004-C)

Reference

[1] Donald J. Berndt, James Clifford. Using Dynamic Time Warping to Find Patterns in Time Series [C]. In Proceedings of the KDD Workshop, Seattle, WA.1994:359- 370.
[2] He Ling, Wu Lingda, Cai Yichao. Overview of Clustering Algorithms in Data Mining. Computer Application Research, No.1, 2007.
[3] R.O. Duda, PE Hart, D G Stork. Pattern Classification (2nd Edition) [M]. NewYork: Wiley, 2001. 452458.
[4] Rodríguez, J.J. & Alonso, C.J. (2004). Interval and dynamic time warping-based decision trees. In Proceedings of the 2004 ACM symposium on Applied computing (SAC), pp.548-552.
[5] Nanopoulos, A., Alcock, R., & Manolopoulos, Y. (2001).Feature-based Classification of Time-series Data. International Journal of Computer Research, pp. 49-61.