Research Article

Short-Term Forecasting of Railway Passenger Flow Based on Clustering of Booking Curves

Minshu Ma, Jun Liu, and Jingjia Cao

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Minshu Ma; mshma@bjtu.edu.cn

Received 13 January 2014; Accepted 17 March 2014; Published 13 April 2014

Academic Editor: Jonathan N. Blakely

Copyright © 2014 Minshu Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For railway companies, the benefits from revenue management activities, like inventory control, dynamic pricing, and so forth, rely heavily on the accuracy of the short-term forecasting of the passenger flow. In this paper, based on the analysis of the relevance between final booking amounts and shapes of the booking curves, a novel short-term forecasting approach, which employs a specifically designed clustering algorithm and the data of both historical booking records and the bookings on hand, is proposed. The empirical study with real data sets from Chinese railway shows that the proposed approach outperforms the advanced pickup model (one of the most popular models in practice) during the early and middle stages of booking horizon when bookings are not concentrated in the final days before departure.

1. Introduction

Short-term forecasting is an essential and fundamental function of the revenue management (RM) systems, which was originated from the American air transportation industry and has been widely spread to some other areas, such as railway transportation, hotel, and power supplier. With the short-term forecasting of the passenger demands, the railway companies as well as the airlines are able to make decisions for the RM tasks such as inventory control, dynamic pricing, and overbooking. The more accurate forecasts are, the more reasonable decisions could be.

Generally, there are three methodologies in dealing with the short-term forecasting problems for passenger transportation of railway companies or airlines, based, respectively, on historical booking models, advanced booking models, and combined booking models. The differences lie in what information they depend on to perform forecasts. For the historical booking models, the final booking amounts of the departed trains or air flights, or the passenger volumes in the past, are what the forecasting depends on. Advanced booking models, unlike this, employ the reservation data corresponding to the passenger flow to be predicted. In other words, to forecast how many passengers have the demand to

be transported at a certain time in the future, the historical booking models focus on how many passengers were transported, while advanced booking models focus on how many passengers have booked. The combined booking models, as prompted by the name, utilize both types of information and, usually, integrate outcomes from multiple models. The above methodologies are also applicable to the forecasting of booking demands for hotels. The papers in both fields provide valuable references for research works in either area.

Among the three categories of models mentioned above, historical booking models have longer history of being studied and applied. The classical time series based models, like the Box-Jenkins model, just belong to this category. In recent years, the combination of these models with some new techniques, like the artificial neural networks, was studied and some meaningful results have been reported [1, 2].

Advanced booking models, based on the information of bookings on hand, were designed specifically for RM oriented forecasting. Regression and pickup models are most common in this category. While regression models try to figure out the quantitative relationships between the final bookings and the bookings on hand in a regular way [3, 4], pickup models, exploiting the unique characteristics of reservation data, forecast the bookings to come in the future by aggregating the

possible increments of bookings that are assumed to be analogous to the increments corresponding to earlier departure date. Research by Wickham [5] shows that the pickup models consistently outperform several other prevalent short-term demand forecasting models for air transportation, including the regression and time series models, in various scenarios.

Since pickup models require the reservation data corresponding to earlier departure date, the selection of those data became an important issue. Irwin introduced the practice within KLM Royal Dutch Airlines [6], where the booking curves of years in the past are utilized in combination with some other information like the possible trend observed in previous months. Tsai proposed a three-stage model based on the idea that, when applying the advanced booking models, the reservation data corresponding to earlier departure date should be chosen according to their similarity with the reservation on hand corresponding to the departure date on which the final arrivals are to be forecasted [7]. That is also the motivation for the research in this paper.

In this paper, clustering, as one of the key data mining techniques, is used to classify the historical reservation data. In the past, clustering as well as some other data mining techniques has been used in forecasting for urban traffic flow [8], air transportation [9], sales in retail merchandising [10], electricity load [11], and so on [12–14]. But they neither are used for short-term forecasting of traffic demand nor are used in connection with advanced booking models.

The organization of this paper is as follows. After the brief review of related works, the relevance between final booking numbers and booking curve shapes will first be analyzed in Section 2. This analysis underlies the clustering algorithm as well as the prediction model introduced in the following two sections, respectively. In Section 5, the data of an empirical study, the method of the study and the analysis on the results of the study that employs real data from Chinese railway will be introduced. The conclusion will be given in Section 6.

2. Analysis of the Relevance between Final Booking Numbers and Booking Curve Shapes

The booking curves, with the time length before departure as the horizontal axis and the accumulated booking numbers by that time as the vertical axis, show the dynamic booking processes. The inclination of a booking curve at a certain point could reflect how concentrated it is for the passengers to book the tickets around that time point. From common sense, we know that the more likely the demand will go high, the more likely the people will book the tickets earlier. That is to say, for a given amount of tickets in stock to be sold, the higher the demand is, probably the more sharply the accumulated booking number increases during the earlier stage. This prompts the assumption that booking curve shapes might somewhat be related to the amounts of the passenger demands.

To verify the relevance between traveling demands and booking curve shapes, several metrics depicting the curve shapes are defined at first. Divide a booking curve into θ

TABLE 1: Correlation analysis result for the selected data sets.

Data set number	1	2	3	4
Pearson's correlation with				
K_1	-0.217**	-0.121**	0.194**	0.645**
K_2	0.472**	0.579**	0.678**	0.595**
K_3	0.773**	0.780**	0.766**	0.821**
Δ_1	-0.409**	-0.414**	-0.208**	0.165**
Δ_2	-0.403**	-0.452**	-0.402**	-0.422**

** Correlation is significant at the 0.01 level (two-tailed).

FIGURE 1: Demonstration of the relevance between final booking amounts and booking curve shapes.

segments with equal time length and assume the averages of the first order differences of each segment to be K_i where $i = 1, 2, \dots, \theta$. Let $\Delta_j = K_{j+1} - K_j$, $j = 1, 2, \dots, \theta - 1$. Then the values of K_i as well as Δ_j can be taken as some metrics of booking curve shapes to a certain extent.

Four sets of booking data from the historical records of the ticketing and reservation system of the Chinese railway with different train numbers and different ODs were selected for the purpose of testing, all covering the period from January 1, 2009, to July 31, 2011. Let $\theta = 3$ and $K_1, K_2, K_3, \Delta_1, \Delta_2$ for each departure date in all data sets were calculated. Then, correlation analysis between them and the final passenger amounts on the corresponding departure date was conducted. The results are shown in Table 1.

According to the results, final passenger amounts show significant relevance with all the metrics at 0.01 level. This coheres with the above assumption.

The three booking curves from real data sets in Figure 1 demonstrate such relevance between final booking amounts and booking curve shapes.

3. Clustering Algorithm for Booking Curves

Since the intersections of booking curves and Y-coordinates correspond to final booking amounts, the extending patterns of booking curves actually imply the relationship between final booking amounts and booking curve shapes. Here clustering is utilized to find such patterns.

Though the regular clustering algorithms have been used in various fields, an algorithm designed to fit the specific

characteristics of the problem is generally better than regular ones. The clustering algorithm designed for booking curves is as below.

Let $B = \{B_i \mid i = 1, 2, \dots, N\}$ be a set of historical booking records, with i representing different departure dates. The less the value of i is, the earlier the departure date would be. Consider that $B_i = (b_{i,1}, b_{i,2}, \dots, b_{i,u})$ is the sequence of booking records for a certain departure date i , while $b_{i,j}$ represents the accumulated numbers having been booked by the $(j - 1)$ th day before departure. $b_{i,1}$ is the final booking number by the departure date or, in other words, the passenger amount. u is the total number of records being utilized, which could be either the length of the whole booking horizon or just the number of selected days showing significance in terms of booking amounts in the booking horizon.

The clustering algorithm employs merging approach, which means that, at the very beginning, each individual constitutes a separate class, as follows:

$$\begin{aligned} CS^{(0)} &= \{C_1^{(0)}, C_2^{(0)}, \dots, C_N^{(0)}\}, \\ C_i^{(0)} &= \{B_i\}, \quad i = 1, 2, \dots, N. \end{aligned} \quad (1)$$

For each step after that, two existing classes are merged into one class until a certain criterion is satisfied. Such a merging approach also makes the update of the classification easier when new data are put into the training set as time goes on.

The merging process is as follows.

Let $CS^{(k)} = \{C_n^{(k)} \mid n = 1, 2, \dots, N^{(k)}\}$ be the set of the classes after having been merged for k steps. $N^{(k)}$ is the amount of classes. Since, for each step, there are only 2 classes being involved in the merging process, $N^{(k)} = N - k$,

$$C_n^{(k)} = \{B_i \mid i = q_{n,1}^{(k)}, q_{n,2}^{(k)}, \dots, q_{n,N_n^{(k)}}^{(k)}; q_{n,l}^{(k)} \in \{1, 2, \dots, N\}\}, \quad (2)$$

where $N_n^{(k)}$ is the amount of individuals in $C_n^{(k)}$, the n th class after k steps of merging.

Step 1. Calculate the center of each class. The arithmetical means of all the curves in the class $C_n^{(k)}$ can be seen as its center, denoted by $\bar{B}_n^{(k)} = (\bar{b}_{n,1}^{(k)}, \bar{b}_{n,2}^{(k)}, \dots, \bar{b}_{n,u}^{(k)})$. Consider

$$\bar{b}_{n,j}^{(k)} = \frac{1}{N_n^{(k)}} \sum_{l=1}^{N_n^{(k)}} b_{q_{n,l}^{(k)},j}, \quad j = 1, 2, \dots, u. \quad (3)$$

Step 2. Calculate the distances between the centers of every two classes. The most convenient way in doing so is to employ the Euclidean distance measurement. However, since the booking records closer to the departure date should provide more reliable information about the final booking amounts, as discussed in [7], a weighted Euclidean distance measurement is used here, in which the weights are decided based on the difference between the booking date and the

departure date. The closer the dates are, the higher the weights are assigned. The measurement is defined as

$$d(X, Y) = \frac{1}{u} \sqrt{\sum_{j=1}^u \frac{1}{j^2} (x_j - y_j)^2}. \quad (4)$$

In (4), $X = (x_1, x_2, \dots, x_u)$, $Y = (y_1, y_2, \dots, y_u)$.

Step 3. Decide the two classes to be merged. Construct a set named S as

$$\begin{aligned} S &= \{(C_m^{(k)}, C_n^{(k)}) \mid C_m^{(k)}, C_n^{(k)} \in CS^{(k)}, \forall B_i \in C_m^{(k)}, \\ &\quad B_j \in C_n^{(k)}, d(B_i, B_j) \leq D'_{\max}\}, \end{aligned} \quad (5)$$

where $D'_{\max}$ is a given threshold for the maximum distance between any two individuals in the future classes after merging.

Then, find $(C_{m'}^{(k)}, C_{n'}^{(k)}) \in S$, which fulfills

$$\forall (C_m^{(k)}, C_n^{(k)}) \in S, \quad d(\bar{B}_{m'}^{(k)}, \bar{B}_{n'}^{(k)}) \leq d(\bar{B}_m^{(k)}, \bar{B}_n^{(k)}). \quad (6)$$

Step 4. Judge the condition for ending clustering. If $S = \emptyset$ or $d(\bar{B}_{m'}^{(k)}, \bar{B}_{n'}^{(k)}) > D''_{\max}$, where $D''_{\max}$ is another given threshold, stop merging process and go to Step 5; otherwise, merge the two classes $C_{m'}^{(k)}$ and $C_{n'}^{(k)}$ and return to Step 1. The set of the classes after merging is as follows:

$$\begin{aligned} CS^{(k+1)} &= \{C_n^{(k)} \mid n = 1, 2, \dots, N^{(k)}; n \neq m'; n \neq n'\} \\ &\cup \{C_{m'}^{(k)} \cup C_{n'}^{(k)}\}. \end{aligned} \quad (7)$$

Step 5. To utilize the clustering results in predicting the final booking amounts, further, the classes are sorted according to the first elements of the class centers, namely, $\bar{b}_{n,1}^{(k)}$.

Assuming that $CS = \{C_1, C_2, \dots, C_\gamma\}$ is the set of the classes after sorting, while γ is the total number of the classes, and the center of C_n is $\bar{B}_n = (\bar{b}_{n,1}, \bar{b}_{n,2}, \dots, \bar{b}_{n,u})$, then $\bar{b}_{1,1} < \bar{b}_{2,1} < \dots < \bar{b}_{\gamma,1}$.

Let $C(B_i) \in \{1, 2, \dots, \gamma\}$ represent the class number in CS that B_i belongs to. Then, based on the historical booking records $B = \{B_1, B_2, \dots, B_N\}$, a series of class numbers for each day, $\{C(B_1), C(B_2), \dots, C(B_N)\}$, can be constructed. Comparing it with the series of the final booking amounts $\{b_{1,1}, b_{2,1}, \dots, b_{N,1}\}$, it can be found that the two series are similar to each other in terms of variation trend. Figure 2 illustrates such comparison for a certain train number and a certain OD during 2010.

Moreover, since the passenger flows generally subject to yearly periodicity, the series of class numbers of booking curves also show such a characteristic. Figure 3 illustrates the similarity between that of 2009 and 2010 for a given train number and given OD. It can be seen that the two curves are close to each other except the section between late January and March where the change of 2010 seems to lag behind that

FIGURE 2: Comparison between the series of passenger amounts and the corresponding class numbers.

FIGURE 3: Comparison of the series of class numbers for consecutive years.

of 2009. This can be explained by the fact that the Chinese Spring Festival, which has tremendous impact on passenger flow, is defined by Chinese lunar calendar. In 2009, the festival was in January 26, while, in 2010, it was in February 14.

4. Prediction Model

The prediction model is based on three sets of data:

- (1) the set of all class centers, namely, $\{\bar{B}_n \mid n = 1, 2, \dots, \gamma\}$;
- (2) the class numbers that the historical booking curves belong to, denoted by $\{C(B_i) \mid i = 1, 2, \dots, N\}$;
- (3) the bookings on hand corresponding to the booking amounts to be predicted. Assuming that there are s days left before the departure date p , for which the final booking number is to be predicted, the bookings on hand will be $B_p = \{b_{p,s}, b_{p,s+1}, \dots, b_{p,u}\}$.

The main idea is as follows. Assuming that the booking curve to be predicted matches one of the patterns drawn from the historical data and represented by the class centers $\{\bar{B}_n \mid n = 1, 2, \dots, \gamma\}$, the bookings on hand, B_p , are compared with the elements in $\{\bar{B}_n\}$ and the class center that matches B_p best can be used to forecast the trend following B_p in the future.

The weighted average quadratic sum $d'(B_p, \bar{B}_n)$, as shown in (8), is used as the metric to decide the most matched class:

$$d'(B_p, \bar{B}_n) = \frac{1}{u-s+1} \sqrt{\sum_{j=s}^u \frac{1}{j^2} (b_{p,j} - b_{n,j})^2}. \quad (8)$$

And the prediction of the class number for B_p , $C(B_p)$, is decided by

$$C(B_p) = \left\{ m \mid 1 \leq m \leq \gamma; d'(B_p, \bar{B}_m) = \min_n d'(B_p, \bar{B}_n) \right\}. \quad (9)$$

Then, the prediction of the final booking amount on the departure date p , denoted by $b'_{p,1}$, can be done in two ways. One is to directly use the first element of $\bar{B}_{C(B_p)}$ as the final result, as follows:

$$b'_{p,1} = \bar{b}_{C(B_p),1}. \quad (10)$$

The other is to use the increments in $\bar{B}_{C(B_p)}$ as the prediction of future increments, as follows:

$$b'_{p,1} = b_{p,s} + (\bar{b}_{C(B_p),1} - \bar{b}_{C(B_p),s}). \quad (11)$$

Since the booking curves being similar to each other at early stage could have quite different results and, at the same time, the classes have been sorted according to their final booking amounts, if the searching scope for the most matched class can be narrowed, the chance to match the wrong class with a similar curve at early stage should be reduced somewhat. For this reason, considering the similarity between class numbers for the same period in consecutive years, the interval between the minimum and maximum class numbers, denoted by $n_{\min}$ and $n_{\max}$, respectively, in the 4 weeks period with the same day in last year as center is set as the searching scope. Formula (9) is then replaced with (12) as follows:

$$C(B_p) = \left\{ m \mid n_{\min} \leq m \leq n_{\max}; \right. \quad (12)$$

$$\left. d'(B_p, \bar{B}_m) = \min_{n \in [n_{\min}, n_{\max}]} d'(B_p, \bar{B}_n) \right\}.$$

What needs to be explained is that “the same day in last year” might not be exactly the same calendar date. Taking the Chinese Spring Festival as an example, it conforms to Chinese lunar calendar. So, for the time period around the Spring Festival, “the same day in last year” should refer to the same date on Chinese lunar calendar.

5. Empirical Study

5.1. Data Used. To test the performance of the proposed model, the 4 data sets once used in the correlation analysis in Section 2 were utilized again for empirical study. The data with departure dates ranging from January 1, 2009, to June 30, 2010, were used for initial training, while those in

TABLE 2: Summary of the models being tested.

Name	Model	Search scope	Calculation of final result
CM-1	Clustering based	[1, γ]	Formula (10)
CM-2			Formula (11)
CM-3		[$n_{\min}$, $n_{\max}$]	Formula (10)
CM-4			Formula (11)
AP	Advanced pickup	—	—

the following two months were used for test. The selected period for test covers the summer vacation season. The actual traveling demands in that period fluctuate dramatically, which makes it harder to predict accurately.

5.2. Methodology. In the study, the two approaches in deciding the searching scope for the most matched class, as illustrated in formulas (9) and (12), were both tested. The two different ways in calculating the prediction of the final booking amounts, as shown in (10) and (11), respectively, were also tested. In addition, the advanced pickup model with exponential averages [7], which is thought to have outstanding performance for short-term prediction and is widely used in revenue management practice, was also used for the purpose of comparison. Table 2 provides the summary of the tested models.

Since parameter values can have significant impact on the results, different parameter values were tested in the study. For the proposed clustering algorithm, there are two parameters, $D'_{\max}$ and $D''_{\max}$. $D'_{\max}$, the maximum distance between any two individuals in the classes to be merged, which acts as the threshold to control which classes could be merged, was tested with all the integers in [1, 6]. $D''_{\max}$, the maximum distance between the centers of the classes to be merged, which controls when should the merging process stop, was tested with the numbers in [0.5, 5] with an interval of 0.5.

For each departure date during the test period, different forecasting horizons ranging from 1 day to 8 days were applied to test the applicability of each model to different forecasting horizons.

For a given set of parameter values and forecasting horizon length, the results were evaluated with the mean values of the relative errors as defined by

$$RE = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} \left(\frac{b_i^f - b_i^r}{b_i^r} \right)^2}, \quad (13)$$

where N_t represents the total number of different departure dates for test, b_i^f represents the forecasting result for the i th departure date, and b_i^r , as a comparison, represents the actual final booking amounts on that departure date.

5.3. Results and Analysis. First of all, the performances of the different approaches in deciding the searching scope and the final forecasting results as shown in (9) to (12) are evaluated through the comparison between CM-1 and CM-4 models.

Let $RE_{CM-1}(D'_{\max}, D''_{\max})$, $RE_{CM-2}(D'_{\max}, D''_{\max})$, $RE_{CM-3}(D'_{\max}, D''_{\max})$, and $RE_{CM-4}(D'_{\max}, D''_{\max})$ represent the RE values for the models CM-1, CM-2, CM-3, and CM-4, respectively, given a certain value combination of the parameters $D'_{\max}$ and $D''_{\max}$. Calculate the differences between them as follows:

$$\delta_{ij}(D'_{\max}, D''_{\max}) = RE_{CM-i}(D'_{\max}, D''_{\max}) - RE_{CM-j}(D'_{\max}, D''_{\max}). \quad (14)$$

The distribution of $\delta_{13}(D'_{\max}, D''_{\max})$ with different parameter value combinations of $(D'_{\max}, D''_{\max})$ is demonstrated in Figure 4(a), while that of $\delta_{24}(D'_{\max}, D''_{\max})$ is shown in Figure 4(b). In the figures, the vertical axes represent the amounts of different value combinations of $(D'_{\max}, D''_{\max})$ with which $\delta_{ij}(D'_{\max}, D''_{\max})$ fall into the corresponding sections defined by the horizontal axes.

From Figure 4, it can be seen that both $\delta_{13}(D'_{\max}, D''_{\max})$ and $\delta_{24}(D'_{\max}, D''_{\max})$ locate mostly on the right side of 0%, which means that the narrower search scope as defined in (12) does improve the accuracy of the forecast in most cases comparing to the unlimited searching scope defined in (9).

To compare the forecasting performances between models employing (10) and (11), $\delta_{12}(D'_{\max}, D''_{\max})$ and $\delta_{34}(D'_{\max}, D''_{\max})$ were calculated. Their distributions are demonstrated in Figures 5(a) and 5(b), respectively.

In Figure 5, though the differences are not as dramatic as those in Figure 4, still, the approach defined by (11), employing the same idea as pickup models, generally outperforms that defined by (10).

Based on the above analysis, the model CM-4, which employs the two better approaches defined by (12) and (11), will be used in comparison with the advanced pickup model.

The comparison of the relative errors between CM-4 and the advanced pickup model for all the 4 data sets is illustrated in Figure 6.

For the 1st and 2nd data sets, CM-4 model shows much better accuracy when forecasting is made more than 5 days ahead of departure date. As forecasting horizon gets shorter or, in other words, as forecasting was made closer to departure date, the relative errors of the advanced pickup model decreased faster than CM-4 model. As a result, the advantage of CM-4 model gets smaller during the middle stage. When forecasting was made just 3 or even less days before departure date, the advanced pickup model surpasses CM-4 model in terms of forecasting accuracy.

For the 3rd data set, both models produced relatively higher error when forecasting is made more than 1 day before departure date. When forecasting horizon lies between 5 and 7 days, CM-4 model shows a little better performance, while, for forecasting made during the following 3 days, the advanced pickup model performs better.

For the 4th data set, the mean relative errors of CM-4 model are totally higher than those of the advanced pickup model.

The above observation can be summarized as the following two facts. First, CM-4 is more suitable to be used for data

FIGURE 4: Distribution of the differences between the RE values of models with different ways in deciding the searching scopes.

FIGURE 5: Distribution of the differences between the RE values of models with different ways to decide the final booking amounts.

sets numbers 1 and 2 and is most unsuitable to be used for data set number 4. Secondly, comparing to the advanced pickup model, CM-4 model shows better forecasting accuracy for longer forecasting horizon.

To find the possible reason for different suitability of the proposed model to different data sets, the booking curves for each departure date in the 4 data sets are drawn in Figure 7. To make the diagram clearer, the curves are sampled with a ratio of 1:5.

As shown in Figure 7, there are more bookings that happened in the early stage of booking horizon for the 1st and 2nd data sets, which means that the corresponding parts of the booking curves have higher diversity and could convey more information about the whole booking process. On the contrary, for the 4th data set, most bookings concentrate in the last 3 days before departure. This leads most of the booking curves to be less distinguishable in the early and middle stages and also to convey less valuable information

FIGURE 6: Comparison between the relative errors of CM-4 and the advanced pickup model.

about all the processes. The distribution of the bookings curves in the 3rd data set seems to be just between the above two situations. Such characteristics can also be observed from the correlation coefficients between the final booking amounts and K_i listed in Table 1.

The different distributions of the booking curves could be used to explain the different performance of CM-4 model for different data sets. First, the clustering algorithm relies heavily on the information conveyed by the booking curves. The more the information they convey, the more reasonable the classifications are. In addition, if the booking curves are more distinguishable, the comparison between the bookings on hand and the class centers in the first step of the prediction process could be more accurate. That is why the proposed clustering based model is more suitable to the 1st and 2nd data sets.

Moreover, according to Table 1, for the first 3 data sets, as the time before departure gets shorter, the relevance between the final booking amounts and K_i decreases continuously. The final booking amounts hold the strongest positive relevance with K_3 which corresponds to the early stage of booking horizon. Then the relevance with K_2 , corresponding to the middle stage, decreases to different extent for different data sets. Further, the relevance with K_1 stands at the lowest level. So, as the time when forecast is made gets closer to departure date, there is more information on hand that generally makes the forecast more accurate. But the improvements on the forecasting accuracy for the proposed model are not as much as

those for the advanced booking model due to the sharply decreasing relevance between the booking amounts and the booking curve shapes, while the relevance underlies the proposed model. That is why CM-4 model shows better forecasting accuracy for longer forecasting horizon but is surpassed when forecasts are made at the final stage of the booking horizon.

6. Conclusion

In this paper, a short-term forecasting approach for railway passenger transportation based on clustering technique is proposed. The approach employs data from both historical booking records and bookings on hand. The empirical study shows that the proposed model outperforms the advanced pickup model, one of the most popular models in practice, during the early and middle stages of booking horizon when bookings are not concentrated in the final days before departure.

Further research works could involve several aspects. One is to further investigate and verify the suitable circumstances of the model with more real data. Secondly, how to choose the most proper one from various forecasting models under different circumstances according to statistical characteristics is also an interesting topic. Besides, it is worthwhile to study the possibility of adapting the proposed model to air transportation.

FIGURE 7: The booking curves in different data sets.

Conflict of Interests

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

Acknowledgments

This research was supported by the National Key Technologies R&D Program funded by the Ministry of Science and Technology of the People's Republic of China (2009BAG12A10) and the Science and Technology Research Program of the Ministry of Railway of China (2012F023).

References

- [1] T. H. Tsai, C. K. Lee, and C. H. Wei, "Design of dynamic neural networks to forecast short-term railway passenger demand," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 6, pp. 1651–1666, 2005.
- [2] T. H. Tsai, C. K. Lee, and C. H. Wei, "Neural network based temporal feature models for short-term railway passenger demand forecasting," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3728–3736, 2009.
- [3] N. Phumchusri and P. Mongkolku, "Hotel room demand forecasting via observed reservation information," in *Proceedings of the Asia Pacific Industrial Engineering & Management Conference*, pp. 1978–1985, Phuket, Thailand, December 2012.
- [4] A. Haensel and G. Koole, "Booking horizon forecasting with dynamic updating: a case study of hotel reservation data," *International Journal of Forecasting*, vol. 27, no. 3, pp. 942–960, 2011.
- [5] R. R. Wickham, *Evaluation of forecasting techniques for short-term demand of air transportation [M.S. thesis]*, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1995.
- [6] M. Irwin, *Short-term revenue forecasting at KLM [M.S. thesis]*, Leiden University, Leiden, The Netherlands, 2010.
- [7] T. H. Tsai, "A temporal case retrieval model to predict railway passenger arrivals," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8876–8882, 2009.
- [8] H. Huang, Q. Tang, and Z. Liu, "Adaptive correction forecasting approach for urban traffic flow based on fuzzy C-mean clustering and advanced neural network," *Mathematical Problems in Engineering*, vol. 2013, Article ID 195824, 7 pages, 2013.
- [9] C. Hueglin and F. Vannotti, "Data mining techniques to improve forecast accuracy in airline business," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pp. 438–442, San Francisco, Calif, USA, August 2001.
- [10] M. Kumar and N. R. Patel, "Using clustering to improve sales forecasts in retail merchandising," *Annals of Operations Research*, vol. 174, no. 1, pp. 33–46, 2010.
- [11] M. Misiti, Y. Misiti, G. Oppenheim, and J. M. Poggi, "Optimized clusters for disaggregated electricity load forecasting," *REVSTAT—Statistical Journal*, vol. 8, no. 2, pp. 105–124, 2010.
- [12] M. Dong, "A tutorial on nonlinear time-series data mining in engineering asset health and reliability prediction: concepts, models, and algorithms," *Mathematical Problems in Engineering*, vol. 2010, Article ID 175936, 22 pages, 2010.
- [13] V. Kedia, V. Thummala, and K. Karlapalem, "Time series forecasting through clustering—a case study," in *Proceeding of the Advances in Data Management*, pp. 183–191, Goa, India, January 2005.
- [14] D. R. Morales and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining," *European Journal of Operational Research*, vol. 202, no. 2, pp. 554–562, 2010.