ISSN 1648-4142 / eISSN 1648-3480

VILNIUS TECH
Vilnius Gediminas Technical University

TRANSPORT

2025
Volume 40
Issue 1
Pages 12–23
https://doi.org/10.3846/transport.2025.20517

Original Article

FORECASTING SHORT-TERM PASSENGER FLOW ON A BUS ROUTE: A SPLITTING–INTEGRATING METHOD BASED ON PASSENGER TRAVEL BEHAVIOR

Xiaoping FANG[1], Mei LIN[2]✉, Weiya CHEN[3], Xin PAN[4]

[1,3,4]*School of Traffic and Transportation Engineering, Rail Data Research and Application Key Laboratory of Hunan Province, Central South University, Changsha, China*
[2]*Institute of Transportation Development Strategy and Planning of Sichuan Province, China*

Highlights:

- this article proposed to use SQL to mine the correlation between the bus card number and the travel time to infer the characteristics of passenger travel behaviour;
- this article established a splitting–integrating prediction method that takes into account both passenger travel behaviour characteristics and time characteristics;
- this article confirmed that the subset of passenger flow data divided according to the characteristics of passenger travel behaviour is beneficial to improve the prediction performance, but the size and characteristics of the subset will affect the prediction performance;
- this article compared and confirmed the adaptability of SARIMA method and RBF method.

Article History:

- submitted 17 February 2019;
- resubmitted 19 December 2019, 27 February 2020;
- accepted 15 June 2020.

Abstract. Short-term passenger flow forecasting is the key to implement real-time dynamic dispatching of buses, which can meet the travel time requirement of passengers with different attributes. In practice, it is difficult to obtain passenger attribute information due to the restriction of bus information systems or other conditions. This article proposes a new perspective on identifying passenger attribute information, that is, the correlation between the bus card number and the travel time is used to analyse passenger travel behaviour. Then using the travel frequency as the splitting boundary, the passenger set is split into different types of subsets, which are predicted by different methods. The total forecast values are obtained by integration, so as to explore the effectiveness of the passenger attribute identification and splitting–integrating method. The result shows that: (1) compared with the forecasting method without considering the passenger travel behaviour, the performance of splitting–integrating method is better, and the passenger attribute identification method is effective; (2) the value of the splitting boundary will affect the size and consistency of the subset, and the optimal value can be sought according to forecast results; (3) different types of subsets should be treated by different forecasting models and combination paths.

Keywords: short-term passenger flow forecasting, bus passenger flow, passenger attribute, passenger travel behaviour, splitting–integrating method.

✉Corresponding author. E-mail: *lm034csu@163.com*

Notations

Variables and functions:

goal – the mean square error target value;
mn – the maximum number of neurons;
R^2 – coefficient of determination;
spreed – expansion speed of radial basis function;
TS – the boundary for extracting passengers.

Abbreviations:

AC – autocorrelation coefficient;
ADF – unit root value;
AIC – Akaike information criterion;
ARIMA – autoregressive integrated moving average;
BP – back propagation;
EMD – empirical mode decomposition;
KNN – *k*-nearest neighbour;
MAPE – mean absolute percentage error;
PAC – partial AC;
PP – Phillips–Perron;
RBF – radical basis function;
RF – random forest;
RMSE – root mean square error;
SARIMA – seasonal ARIMA;

SC – Schwarz criterion;
SQL – structured query language;
SSA – singular spectrum analysis;
SVM – support vector machines.

1. Introduction

Intelligent bus is developing rapidly in many cities and has become a future development trend. In addition, the short-term passenger flow forecasting is the key to the development of intelligent bus. However, there is a contradiction between the diversity of passenger travel demand and the stabilization of bus service time, and different passengers have different requirements for travel time. Therefore, it is not enough to only consider the travel time of passenger flow to make short-term forecasting. If short-term forecasting can be made based on the passenger travel behaviour and attributes, it can not only better respond to passengers demand for travel time, but also effectively alleviate the contradiction.

Short-term passenger flow forecasting is a non-linear and time-varying problem, which refers to forecast the passenger flow no more than 15 min ahead, and needs to be based on the feature of passenger flow (Qiu, Yang 2013; Vlahogianni, Karlaftis 2011; Bai 2017). Influenced by many complicated factors, both natural and human, passenger travel behaviour has markedly time-varying feature. Therefore, short-term forecasting is more difficult than the medium-term or long-term forecasting (Wang *et al.* 2015a; Teng, Chen 2015). Its essence lies in the randomness of passenger travel and the diversity of passenger attributes. In addition, it is theoretically considered that the feature of the data is descriptive indicator of object attribute. If the object attribute corresponding to the data, that is, the passenger attributes, can be effectively analysed, then the generation and distribution of the data can be essentially recognized, which is conducive to identifying the feature of the data.

In reality, it is very hard to obtain passenger attribute information due to the limitation of bus information systems or other application conditions (Haworth, Cheng 2012). For the bus card data collected by the bus systems, it generally only includes the time and date data, and occasionally includes the card number or the location data. However, most of the bus cards are not bound with the passenger's name, so it is difficult to collect the passenger's personal information. Therefore, most of the collected passenger flow data is incomplete. In this context, how to identify passenger attribute information based on the collected data, so as to identify the feature of short-term passenger flow data more effectively, is a key problem.

The rest of this article is organized as follows: current Section 1 is introduction to the problem; in Section 2, it is the literature review; in Section 3, the splitting–integrating method is presented in detail; in Section 4, application of this method and results are elaborated; finally, conclusions are present at the Section 5.

2. Literature review

The attribute of passenger flow data is divided into 3 basic attributes: (1) classified, (2) ordinal, and (3) numerical attribute (Zuo 2016). Classified attribute is some names, symbols, such as bus card number, card type, or location. Ordinal attribute is ordered, like travel time, date, etc. Numerical attribute data is expressed by actual values, such as the number of passengers, the times or frequency of travel. If there are errors or omissions in the collection and transmission of some attribute values, the collected data is incomplete data that is occasionally or skipped missing. If the relevant fields are not considered in the design of information system, and all values of this attribute cannot be collected, then the collected data is incomplete data with limited attribute.

The way to identify passenger attribute information from incomplete data depends on the incompleteness of the data. For occasional or skipped incomplete data, the passenger attribute information can be effectively identified after using the existing data to deduce and fill in. However, for incomplete data with limited attribute, especially when the information related to passenger attribute is missing, the way to identify passenger attribute information is either to upgrade the system or to analyse the correlation between existing attribute information and passenger attribute information (Zhong *et al.* 2006).

At present, scholars usually perform appropriate correlation analysis based on the feature of the existing data, analyse passenger travel behaviour, and identify passenger attribute information. For example, Du & Aultman-Hall (2007) used a heuristic method to analyse the correlation between the location data, and inferred the boarding and alighting behaviour of each passenger. However, when the existing data is only indirectly related to passenger attribute, such as bus card number, some scholars have tried to analyse passenger travel behaviour, and identified passenger attribute information by analysing the frequency of card number appearing within a certain period. For example, Lu (2016) analysed the passenger travel behaviour based on the frequency of card number appearing, which appeared 20 times in a month, and regular passengers and random passengers were inferred. However, the statistical standard of the frequency of card number appearing will affect the analysis of passenger travel behaviour, thus affect the quality of passenger attribute identification and forecast results.

It is worth noting that the time-varying feature of short-term passenger flow is significant, and the travel behaviour of different passengers vary greatly (Wang *et al.* 2015a). These make the components of the forecast object complicated, and sometimes using a single model to forecast the total passenger flow dataset may not achieve the satisfied results. If the passenger flow with different features is split and each subset is predicted by different models, it's possible to improve the prediction effect. Therefore, the rationality of passenger attribute identification and splitting method can be judged by the forecast result.

Clustering analysis that maximizes homogeneity within groups and heterogeneity among groups based on predefined metrics is a typical classification method (Sfetsos, Siriopoulos 2004). Chen *et al.* (2019) used *k*-means clustering algorithm by setting multiple clustering values to split weekday passenger flow into the types of passenger flow of peak time, secondary peak time, flat peak time, secondary low peak time and low peak time, and found the best category number by comparing the test coefficients. In the same way, consider setting the statistical standard multiple times, the total passenger flow set is split into subsets of different sizes, and the best statistical standard and the best subset are found by forecast results.

The motivation of this article is to explore whether the forecasting method considering the passenger attribute information can improve the forecast performance. Compared with the forecasting method that does not consider passenger travel behaviour and attribute information, the rationality of passenger attribute information identification and the effectiveness of the splitting–integrating method are verified.

3. Methodology

Methodology is the basis for passenger flow forecasting. 1st-of-all, the principle of the splitting–integrating forecasting method is described in detail. Then the experimental data and its feature are analysed. In addition, an adaptive forecasting model is selected to pave the way for the application of the splitting–integrating method.

3.1. Splitting–integrating forecasting method

The splitting–integrating forecasting method based on incomplete data has 3 steps, as shown in Figure 1.

Step 1. Data decomposition. For incomplete passenger flow, splitting it into different subsets {1, 2, ..., *m*} according to the correlation between data attributes.

Step 2. Establish forecasting model. The applicable models {N_1, N_2, ..., N_n} are selected to forecast each subset, and forecasting models $N_{xy}\left(x=\{1, 2, ..., n\}; y=\{1, 2, ..., m\}\right)$ are established.

Step 3. Integrate the forecast values of each subset, and obtain the optimal combination model. By adopting the fully combined path of forecasting model, the forecast values of subsets are integrated. If the forecast performance is not optimal, the combination path is reselected until the optimal forecast performance is obtained.

3.2. Data and feature analysis

This article takes the bus card consumption records of the No 104 bus in Changsha (China), as an example. Since the buses start and end at Changsha railway station and Jiufeng park station, which pass through densely populated areas such as schools, hospitals and subways. A large number of passengers swipe their card every day, thus the collected data has different features. This can be used to explore the feasibility of splitting–integrating forecasting method.

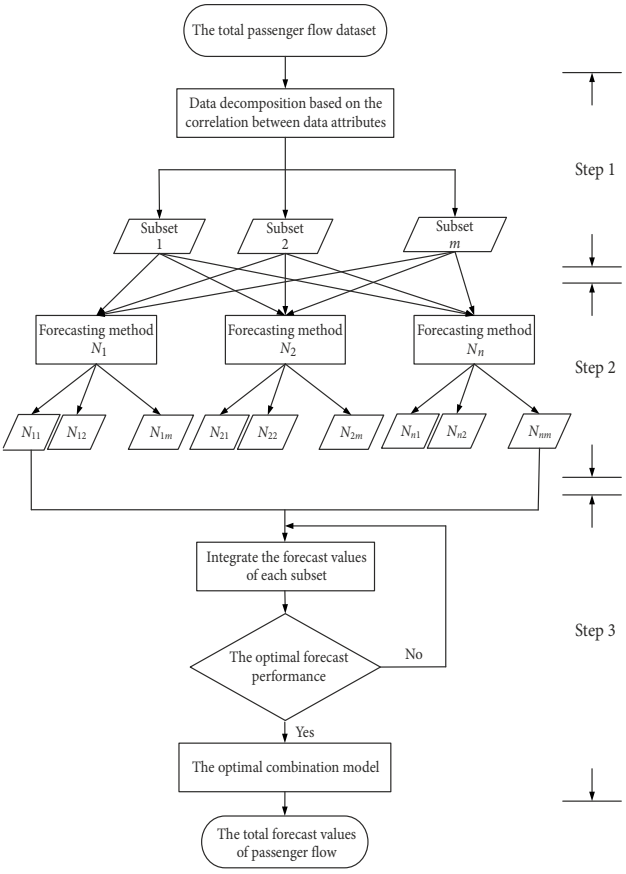


Figure 1. The principle of the splitting–integrating forecasting method

The passenger flow information was collected through the bus system, which was collected by 20 buses in real time 3–30 June 2017, constituting 216749 records in all. Each record includes the number of passengers, bus car number, bus card number, boarding time and boarding date. In addition, each attribute has one-to-one correlation. Among them, the boarding date is any day (3–30 June 2017), and the boarding time is any time from 6:00 to 22:15 daily. Table 1 is the total daily passenger flow volume on working day and on non-working day. It shows that the weekly passenger flow volume is basically maintained at about 54000, of which the passenger volume is about 14000 on non-working days and about 40000 on working days. It can be seen that the overall weekly passenger flow volume is relatively stable.

The time-varying feature of the passenger flow data is the important basis for the selection of forecasting model. Taking 15 min as the statistical unit, the daily operation period (6:00...20:15) is divided into 66 periods. And the total passenger flow volume is allocated to each period to form original time series, which is shown in Figure 2. It can be seen that the time series shows a relatively stable and periodic time-varying features. As can be seen from Figure 3, the time-varying features of daily passenger flow are non-stationary and random.

Table 1. Descriptive statistics of the total passenger flow dataset

Week	On non-working days			On working days			Total	
	Date	Passenger flow [number]	Percentage of total passenger flow [%]	Date	Passenger flow [number]	Percentage of total passenger flow [%]	Passenger flow [number]	Percentage [%]
1st	3–4 June	14992	6.92	5–9 June	41365	19.08	56357	26.00
2nd	10–11 June	14070	6.49	12–16 June	40294	18.59	54364	25.08
3rd	17–18 June	14460	6.67	19–23 June	39554	18.25	54014	24.92
4th	24–25 June	12801	5.91	26–30 June	39213	18.09	52014	24.00
Total	–	56323	25.99	–	160426	74.01	216749	100.00

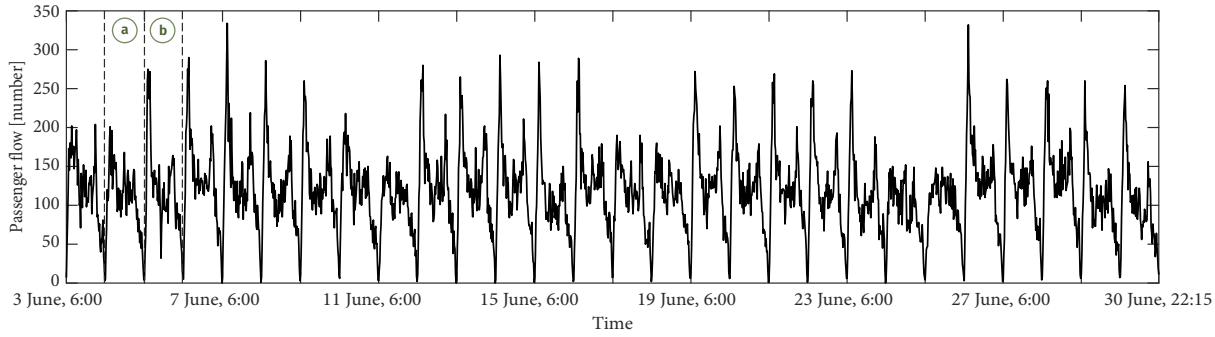


Figure 2. The original time series (3–30 June 2017): (a) – represents 4 June 2017; (b) – stands for 5 June 2017

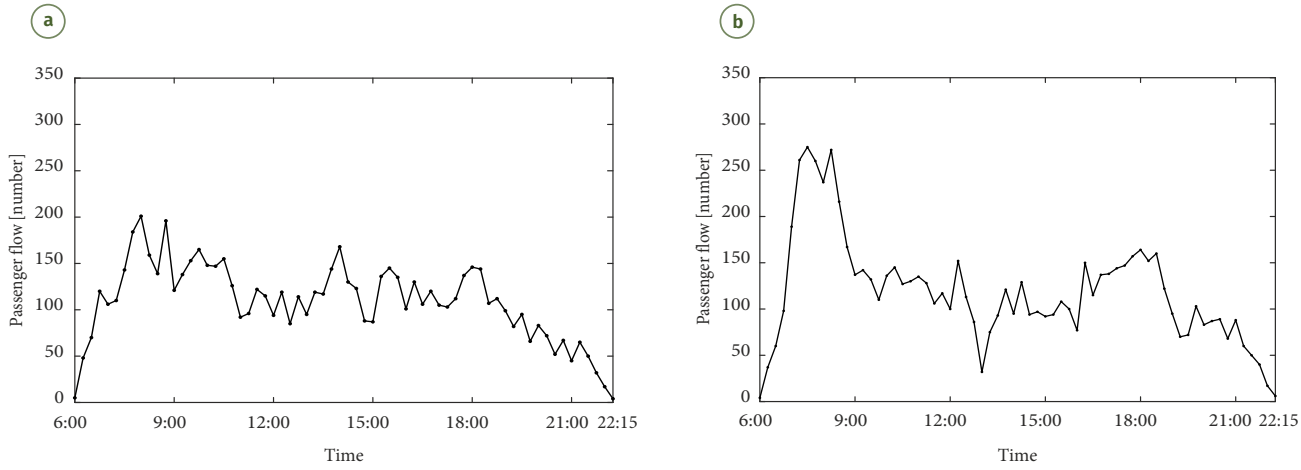


Figure 3. The distribution of the original time series: (a) – one non-working day; (b) – one working day

3.3. Forecasting models

A forecasting model refers to use mathematical language or formula to describe the quantitative relationship of historical data. For the short-term passenger flow forecasting, scholars have proposed many models. As reviewed in Qiu & Yang (2013), Wang *et al.* (2015) and Bai (2016), time series modelling is roughly based on 2 types of techniques.

The 1st type is based on mathematical and physical statistics, including moving average (Meng *et al.* 2018), Kalman filtering (Zhang *et al.* 2011), ARIMA (Zuo 2016), SARIMA (Wang *et al.* 2015a) and so on. Among them, Kalman filter is applicable to linear and stable data (Wang *et al.* 2016). ARIMA and SARIMA are suitable for non-linear stationary data (Zhao 2018). What's more, SARIMA is more applicable when the data has periodic feature (Williams, Hoel 2003).

The 2nd type is based on biological and computer simulation techniques, mainly including RF (Li *et al.* 2017), KNN (Dou 2011), SVM (Sun *et al.* 2015), BP neural network (Zhu 2017), RBF neural network (Feng *et al.* 2015) and so on. These models don't take mathematical derivation as the core, but pay more attention to the fitting effect (Qiu *et al.* 2013). Among them, both RF and KNN can establish models based on the feature of data, but need to set category parameters by experience. SVM, BP neural network and RBF neural network can be applied to data with any features (Tsai *et al.* 2009; Smith *et al.* 2002; Wang *et al.*

2015a). However, the algorithm of SVM is complex and time-consuming, especially in short-term passenger flow forecasting. BP neural network has the disadvantage that learning convergence takes a long time and can't guarantee to reach the global optimum. RBF neural network has the ability of global approximation and can deal with complex non-linear problems without prior knowledge (Feng *et al.* 2015).

Research result shows that there is no one forecasting model, which can predict all types of data well. It is necessary to choose an appropriate forecasting model according to the data environment (Smith *et al.* 2002; Gao, Er 2005). The above analysis shows that the experimental data has periodic, non-linear, non-stationary, and random feature. In addition, according to the characteristics and application environment of the above forecasting models, the SARIMA model based on the 1st type of technique has good applicability to periodic and non-linear data. The RBF neural network model based on the 2nd type of technique can process data containing complex feature. In addition, both models can adopt dynamic and rolling mode to better forecast short-term and small amounts of the data. In addition, in order to fully verify the effectiveness of the splitting–integrating method, the SARIMA model and RBF neural network model are selected as the representatives from the 2 types of techniques. The feasibility of the selected model is judged by forecasting the experimental data.

3.3.1. SARIMA model

Sample data within the 1st to 3rd week is used to establish the SARIMA model, and sample data within the 4th week for out-of-sample testing. The SARIMA model requires the data to be stable. If not, the data needs to be smoothed (Wang *et al.* 2015b; Ma *et al.* 2016).

The stability of data is judged by the ADF, sequence diagram, and autocorrelation diagram. As shown in Table 2, for the ADF test, *Prob* > 0.05. The time series shown in Figure 2 shows a non-linear variation. The AC shown in Figure 4 has no truncation, and s = 66. These indicate that the series is not stable, seasonal and non-seasonal differential processing are required.

The differenced series is obtained after 1st-order seasonal and 1st-order non-seasonal differential processing. Then its stability is tested. As shown in Table 3, the ADF test rejects the null hypothesis at a significance level of 1%. As shown in Figure 5, the series fluctuates around zero without an increasing or decreasing trend. The AC shown in Figure 6 presents a 1st-order truncation. Therefore, the series is stable, and meets the requirements of the SARIMA model.

The parameters of the model that need to be identified are the non-seasonal difference d, seasonal difference D, lag orders p, q, P and Q. The optimal values of parameters are judged using R^2, AIC, and SC. d = 1 and D = 1 are obtained through the above differential processing. It can be preliminarily judged from Figure 6 that the values of p and q are 1, 2, 3 and 1, 2, respectively, and the values of P and Q are 1, 2 and 1, 2, respectively. By combining p, q, P and Q, when p = 3, q = 2, P = 2 and Q = 1, the *SARIMA*(3, 1, 2)(2, 1, 1)$_{66}$ model is optimal. The test results are shown in Table 4, and the model is:

Table 2. The original time series of ADF test

Method	Statistic	Prob
ADF–Fisher χ^2	30.3238	0.4492
PP–Fisher χ^2	32.3437	0.3517

Table 3. The differenced series of ADF test

		t-statistic	Prob
Extended Dickey–Fuller test statistics		–13.97921	0.0000
ADF test	1%	–3.437314	–
	5%	–2.864503	–
	10%	–2.568401	–

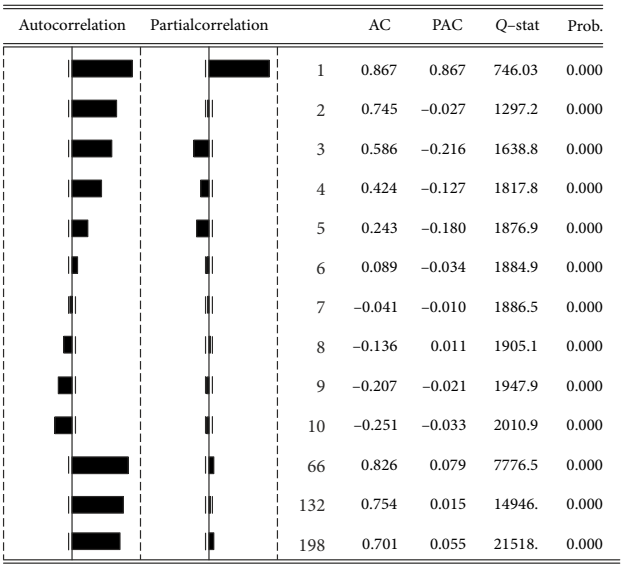


Figure 4. The original time series of autocorrelation diagram

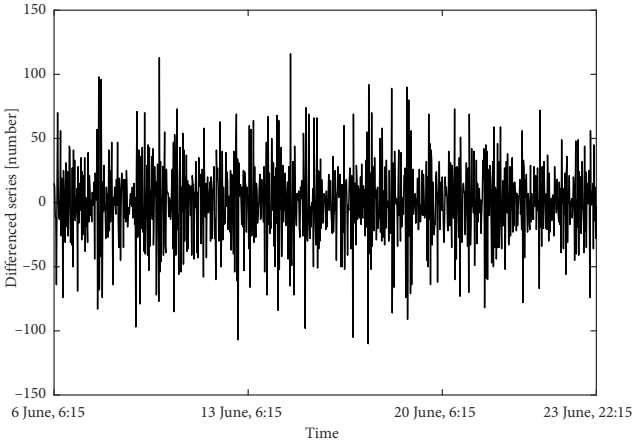
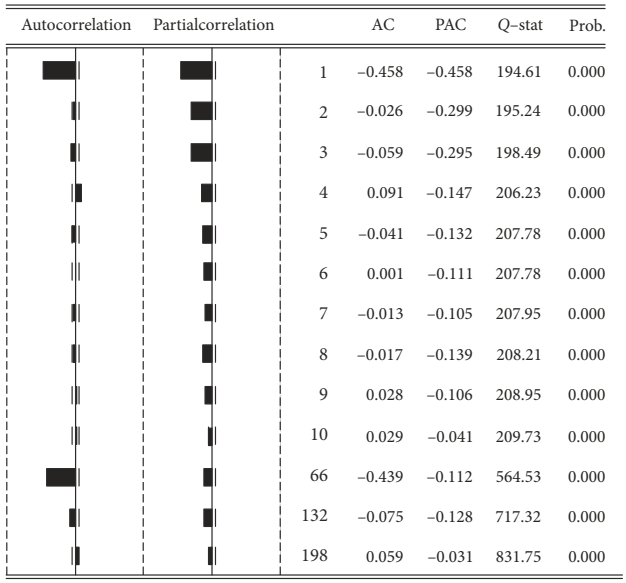


Figure 5. The differenced series of sequence diagram

$$\left(1+0.615\cdot B+0.049\cdot B^2+0.131\cdot B^3\right)\times$$
$$\left(1+0.125\cdot B^6\cdot 6+0.151\cdot B^1\cdot 32\right)\times$$
$$\left(1-B\right)\cdot\left(1-B^6\cdot 6\right)\cdot y_t=\left(1+0.241\cdot B+0.55\cdot B^2\right)\times$$
$$\left(1+0.906\cdot B^6\cdot 6\right)\cdot\varepsilon_t+0.0007, \quad (1)$$

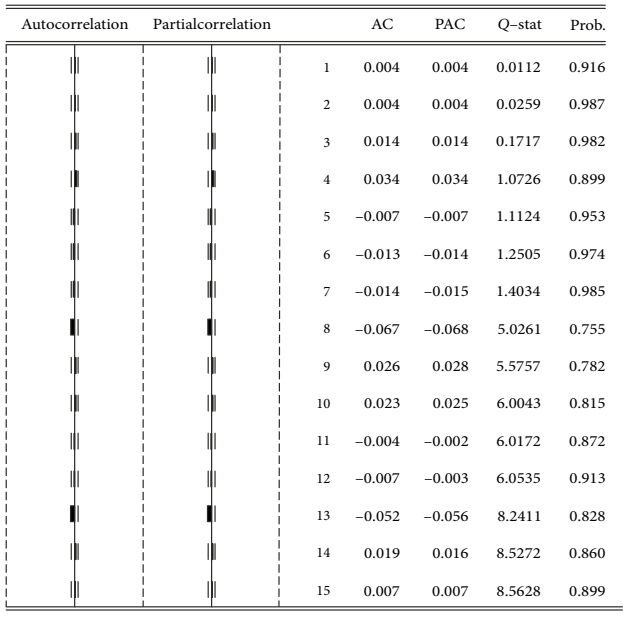
where: y_t is the forecast value; B is the backward shift operator $\left(B^n\cdot y_t=y_{(t-n)}\right)$; ε_t is the residual sequence, which should be independent.

The correlation of the residuals is tested using the Q statistic and the results are shown in Figure 7. Both the AC and PAC are within the acceptable range, indicating that the *SARIMA*(3, 1, 2)(2, 1, 1)$_{66}$ model is ideal.



Autocorrelation	Partialcorrelation		AC	PAC	Q-stat	Prob.
		1	−0.458	−0.458	194.61	0.000
		2	−0.026	−0.299	195.24	0.000
		3	−0.059	−0.295	198.49	0.000
		4	0.091	−0.147	206.23	0.000
		5	−0.041	−0.132	207.78	0.000
		6	0.001	−0.111	207.78	0.000
		7	−0.013	−0.105	207.95	0.000
		8	−0.017	−0.139	208.21	0.000
		9	0.028	−0.106	208.95	0.000
		10	0.029	−0.041	209.73	0.000
		66	−0.439	−0.112	564.53	0.000
		132	−0.075	−0.128	717.32	0.000
		198	0.059	−0.031	831.75	0.000

Figure 6. The differenced series of autocorrelation diagram



Autocorrelation	Partialcorrelation		AC	PAC	Q-stat	Prob.
		1	0.004	0.004	0.0112	0.916
		2	0.004	0.004	0.0259	0.987
		3	0.014	0.014	0.1717	0.982
		4	0.034	0.034	1.0726	0.899
		5	−0.007	−0.007	1.1124	0.953
		6	−0.013	−0.014	1.2505	0.974
		7	−0.014	−0.015	1.4034	0.985
		8	−0.067	−0.068	5.0261	0.755
		9	0.026	0.028	5.5757	0.782
		10	0.023	0.025	6.0043	0.815
		11	−0.004	−0.002	6.0172	0.872
		12	−0.007	−0.003	6.0535	0.913
		13	−0.052	−0.056	8.2411	0.828
		14	0.019	0.016	8.5272	0.860
		15	0.007	0.007	8.5628	0.899

Figure 7. Q-test of residual sequence

Table 4. The test results of model

Variable	Coefficient	Std. error	t-statistic	Prob.
C	0.000696	0.015998	0.043490	0.9653
AR(1)	–0.614902	0.178889	–3.437340	0.0006
AR(2)	–0.049119	0.052769	–0.930821	0.3522
AR(3)	–0.130888	0.041154	–3.180465	0.0015
SAR(66)	–0.124864	0.035549	–3.512417	0.0005
SAR(132)	–0.151270	0.034388	–4.398962	0.0000
MA(1)	–0.241215	0.178741	–1.349521	0.1776
MA(2)	–0.549884	0.163886	–3.355278	0.0008
SMA(66)	–0.906308	0.010542	–85.97457	0.0000
R^2	0.727938	AIC	8.490312	
Adjusted R^2	0.725144	SC	8.543643	

3.3.2. RBF neural network model

The steps of establishing the RBF neural network model are divided into 3 steps.

Step 1. Determining training and testing data.

Training data $X_n=\left\{x_n\mid x_n\in X_{kt},\ k\right.$ is the date within the 1st to 3rd week, t is the period, $\left.n\in\left\{1,2,\ldots,990\right\}\right\}$.

Testing data $Y_n=\left\{y_n\mid y_n\in Y_{kt},\ k\right.$ is the date within the 4th week, t is the period, $\left.n\in\left\{1,2,\ldots,330\right\}\right\}$.

Step 2. Training the network and determining the parameter.

The rolling training is that the output values are fed back into the network structure as part of the input values (Feng *et al.* 2015; Xie *et al.* 2013). When the RMSE between the output values and the observed values reaches a minimum, then the optimal values of the *goal*, *spread*, *mn* can be obtained (Wang, Cheng 2016). Rolling training is carried out: the input data $\left(X_h=\left\{x_{1+g},x_{2+g},\ldots,x_{66+g}\right\}\in X_n\right)$, the expected output data $(T_j=\left\{x_{67+g}\right\},\ g\in\left\{0,1,2,3,\ldots,66\cdot m-67\right\}$, and m is the number of the date) are input into the network, that is, $net=newrb\left(X_g,T_j,goal,spread,mn\right)$.

In order to determine the network parameters based on the sample data in the experiment, the initial range of parameters is firstly obtained by a pre-analysis: *goal* ∈ {0.0001, 0.001, 0.01, 0.1}, *spread* ∈ [0.9, 1.0], *mn* ∈ [30, 70]. Then the network is repeatedly trained by trial in the parameter range of *spread* with a step size of 0.01 and *mn* with a step size of 1. Among them, the RMSE values obtained by the network with *goal* = 0.001, *spread* ∈ [0.9, 1.0], and *mn* ∈ [30, 70] are shown in Table 5. Therefore, the optimal parameter values of RBF neural network are: *goal* = 0.001, *spread* = 0.92, *mn* = 50.

Step 3. Forecasting.

The testing data is input into the determined network for simulation forecast.

3 widely used performance measurement indicators are selected to evaluate the feasibility of forecasting method: RMSE, MAPE, and R^2 (Bai *et al.* 2017; Sun *et al.* 2015; Tsai *et al.* 2009):

Table 5. RMSE of the RBF neural network model with different parameters (*goal* = 0.001)

mn \ *spread*	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00
30	22.548	22.411	22.847	21.892	21.988	22.030	21.794	22.694	22.173	21.826	21.536
31	22.456	21.903	22.791	21.841	21.972	22.012	21.779	22.652	22.064	22.103	21.451
32	22.447	21.885	22.694	21.972	21.946	21.815	21.885	22.240	22.028	22.125	21.608
33	22.436	21.711	22.579	21.909	21.958	21.794	21.647	22.220	21.942	22.550	21.692
34	22.427	21.712	22.294	21.752	21.858	21.900	21.950	22.140	21.775	22.727	21.674
35	22.115	21.453	21.818	21.766	21.767	21.772	21.734	21.892	21.562	22.549	21.553
36	21.997	21.316	21.709	22.034	21.253	21.599	21.736	21.873	21.470	22.487	21.556
37	21.724	21.222	21.669	22.127	21.301	21.711	21.680	21.706	21.262	22.306	21.237
38	21.607	21.050	21.662	22.107	21.223	21.798	21.935	21.649	21.300	22.238	21.023
39	21.491	21.087	21.476	22.395	21.022	21.798	21.896	21.671	20.993	22.125	20.974
40	21.581	21.116	21.046	22.724	21.054	22.159	22.006	21.682	20.977	22.135	20.899
41	21.559	21.115	20.969	22.357	20.998	21.893	21.844	21.993	21.056	21.956	20.940
42	21.715	20.914	21.036	22.563	21.095	21.761	21.847	21.874	21.039	22.012	21.103
43	21.640	20.906	20.872	22.597	21.103	21.600	21.736	21.856	21.010	21.871	21.028
44	21.646	21.044	20.696	22.337	21.107	21.514	21.685	21.878	21.067	21.574	20.678
45	21.394	21.028	20.675	22.156	20.892	21.507	21.945	21.896	21.041	21.628	20.724
46	21.303	20.989	20.357	22.171	20.908	21.359	21.914	21.927	21.152	21.622	20.692
47	21.036	21.091	20.340	22.057	20.857	20.940	21.768	21.924	21.053	21.601	20.786
48	20.918	21.188	20.332	21.994	20.880	20.567	21.740	21.685	21.179	21.603	20.830
49	21.016	20.945	20.341	21.994	20.932	20.549	21.797	21.700	21.169	21.694	20.807
50	21.244	20.991	20.255	21.820	20.806	20.594	21.654	21.805	21.136	21.727	20.651
51	21.239	21.009	20.359	21.793	20.899	20.686	21.639	21.946	21.027	21.690	20.673
52	21.280	20.846	20.399	21.738	20.979	20.866	21.638	21.987	21.043	21.692	20.931
53	21.422	20.904	20.474	21.772	21.017	20.899	21.652	22.099	21.158	21.567	20.845
54	21.355	20.761	20.395	21.810	20.921	21.007	21.664	22.134	21.174	21.570	20.935
55	21.355	20.920	20.386	21.816	20.992	20.903	21.707	22.155	21.218	21.522	20.874
56	21.343	20.923	20.373	21.817	21.160	20.983	21.738	22.122	21.231	21.610	20.860
57	21.552	20.783	20.336	21.783	21.200	21.266	21.751	22.042	21.154	21.562	20.931
58	21.556	20.811	20.319	21.886	21.134	21.247	21.768	22.040	21.071	21.617	20.921
59	21.568	20.795	20.325	21.781	21.156	21.361	21.958	22.088	21.075	21.673	20.987
60	21.489	20.805	20.317	21.858	21.127	21.341	21.818	22.085	21.149	21.752	21.014
61	21.648	20.932	20.256	21.828	20.920	21.418	21.821	22.088	21.129	21.788	21.051
62	21.623	21.181	20.288	21.952	20.958	21.459	21.839	22.153	21.301	21.786	20.899
63	21.586	21.156	20.359	21.623	21.041	21.564	21.497	21.775	21.449	21.659	20.949
64	21.547	21.153	20.453	21.586	20.998	21.438	21.420	21.649	21.556	21.521	21.075
65	21.618	21.250	20.602	21.645	21.103	21.357	21.525	21.757	21.822	21.562	21.094
66	21.598	21.185	20.894	21.635	21.153	21.332	21.530	21.769	21.866	21.305	21.034
67	21.677	21.231	21.253	21.704	21.096	21.199	21.578	21.732	21.973	21.240	20.944
68	21.688	21.177	21.439	21.752	21.224	21.092	21.838	21.711	21.858	21.150	20.973
69	21.832	21.373	21.471	21.708	21.177	21.111	22.090	21.522	21.756	21.124	20.980
70	21.767	21.457	21.481	21.685	21.281	21.070	22.126	21.522	21.956	21.071	21.038

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}};$$

$$MAPE = \frac{100}{n}\cdot\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|;$$

$$R^2 = \frac{r_1}{r_2 \cdot r_3},$$

where:

$$r_1 = \left(n\cdot\sum_{i=1}^{n}\hat{y}_i\cdot y_i - \sum_{i=1}^{n}\hat{y}_i\cdot\sum_{i=1}^{n}y_i\right)^2; \quad (2)$$

$$r_2 = n\cdot\sum_{i=1}^{n}\hat{y}_i^2 - \left(\sum_{i=1}^{n}\hat{y}_i\right)^2; \quad (3)$$

$$r_3 = n\cdot\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2 \quad (4)$$

y_i is the observed value; $\hat{y}_i$ is the forecast value.

Equations (2) and (3) calculate the average absolute and relative errors. Equation (4) measures the goodness of fit.

The total forecast values are substituted into the indicators, the performance values are obtained, as shown in Table 6. The R^2 values are close to 1, indicating that the models have a good fit. According to the definition of MAPE, when 20 < MAPE < 50%, the forecast results are reasonable, and when 10 < MAPE < 20%, the model can get good results (Cao, Liang 2015). It can be seen that the 2 forecasting models are acceptable.

4. Application

4.1. Data decomposition

According to the theory of splitting–integrating method and experimental data, the card number data is used as the main index to analyse the passenger travel behaviour and identify passenger attribute information. The time and date data of the same card number are extracted by adopting SQL. By analysing the correlation of card number, time and date data, it is found that the more frequent the same card number appearing, the more similar the time it appears. On the whole, it presents the similarity variation weekly and daily. It can be concluded that the passengers with this kind of swiping records are likely to be commuters, which have obvious regularity. For the same card number that appears less frequently, the data shows random variation, and it is difficult to identify passenger attribute information.

Taking 5 working days per week as the statistical standard, the number of days that each card number appears is used as the splitting boundary. The total passenger set is divided into regular passenger subset and irregular passenger subset, the rationality of passenger attribute feature identification and the optimal splitting boundary are determined by the subset size and the forecast performance.

Assume that the boundary for extracting the data is *TS*. When the number of days that each card number appears ≥*TS*, the data is extracted. Logically, the possible values of *TS* are {2, 3, 4, 5}. When *TS* = 2, passengers who travel more than 2 days are extracted, and the rest are passengers who travel one day. However, for *TS* = 2, the statistical standard is too low to judge the passenger travel features (Due to the fact that the bus card doesn't have the real name). When *TS* = 3 and *TS* = 4, the proportion of passenger volume within the boundary to the total passenger volume is 27.67% and 23.07% respectively, forming statistical feature and showing regular feature. Therefore, both *TS* = 3 and *TS* = 4 are considered as the splitting boundary of this article. When *TS* = 5, the proportion of passenger volume within the boundary to the total passenger volume is only 8.34%. This percentage is too small to make meaningful.

For the passenger flow during non-working days, there is no subset with obvious statistical feature. Therefore, only the passenger flow during working days is predicted using the splitting–integrating forecasting method.

4.2. Forecasting

In this section, *TS* = 4 is taken as the splitting boundary, the SARIMA model and RBF neural network model are used to forecast subsets respectively. The path of the fully combined, namely the SARIMA–SARIMA, the RBF–RBF, the SARIMA–RBF and the RBF–SARIMA, are adopted. Additionally, the modelling steps of SARIMA and RBF neural network are consistent for the data with different feature, and the parameters of the models are determined by the sample data.

According to the modelling steps above, the optimal model obtained for the regular passenger flow subset is $SARIMA(1, 1, 1)(2, 1, 2)_{66}$ model and for the irregular passenger flow subset, $SARIMA(2, 1, 2)(2, 1, 3)_{66}$ model. Similarly, the RBF neural network models are established, the optimal parameter values of the model for regular passenger flow are: *goal* = 0.001, *spread* = 2.27, *mn* = 22, and the optimal parameter values of the model for irregular passenger are: *goal* = 0.001, *spread* = 1.14, *mn* = 30.

After subsets are predicted by different models, the forecast values are integrated to obtain the total forecast values. Combining the forecast values of regular passenger flow obtained by $SARIMA(1, 1, 1)(2, 1, 2)_{66}$ model with the forecast values of irregular passenger flow obtained by $SARIMA(2, 1, 2)(2, 1, 3)_{66}$ model, thus the total forecast values are obtained by SARIMA–SARIMA model. Similarly, the total forecast values of the RBF–RBF model, SARIMA–RBF model, and RBF–SARIMA model are obtained.

4.3. Results and analysis

The forecast values are substituted into the indicators to get the forecast performance values. In order to evaluate the performance of the models in multiple dimensions, the tests are partitioned into in-sample and out-of-sample test, which are listed in Table 7:

- from the R^2 values in Table 7, it can be seen that the fitting degree obtained by the 4 combined models are very high, up to 0.98. This shows that the SARIMA model and RBF neural network model can be used to forecast subsets and obtain high-quality forecast values;

Table 6. Performance values of direct forecasting method

Forecasting model	In-sample			Out-of-sample		
	RMSE [%]	MAPE [%]	R^2	RMSE [%]	MAPE [%]	R^2
SARIMA	16.58	15.04	0.91	35.48	35.79	0.82
RBF	9.67	17.18	0.97	20.25	19.11	0.87

- compared the performance values obtained by the SARIMA model, the relative improvement of RMSE in-sample and out-of-sample obtained by the SARIMA–SARIMA model is 3.14 and 11.01%, respectively, and the relative improvement of MAPE is 3.26% and 17.68%, respectively. Compared the performance values obtained by the RBF model, the relative improvement of RMSE in-sample and out-of-sample obtained by the RBF–RBF model is 3.01% and 0.81%, and the relative improvement of MAPE is 7.51% and 1.96%, respectively. This means that compared with the forecasting method that does not consider passenger travel behaviour, the splitting–integrating forecasting method can effectively improve the prediction effect. In addition, the method of passenger attribute information identification is reasonable;
- since the SARIMA can't forecast the random data well, the performance values of SARIMA–RBF model are better than that of RBF–SARIMA model as a whole. Therefore, the SARIMA is more suitable for predicting data with regularity. In addition, the performance values obtained by 3 combined models composed of RBF neural network are better than those obtained by SARIMA–SARIMA model, and the RBF–RBF model is the optimal model. This indicates that forecasting models have different adaptability to the data with different features. The RBF neural network model has better adaptability to data with regular and random features;
- when TS = 4, the RMSE, MAPE and R^2 out-of-sample obtained by the RBF–RBF model is 19.44%, 17.15% and 0.88, respectively. When TS = 3, the RMSE, MAPE and R^2 out-of-sample obtained by the RBF–RBF model is 21.16%, 24.20% and 0.85, respectively. It can be seen that when the splitting boundary is TS = 4, all 3 indicators are better. This indicates the boundary has an effect on the splitting–integrating method. The stricter the splitting boundary is, the more regular the data will be, and the better the adaptability of forecasting model will be.

Given that the smaller the evaluation period, the closer to real time, which is more practical for a smart public transport system. The above has compared the overall performance, but a better comparison would involve the daily forecast performance. Figure 8a and Figure 8c are a comparison of daily R^2 values obtained by the SARIMA model and the SARIMA–SARIMA model. It can be seen that the performance values obtained by the splitting–integrating method are better. Figure 8b and Figure 8d show that the observed values are close to the forecast values.

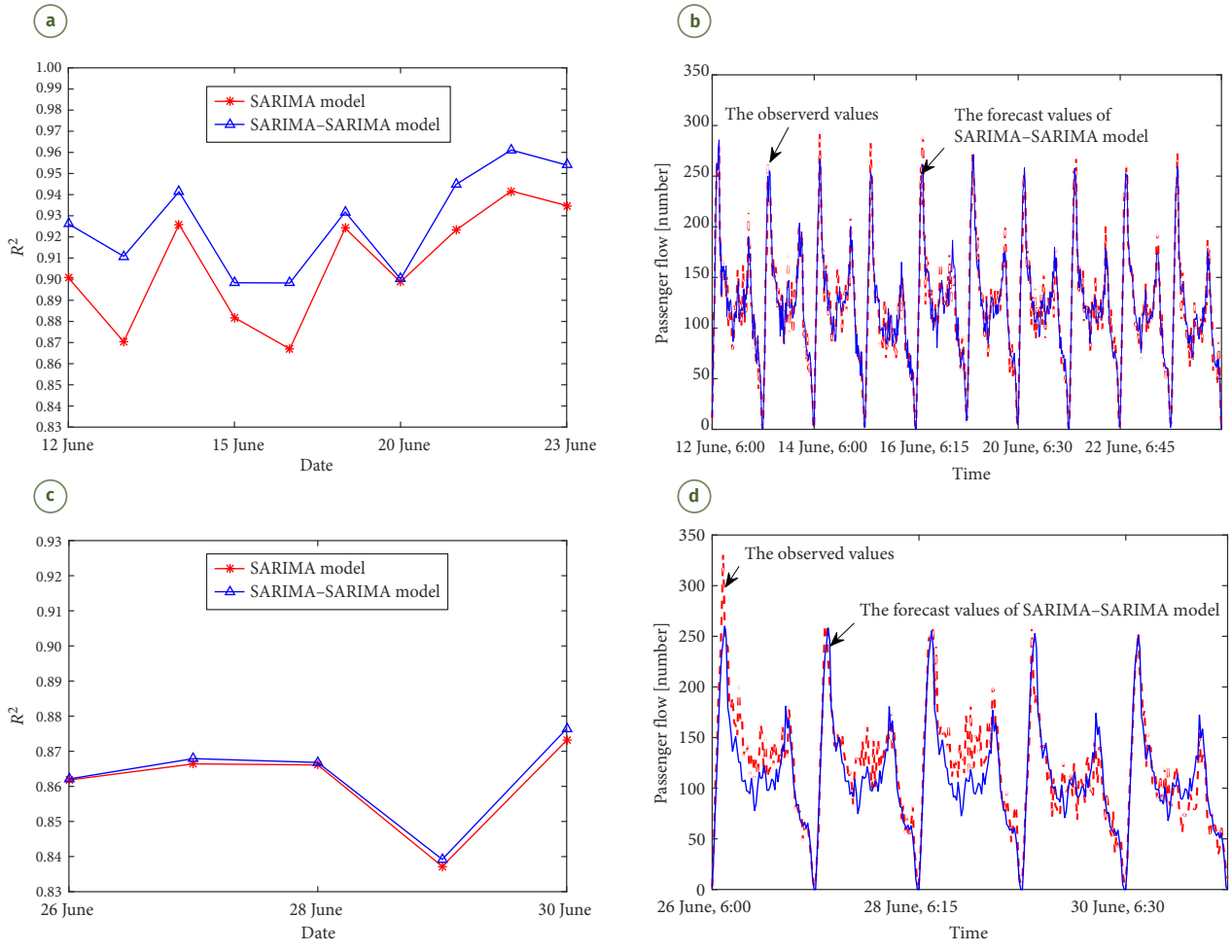


Figure 8. Forecast results of SARIMA model and SARIMA–SARIMA model:
(a) – the comparison of R^2 in-sample; (b) – the comparison of observed values and forecast values in-sample;
(c) – the comparison of R^2 out-of-sample; (d) – the comparison of observed values and forecast values out-of-sample

This indicates that splitting the data does improve forecast performance. In addition, comparing the forecast results in-sample (Figure 8a and Figure 8b) and out-of-sample (Figure 8c and Figure 8d), it's found that the former is better. This illustrates that passenger flows have uncertain variation in the future.

Similarly, the forecast results of the RBF model and the RBF–RBF model are shown in Figure 9. By comparing Figure 8 and Figure 9, it can be found that the RBF–RBF model is better than the SARIMA–SARIMA model, and the RBF model is better than the SARIMA model. This indicates that data with different features are better treated by different forecast models. Additionally, the forecast results in-sample are better than those out-of-sample, further indicating that the variation of passenger flow data is uncertain. Therefore, the prediction should be updated by rolling to maintain the accuracy of the prediction.

5. Conclusions

Main conclusions are:

- passenger flow data with different features are selective for forecasting model and combined path. According to forecast performance, the SARIMA model is more suitable for data with regularity, while the RBF model can be used for data with randomness and complexity. In addition, the RBF-RBF model has the highest applicability;

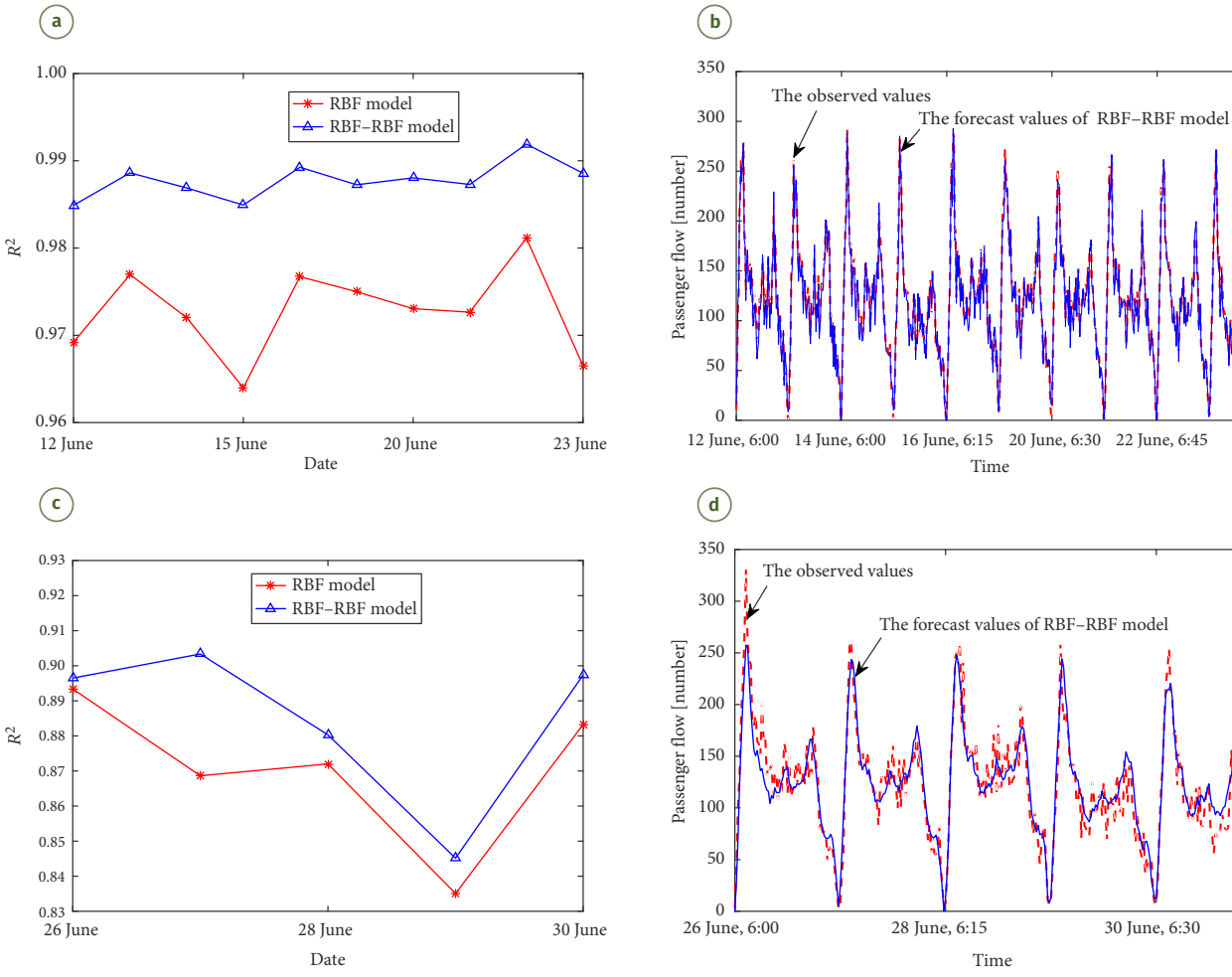


Figure 9. Forecast results of RBF model and RBF–RBF model:
(a) – the comparison of R^2 in-sample; (b) – the comparison of observed values and forecast values in-sample;
(c) – the comparison of R^2 out-of-sample; (d) – the comparison of observed values and forecast values out-of-sample

Table 7. Performance values of splitting–integrating method (TS = 4)

Forecasting model	In-sample			Out-of-sample		
	RMSE [%]	MAPE [%]	R^2	RMSE [%]	MAPE [%]	R^2
SARIMA–SARIMA	13.44	11.78	0.94	24.47	18.11	0.83
RBF–RBF	6.66	9.67	0.98	19.44	17.15	0.88
SARIMA–RBF	8.51	10.78	0.97	20.08	16.96	0.87
RBF–SARIMA	11.78	10.32	0.95	24.06	17.84	0.84

- the location of the splitting boundary (the value of *TS*) has an impact on forecast performance, and the optimal value can be sought based on forecast performance. The higher the value of *TS*, the smaller the subset extracted, and the greater the consistency of the subset. When the subset is too small or too large (*TS* = 2, respectively, 5), the forecast performance isn't ideal. While the best forecast performance appears at *TS* = 4, and the size and consistency are moderate;
- analysing the attribute of data object, understanding the fact behind the data, and trying to adopt the splitting–integrating forecasting path are conducive to improving the forecast performance. Compared with the forecasting method that does not consider passenger travel behaviour and attribute information, the splitting–integrating method achieves better forecast performance. After the passenger flow data is split, the RMSE and MAPE of the SARIMA model is reduced by 11.01% and 17.68%, respectively, and those of the RBF neural network model is reduced by 0.81% and 1.96%, respectively;
- uncertain variation of short-term passenger flow data will affect the prediction effect. All the results in-sample are better than those out-of-sample in the experiment, which show that although the time series follows certain rules, it also has time-varying feature. In applications, if the prediction accuracy needs to be guaranteed, rolling forecast is needed, so that the prediction period isn't too long;
- the short-term passenger flow forecasting considering passenger travel behaviour and attribute information can improve the forecast performance from the perspective of passenger demand, rather than only from the perspective of time of travel. In the context of missing passenger attribute information, this article uses the correlation of card number and the time data to analyse passenger travel behaviour and identify passenger attribute. The results obtained by the SARIMA–SARIMA model and the RBF–RBF model show that the forecasting method considering passenger attribute is better, and the method of passenger attribute identification is effective. This is the contribution of this article. In addition, this article also makes a certain contribution in exploring the adaptability of the data feature and forecasting model. However, the subset still contains different vibration modes, even the regular passenger flow subset also contains high-frequency, low-frequency, non-stationary vibration, resulting in the forecast performance value may not be the best. The EMD, Wavelet, SSA and other decomposition methods can separate the vibration, frequency and noise of the subset, which may further improve the prediction effect. This will be the direction of the next exploration.

Acknowledgements

This research was supported by Natural Science Foundation of Hunan Province (Grant No 2018JJ2537) and Science Progress and Innovation Program of Hunan DOT (Grant No 201723).

This research was also supported by National Natural Science Foundation of China (Grant No 61203162) and Science Progress and Innovation Program of Hunan DOT (Grant No 201949).

Author contributions

Xiaoping Fang, *Weiya Chen* and *Mei Lin* designed the ideas of the study.

Xiaoping Fang and *Weiya Chen* helped to modify the manuscript.

Mei Lin and *Xin Pan* were responsible for data collection and analysis.

Mei Lin established the models and conducted the data experiments.

Disclosure statement

All authors of this article don't have any competing financial, professional, or personal interests from other parties.

References

Bai, L. 2016. Research on the computer algorithm application in urban rail transit holiday passenger flow prediction, in *2016 International Conference on Network and Information Systems for Computers* (*ICNISC*), 15–17 April 2016, Wuhan, China, 233–236. https://doi.org/10.1109/ICNISC.2016.058

Bai, L. 2017. Urban rail transit normal and abnormal short-term passenger flow forecasting method, *Journal of Transportation Systems Engineering and Information Technology* (1): 127–135. https://doi.org/10.16097/j.cnki.1009-6744.2017.01.019 (in Chinese).

Bai, Y.; Sun, Z.; Zeng, B.; Deng, J.; Li, C. 2017. A multi-pattern deep fusion model for short-term bus passenger flow forecasting, *Applied Soft Computing* 58: 669–680. https://doi.org/10.1016/j.asoc.2017.05.011

Cao, C.; Liang, Y.-S. 2015. High-speed railway short-term passenger flow forecasting method based on EMD-BPN, *Technology & Economy in Areas of Communications* (1): 10–12. https://doi.org/10.19348/j.cnki.issn1008-5696.2015.01.003 (in Chinese).

Chen, W.; Pan, X.; Fang, X. 2019. Short-term prediction of passenger flow on bus routes based on k-means clustering combination models, *Journal of South China University of Technology* (*Natural Science Edition*) (4): 83–89. (in Chinese).

Dou, F. 2011. *Research on Holiday Train Distribution Model and Algorithm*. MSc Thesis. Beijing Jiaotong University, China. (in Chinese).

Du, J.; Aultman-Hall, L. 2007. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues, *Transportation Research Part A: Policy and Practice* 41(3): 220–232. https://doi.org/10.1016/j.tra.2006.05.001

Feng, B.; Bao, X.; Wang, Q. 2015. Research of railway passenger volume forecast based on grey and neural network, *Journal of Railway Science and Engineering* (5): 1227–1231. https://doi.org/10.19713/j.cnki.43-1423/u.2015.05.036 (in Chinese).

Gao, Y.; Er, M. J. 2005. NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches, *Fuzzy Sets and Systems* 150(2): 331–350. https://doi.org/10.1016/j.fss.2004.09.015

Haworth, J.; Cheng, T. 2012. Non-parametric regression for space–time forecasting under missing data, *Computers, Environment and Urban Systems* 36(6): 538–550. https://doi.org/10.1016/j.compenvurbsys.2012.08.005

Li, L.-H.; Zhu, J.-S.; Qiang, L.-X.; Qiao, Q.-J. 2017. Study on forecast of high-speed railway short-term passenger flow based on random forest regression, *Railway Transport and Economy* (9): 12–16. https://doi.org/10.16668/j.cnki.issn.1003-1421.2017.09.03 (in Chinese).

Lu, X. Q. 2016. *Bus Passenger Flow Analysis and Prediction Based on Transport IC Card Big Data*. MSc Thesis. Guangdong University of Technology, Guangzhou, Guangdong, China. (in Chinese).

Ma, H. H.; Guo, Q. R.; Ding, C. C. 2016. *Eviews Statistical Analysis and Application*. China: Beijing. (in Chinese).

Meng, P.-C.; Li, X.-Y.; Jia, H.-F.; Li, Y.-Z. 2018. Short-time rail transit passenger flow real-time prediction based on moving average, *Journal of Jilin University* (*Engineering and Technology Edition*) (2): 448–453. https://doi.org/10.13229/j.cnki.jdxbgxb20161256 (in Chinese).

Qiu, D.-G.; Yang, H.-Y. 2013. A short-term traffic flow forecast algorithm based on double seasonal time series, *Journal of Sichuan University* (*Engineering Science Edition*) (5): 64–68. https://doi.org/10.15961/j.jsuese.2013.05.001 (in Chinese).

Sfetsos, A.; Siriopoulos, C. 2004. Combinatorial time series forecasting based on clustering algorithms and neural networks, *Neural Computing & Applications* 13(1): 56–64. https://doi.org/10.1007/s00521-003-0391-y

Smith, B. L.; Williams, B. M.; Oswald, R. K. 2002. Comparison of parametric and nonparametric models for traffic flow forecasting, *Transportation Research Part C: Emerging Technologies* 10(4): 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8

Sun, Y.; Leng, B.; Guan, W. 2015. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system, *Neurocomputing* 166: 109–121. https://doi.org/10.1016/j.neucom.2015.03.085

Teng, J.; Chen, S. 2015. Modified bus passenger flow forecasting model based on integrating ARIMA with neural network, in *CICTP 2015: Efficient, Safe, and Green Multimodal Transportation*, 25–27 July 2015, Beijing, China, 1300–1310. https://doi.org/10.1061/9780784479292.119

Tsai, T.-H.; Lee, C.-K.; Wei, C.-H. 2009. Neural network based temporal feature models for short-term railway passenger demand forecasting, *Expert Systems with Applications* 36(2): 3728-3736. https://doi.org/10.1016/j.eswa.2008.02.071

Vlahogianni, E.; Karlaftis, M. 2011. Temporal aggregation in traffic data: implications for statistical characteristics and model choice, *Transportation Letters: the International Journal of Transportation Research* 3(1): 37–49. https://doi.org/10.3328/TL.2011.03.01.37-49

Wang, D.; Hu, K.; Han, X. 2015a. Research on the application of intelligent algorithm in short-term traffic flow forecast, in *Proceedings of the 2015 International Conference on Automation, Mechanical Control and Computational Engineering*, 24–26 April 2015, Ji'nan, China, 275–280. https://doi.org/10.2991/amcce-15.2015.50

Wang, Y.; Han, B.-M.; Zhang, Q.; Li, D.-W. 2015b. Forecasting of entering passenger flow volume in Beijing subway based on SARIMA model, *Journal of Transportation Systems Engineering and Information Technology* (6): 205–211. https://doi.org/10.16097/j.cnki.1009-6744.2015.06.031 (in Chinese).

Wang, P.; Wu, C.; Gao, X. 2016. Research on subway passenger flow combination prediction model based on RBF neural networks and LSSVM, in *2016 Chinese Control and Decision Conference* (*CCDC*), 28–30 May 2016, Yinchuan, China, 6064–6068. https://doi.org/10.1109/CCDC.2016.7532085

Wang, W.; Cheng, H. 2016. Application of RBF neural network in the forecast of Shanghai railway short-term passenger flow, *Intelligent Computer and Applications* (6): 79–83. (in Chinese).

Williams, B. M.; Hoel, L. A. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results, *Journal of Transportation Engineering* 129(6): 664–672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664)

Xie, Z.-Y.; Jia, L.-M.; Qin, Y.; Wang, L. 2013. Passenger flow parameter prediction algorithm of comprehensive passenger transport hub based on RBF neural network, *Transactions of Beijing institute of Technology* 33(S1): 44–47. Available from Internet: http://journal.bit.edu.cn/zr/en/article/id/2013S111 (in Chinese).

Zhang, C.-H.; Song, R.; Sun, Y. 2011. Kalman filter-based short-term passenger flow forecasting on bus stop, *Journal of Transportation Systems Engineering and Information Technology* (4): 154–159. https://doi.org/10.16097/j.cnki.1009-6744.2011.04.019 (in Chinese).

Zhao, Z. 2018. Research on passenger flow prediction of urban rail transit based on combined model, *Science and Technology & Innovation* 4: 87–88. https://doi.org/10.15913/j.cnki.kjycx.2018.04.087 (in Chinese).

Zhong, M.; Sharma, S.; Lingras, P. 2006. Matching patterns for updating missing values of traffic counts, *Transportation Planning and Technology* 29(2): 141–156. https://doi.org/10.1080/03081060600753461

Zhu, X. X. 2017. *Forecast of Short-Term Bus Passenger Flow Based on IC Card Data*. MSc Thesis. Dalian University of Technology, Liaoning, China. (in Chinese).

Zuo, K. L. 2016. *Research on Passenger Flow Forecasting Method and Application in Different Level of Time Based on IC Smart Card Data*. MSc Thesis. Southeast University, Nanjing, Jiangsu, China. (in Chinese).