International Journal of Hybrid Information Technology
Vol.9, No.2 (2016), pp. 215-226
http://dx.doi.org/10.14257/ijhit.2016.9.2.19

A Combined Forecasting Model for Passenger Flow Based on GM and ARMA

Yunjian Jia[1,a], Peihua He[1,b], Shuguang Liu[2,c] and Lei Cao[2,d]

[1]*College of Communication Engineering, Chongqing University, China*
[2]*Institute of Electronic Information & Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, China*
[a]*yunjian@cqu.edu.cn*, [b]*hepeihua@cqu.edu.cn*, [c]*liushuguang@cigit.ac.cn*, [d]*caolei@cigit.ac.cn*

Abstract

In this paper, we first comparative analysis the existing prediction methods. Based on the GM and ARMA, we propose a new combined forecasting model which integrated the advantage of the GM is suitable for medium and long term forecast, the GM algorithm is simple and the ARMA is suitable for short time forecast. Moreover, we use the rail traffic data to verify this model. The results show that the combined forecasting model we proposed is of high forecast precision, and the combined forecasting model is better than the single forecasting model.

Keywords*: Short-time forecast, GM, ARMA, combination forecasting, passenger flow*

1. Introduction

With the rapid development of Chinese economy, accelerating urbanization brings on many such thorny problems as city space, population density, and transportation. To solve the problem of urban traffic pressure, rail systems is being widely used. Compared with other modes of transport, rail system is quicker, safer and more convenient.

The existing prediction methods mainly include Artificial Neural Network [1-2], Support Vector Machine [3], GM [4] and *etc*. These methods began early and have come to be quite mature. However, each of them has defects which make the accuracy low, and they are generally used in medium and long term prediction. Compared with these prediction methods for medium and long-term, prediction methods for short-term is less and start later. Short-term prediction methods mainly include: Kalman Filtering [5], Wavelet Transform [6], Neural Network [7], Autoregressive Moving Average (ARMA) [8], etc. However, all of them have defects: Kalman Filtering requires that the deviation of the data series should be in linear distribution, thus it is not applied to complicated nonlinear distribution; As for Neural Network and Wavelet Transform, the problem is that the modification of observation scale has great affect on the characteristics of sequence and the selection of algorithm parameter; ARMA model is easy in algorithm yet has significant deviation in the prediction of passenger flow with complicated variations. Restricted by their own characteristic and applicable condition, Single forecasting models could not achieve satisfactory effects in prediction precision, real-time character and portability. Combination forecast model which is a research hotspot can overcome these limitations of single model, also it can integrate the advantage of various models and increase forecasting accuracy.

In view of the existing prediction methods are most for medium and long-term prediction and lack of real-time, we propose a combination forecast model based on GM and ARMA. This combination model has the advantage of high precision, simple operation and easy grasping; also the combination model is real-time. At last, we take the

ISSN: 1738-9968 IJHIT

rail traffic passenger flow data of Chongqing as the research object to verify this combination model proposed. The results showed that the model proposed work well.

2. Principle Introduction

2.1. Grey Model

Grey system theory, proposed by Deng Julong, is one of the major methods for studying and solving problems under uncertainty. The basic idea of gray prediction is to use a known data sequence according to certain rules to constitute a dynamic and non-dynamic white block, and then build the gray model in term of changes and rules. GM (1,1) is the most commonly used model, which consists of a first-order differential equation model that contains only constitute a single variable. Assume that:

$$M^{(0)} = \left(M_{(1)}^{(0)}, M_{(2)}^{(0)}, ..., M_{(i)}^{(0)}\right) \tag{1}$$

Where $M_{(k)}^{(0)} \geq 0, k = 1, 2, ..., n$, is an original sequence.

And $M^{(1)}$ is the 1-AGO sequence of $M^{(0)}$ as follows:

$$M^{(1)} = \left(M_{(1)}^{(1)}, M_{(2)}^{(1)}, ..., M_{(i)}^{(1)}\right) \tag{2}$$

Where $M_{(k)}^{(1)} = \sum_{m=1}^{k} M_{(m)}^{(0)}, k = 1, 2, ..., n$.

Grey model can be defined as:

$$\frac{\mathrm{d}M^{(1)}}{dt} + aZ^{(1)}(k) = \mathrm{u} \tag{3}$$

Estimates grey parameter $\hat{a}, \hat{u}$ using least squares method, according to Eq(3), then

$$\hat{M}_{(k+1)}^{(1)} = [M_{(1)}^{(0)} - \frac{\hat{u}}{\hat{a}}]e^{-\hat{a}k} + \frac{\hat{u}}{\hat{a}}, (k = 1, 2, ..., n) \tag{4}$$

Calculate the value of $\hat{M}_{(k+1)}^{(1)}$, then we can get the predicted value of the primitive data at the time (k+1):

$$\hat{M}_{(k+1)}^{(0)} = \hat{M}_{(k+1)}^{(1)} - \hat{M}_{(k)}^{(1)}, (k = 1, 2, ..., n) \tag{5}$$

2.2. ARMA Model

For a smooth, normal, zero-mean time series $\{\mathrm{x}_t\}$, the value for each x_t is not only related to its p step, but also to the stimulation $\varepsilon_{\mathrm{t}-1}, \varepsilon_{\mathrm{t}-2}, \ldots, \ \varepsilon_{\mathrm{t}-\mathrm{q}}$ $(p, \ q = 1, 2 \ldots\ldots)$ of q step. According to the idea of multiple linear regressions we can get a general ARMA model:

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-1} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{6}$$

Where ε_t is white noise sequence with the zero mean and variance σ^2 ; $\varphi_i \left(i = 1, 2, \ldots, p\right)$ is called autoregressive coefficients, $\theta_{\mathrm{j}} \left(\mathrm{j} = 1, 2, \ldots, q\right)$ is called moving average coefficients.

The Eq(6) is called p orders Autoregressive q orders Moving Average Model, denoted by ARMA (p, q). For further explanation, define delay operator B as $\mathrm{B}x_t = x_{t-1} \mathrm{B}^k x_t = x_{t-k}$, so ARMA model can be abbreviated as:

$$\phi \mathrm{B}x_{\mathrm{t}} \ = \ \Theta(\mathrm{B})\varepsilon_{\mathrm{t}} \tag{7}$$

The formula $\phi B = 1 - \varphi_1 B - \varphi_1 B^2 - ... - \varphi_p B^p$ refers to the p-order autoregressive coefficients polynomial, and the formula $\Theta(B) = 1 - \theta_1 B - \theta_1 B^2 - ... - \theta_1 B^q$ is the **Error! Reference source not found.**q-order moving average coefficients polynomial.

2.3. Combination Forecasting Model

Combination forecasting refers to combining all the forecasting results achieved by several forecasting methods through building a combined forecasting model. Compared with single forecasting models, combined forecasting model can effectively reduce the impact from random factors so as to increase the accuracy and stability of forecasting because it can maximally utilize all kinds of forecasting samples information.

Suppose that there are k $(k \geq 2)$ kinds of forecasting methods to the same forecasting problem. Using y_t, f_{it} and e_{it} $(e_{it} = y_t - f_{it}; i = 1,2,...,k; t = 1,2,...,n)$ to represent the t[th] actual observed value, the forecast value and error of the i[th] method respectively, then the weight of the i[th] method in the combination forecasting is $w_i = \left\{ \sum_{i=1}^{k} w_i = 1 \right\}$. Suppose the forecast value and error of the t[th] combination forecasting is f_{ct} and e_{ct} $(t = 1,2,...,n)$, there will be $f_{ct} = \sum_{i=1}^{k} w_i f_{it}, e_{ct} = y_t - f_{ct} = \sum_{i=1}^{k} w_i e_{it}$. If the sum of squares of the forecasting method's forecast error is e_c^2, then there will be:

$$e_c^2 = \sum_{t=1}^{n} e_{ct}^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} [w_i w_j \{ \sum_{t=1}^{n} e_{it} e_{jt} \}] = w^T E w \tag{8}$$

Where $w = (w_1, w_2, ..., w_k)^T$ is the combination weight vector; $E = (c_{ij})_{k \times k}$ is the matrix of the forecast error, achieved by k kinds of single forecast error and $c_{ij} = \sum_{t=1}^{n} e_{it} e_{jt}$ here.

From the formula above, we know that the numerical value of the sum of squares of the forecast methods' forecast error is concerned with forecast error information E and the combination weight vector w. The information matrix E of the forecast error is decided by the k kinds of forecast methods participating in the combination. Once the information matrix E of the forecast error is determined, the determination of combination weight vector through linear programming is to look for the optimal combination weight vector under the circumstance of the minimal combination forecast error.

$$\begin{cases} \min e_c^2 = w^T E w, \\ s.t. R_k^T w = 1, \\ w > 0, \end{cases} \tag{9}$$

Where $R_k = (1,1,...,1)_{1 \times k}^T$ and its elements are all 1.

As long as all the single forecast methods' errors are known, it is possible to work out the most effective weight vector which multiplied with single forecast value, resulting in the combination forecast.

Most of the current forecast models focus on middle or long term forecast. The thesis uses the generalizability of the grey level model and the high-level accuracy of ARMA model to construct a new combination model to make prediction and analysis of the passenger flow volume of the light rail.

Construct grey level model based on the historic passenger flow volume of the light rail for the past i days, and forecast the passenger flow volume of the $i+1^{th}$ day as M_{i+1}. Count the passenger flow volume and work out the ratio r_{it} of the of passenger flow volume of t^{th} phase of i^{th} day and the total passenger flow volume.

$$r_{it} = \frac{H_{it}}{M_i}, i = 1, 2, ..., k \tag{10}$$

$$\bar{r}_t = \frac{1}{n}\sum_{i=1}^{n} r_{it}, i = 1, 2, ..., n \tag{11}$$

M_i represents the total passenger flow volume of the i^{th} day, H_{it} demonstrates the passenger flow volume of t^{th} phase of i^{th} day, $\bar{r}_t$ illustrates the average ratio of the passenger flow volume of t^{th} phase in whole day.

Then, the passenger flow volume G_t of the t^{th} phase of $i+1^{th}$ day can be presented as:

$$G_t = M_{i+1} \cdot \bar{r}_t, t = 1, 2, ..., n \tag{12}$$

Construct ARMA model to forecast the passenger flow volume of the t^{th} phase of $i+1^{th}$ day and mark it as H_t. Then the combination forecasting model can be constructed based on the grey level model and the ARMA model as:

$$S_t = \alpha \cdot G_t + \beta \cdot H_t, t = 1, 2, ..., n. \tag{13}$$

Where α, β is the combination weighting coefficient of this combination model; G_t is the passenger flow volume of the t^{th} phase, forecasted by the grey level model; H_t is the passenger flow volume of the t^{th} phase, forecasted by the ARMA model.

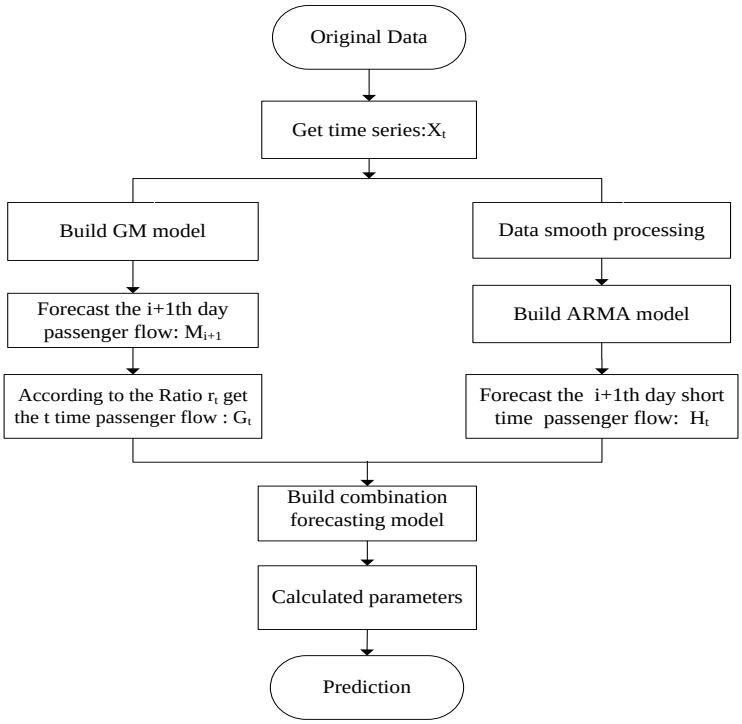

Figure 1. The Flow Chart of the Combination Forecasting Scheme

Work out the combination weighting coefficient α, β through the method of linear programming according to G_t , H_t . Then, using the above formula, a combination forecasting model under the minimal sum of squares is resulted. Refer to flow figure 1.

3. Cases of Application of the Model

3.1 Data Sets used by the Model

In order to evaluate the forecasting ability of this combination forecasting model, this thesis focuses on the passenger flow volume of Lianglukou in Chongqing in June, 2012, sampling every 10 minutes from 7 am to 10 pm. It took 4 weeks (5 days per week, excluding the weekends), having 1,800 pairs of passenger flow volume which are the original time series.

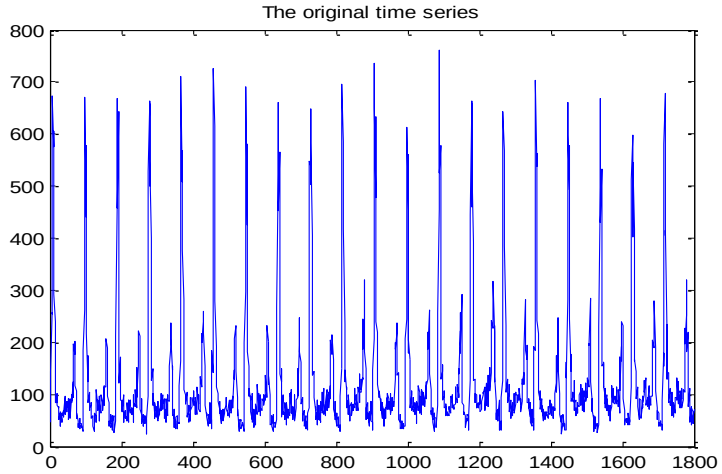

Figure 2. The Original Time Series

This thesis uses the passenger flow data of the first 19 working days of June as the basic data, making a short-term forecasting of the passenger flow of the 12 phases in the 20th working day from 10 am to 12 am. First we use the grey level model to forecast the passenger flow data of the 20th working day. Since the passenger flow data is cyclic, the passenger flow volume of the time t is forecasted according to the ratio of the historic t time. Then, after data preprocessing, the ARMA model is built to forecast the passenger flow volume of the tth phase of the 20th day. At last, this thesis constructs a combination forecasting model based on the passenger flow volume of the tth phase of the 20th day forecasted by the grey level model and the ARMA model. Then, the passenger flow of the 12 phases in the 20th working day from 10 am to 12 am can be forecasted though this combination model.

3.2. GM (1,1) Model Predictions

Taking rail traffic passenger data in the first 19 days of June 2012 as the data source to predict the 20th day's and ultimately determines the parameters $\hat{a} = -0.0064$, $\hat{u} = 10638$. Take the $\hat{a}, \hat{u}$ values into the formula we can get

$$\hat{M}_{(k+1)}^{(1)} = 1678921.2347\exp(0.0063778k)\text{-}1667965.2347(k = 1,2,...,n)$$

Using the constructed GM (1,1) model to predict the 20th day railway traffic passenger flow, we can get M_{20} is 12049.

3.3. ARMA Model Predictions

(1) From Figure 2, we can see the original sequence is non-stationary series, which need a smooth process, where the smooth processing we used two difference method. Generally, d-order difference: $\nabla^d X_t = (1 - B)^d X_t$, where ∇^d **Error! Reference source not found.** is called d-order differential operator.

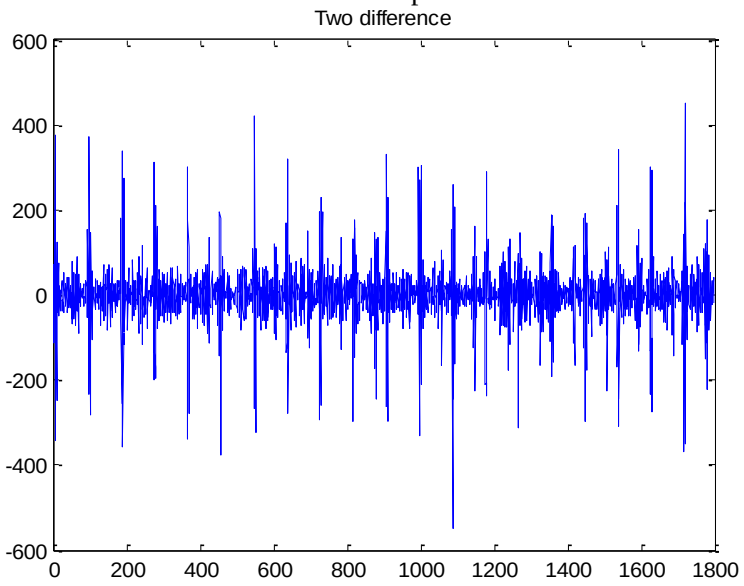

Figure 3. The Two Difference

(2) Determine the model by AIC criterion for ARMA model;

(3) After several experiments, the ARMA (4, 1) model is choosed and parameters are obtained by using the method of least squares estimation:

$$\varphi_1 = 1.243, \varphi_2 = -0.1877, \varphi_3 = 0.002796, \varphi_4 = -0.1168 , \theta_1 = 0.2166$$

Thereby ARMA model is established as follows:

$$(1 - 1.243B + 0.1877B^2 - 0.002796B^3 + 0.1168B^4)x_t = (1 - 0.2166B)\varepsilon_t$$

(4) Forecast: use the established forecasting model to predict each phase passenger number of the 20th day railway traffic.

3.4. Combination Forecasting Model

Respectively work out the i $(i = 1,2,...,19)^{th}$ day first, use the proportion r_{it} of the total traffic in the traffic in one day, and then a weighted average $\bar{r}_t$ is obtained, then the 20th day railway traffic can be expressed as:

$$G_t = M_{20} \cdot \bar{r}_t, t = 1,2,...,n$$

Gray model predicts 18 points denoted as $G_1, G_2, ...G_{18}$, ARMA model predicts 18 points denoted as $H_1, H_2, ...H_{18}$. The prediction error recorded as $\begin{cases} e_{11}, e_{12}, ..., e_{118} \\ e_{21}, e_{22}, ..., e_{218} \end{cases}$.The establishment of combination forecasting model is:

$$S_t = \alpha G_t + \beta H_t, t = 1, 2, ..., n$$

get the error message matrix E:

$$\mathrm{E}=\begin{bmatrix} \sum_{t=1}^{18} e_{1t}^2 & \sum_{t=1}^{18} e_{1t} e_{2t} \\ \sum_{t=1}^{18} e_{1t} e_{2t} & \sum_{t=1}^{18} e_{1t}^2 \end{bmatrix}$$

Where $\begin{cases} \min e_c^2 = [\alpha, \beta] E [\alpha, \beta]^T, \\ \alpha + \beta = 1, \\ \alpha, \beta > 0, \end{cases}$

Though linear programming approach the combining weights finalized coefficient values are 0.5444917, 0.4555083.

Each of the predictive value and the relative error of the gray model, ARMA model, the combined model are collected in Table 1.

Table 1. The Predicted Value and the Actual Value Comparison of 7: 00-10: 00

t	actual value	GM		ARMA		GM-ARMA	
		predict value	relative error	predict value	relative error	predict value	relative error
1	42	36	0.132857	47	0.119047	41	0.023809
2	52	59	0.134615	58	0.115384	58	0.115384
3	92	104	0.130434	107	0.133043	105	0.121304
4	150	168	0.12	169	0.126666	168	0.12
5	170	176	0.035294	160	0.058823	168	0.011764
6	241	272	0.108630	263	0.091286	267	0.087883
7	268	281	0.048507	257	0.041044	270	0.010462
8	596	629	0.055369	635	0.065436	631	0.058724
9	632	670	0.060126	664	0.050632	667	0.055379
10	698	714	0.022922	683	0.021489	700	0.012865
11	516	547	0.060077	541	0.048449	544	0.050263
12	512	500	0.052734	555	0.083984	525	0.025390

13	289	259	0.083806	326	0.098027	288	0.014602
14	215	230	0.069767	232	0.079069	230	0.062767
15	148	158	0.104864	171	0.085405	164	0.068108
16	118	145	0.111864	95	0.104915	122	0.053898
17	104	120	0.093846	87	0.103461	105	0.010615
18	109	116	0.064220	100	0.092568	108	0.020174

It can be seen that the relative error of the combined model is significantly lower than the predicted value of the gray model and the ARMA model. Use the combined model we get MAPE=0.050188, respectively, lower than predicted with gray model MAPE=0.082774 and with ARMA model prediction MAPE=0.084374.

3.5. Prediction

Construct the model of combined forecasting model as:

$$S_t = 0.5444917G_t + 0.4555083H_t, t = 1, 2, ..., n$$

Where G_t is the value predicted by the model with gray, H_t is the predict results of the ARMA model. Then predict the total value from 10:00 am to 12:00 am on the 20th traffic, data obtained are shown in Table 2

Table 2. Prediction Results of 10: 00-12: 00

t	actual value	GM		ARMA		GM-ARMA	
		predict value	relative error	predict value	relative error	predict value	relative error
1	102	97	0.049020	95	0.068627	96	0.057951
2	83	90	0.084337	91	0.096386	90	0.089825
3	88	94	0.068182	97	0.102273	95	0.083711
4	85	92	0.082353	81	0.047059	86	0.023405
5	86	80	0.069767	81	0.05814	80	0.064471
6	80	75	0.062500	84	0.05	79	0.011255
7	97	90	0.072165	88	0.092784	89	0.081557
8	82	73	0.109756	71	0.134146	72	0.120866
9	87	78	0.103448	75	0.137931	76	0.119155
10	60	64	0.066667	54	0.1	59	0.009251
11	70	78	0.114286	64	0.085714	71	0.023184
12	62	66	0.064516	56	0.096774	61	0.008953

Each of the predictive value of the gray model, ARMA model, the combined model are collected in Figure 4. Figure 4 is the result of the relative error comparison for the gray model, ARMA model, the combined model.

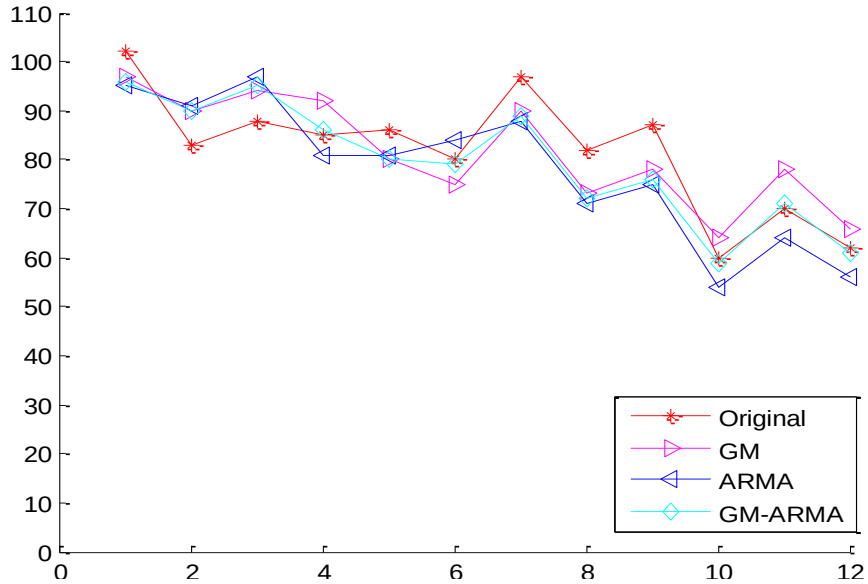

Figure 4. Phases of the Predicted Value and the Actual Value Comparison

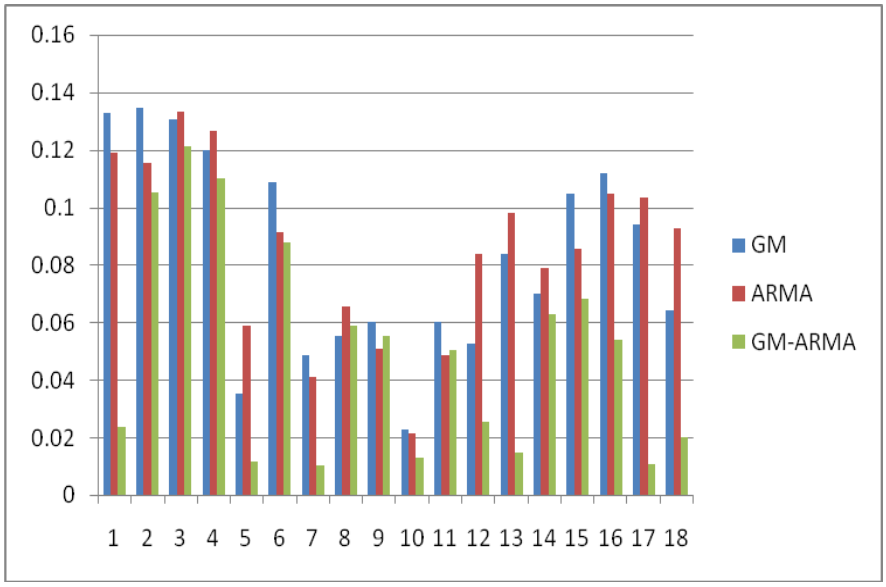

Figure 5. Comparison of the Relative Error for Each Prediction

It can be seen that the relative error of the combined model is significantly lower than the predicted value of the gray model and the ARMA model. Use the combined model we get MAPE =0.057799, respectively, lower than predicted with gray model MAPE = 0.078916 and with ARMA model prediction MAPE = 0.089153.

4.Conclusions

In this paper, we first comparative analysis the existing prediction methods, we find that the existing prediction methods are most for medium and long term prediction, and most of them have defects such as low prediction precision, process complexity, more

limitations and so on. In this paper, we present a new combined forecasting model based on GM and ARMA. The proposed scheme take advantage of GM that the algorithm is simple and is suitable for medium and long term forecast , also the combined forecasting model makes good use of the ARMA for it has high precision and better performance on short-term prediction. Then we use the actual rail traffic data to verify this mode. Simulation results show that our model can accurately describe and predict passenger flow, and the MAPE is lower than the GM model and the ARMA model, it means that the new combined forecasting model we proposed perform better than single model. In other words, the proposed scheme has great improvement compared with other prediction methods. Therefore, our method can increase forecasting accuracy and is expected to have reference value in the forecast field.

Acknowledgements

This work is sponsored in part by the Decision-Making and Consulting Program of Chongqing, China (No. cstc2014jccxC0017).

References

[1] Y. Wei and M. C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks", Journal of Transportation Research Part C: Emerging Technologies, vol. 21, no. 1, **(2012)**, pp. 148-162.
[2] H. Yin, S. C. Wong, J. Xu, *et al.*, "Urban traffic flow prediction using a fuzzy-neural approach", Journal of Transportation Research Part C: Emerging Technologies, vol. 10, no. 2, **(2002)**, pp. 85-98.
[3] M. Castro-Neto, Y. S. Jeong, M. K. Jeong, *et al.*, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions", Journal of Expert systems with applications, vol. 36, no. 3, **(2009)**, pp. 6164-6173.
[4] J. L. Deng, "Grey forecasting and decision-making", Wuhan: Huazhong University of science and technology press, **(1986)**.
[5] C. H. Zhang, R. Song and Y. Sun, "The short-term passenger flow prediction based on kalman filtering", Journal of Transportation systems engineering and information technology, vol. 11, no. 4, **(2011)**, pp. 154-159.
[6] C. L. Ren, C. X. Cao and J. Li, "Short-term traffic prediction research based on wavelet neural network", Journal of Science technology and engineering, vol. 11, no. 21, **(2011)**, pp. 5099-5103.
[7] W. Zheng, D. H. Lee and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach", Journal of Transportation engineering, vol. 132, no. 2, **(2006)**, pp. 114-121.
[8] K. Kalpakis, D. Gada and V. P. Unta, "Distance measures for effective clustering of ARMA time series", Proc of the IEEE International Conference on Data Mining. San Jose, USA: [s. n.], **(2001)**, pp. 273 - 276.

Authors

Yunjian Jia, He received his B.S. degree from Nankai University, China, and his M.E. and Ph.D. degrees in Engineering from Osaka University, Japan, in 1999, 2003 and 2006, respectively. From 2006 to 2012, he was with Central Research Laboratory, Hitachi, Ltd., where he engaged in research and development on wireless networks, and also contributed to LTE/LTE-Advanced standardization in 3GPP. He is now a professor at the College of Communication Engineering, Chongqing University, Chongqing, China. He is the author of more than 60 published papers, and 20 granted patents. His research interests include radio access technologies, mobile networks, and IoT. Dr. Jia has won several prizes from industry and academia including the IEEE Vehicular Technology Society Young Researcher Encouragement Award, the IEICE Paper Award, the Yokosuka Research Park R&D Committee YRP Award, and the Top 50 Young

Inventors of Hitachi. Moreover, he was a research fellowship award recipient of International Communication Foundation in 2004, and Telecommunications Advancement Foundation Japan in 2005.

Peihua He, She received her B.E degree in Communication Engineering from Chongqing University, China, in 2014. She is currently working toward her M.E degree in Information and Communication Engineering in the same university. She mainly engages in the big data, data mining and wireless communication.

Lei Cao, He received a double bachelor degree majoring in biological technology and communication engineering from Chongqing University of Posts and Telecommunications, in 2011. He received his Master's degree in Communication and Information System from Institute of Electronic Information & Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, China in 2014. His research directions include the big data ,indoor positioning and data mining.