Traffic Accident Time Series Prediction Model Based on Combination of ARIMA and BP and SVM

Xiaorui Shao
Information collaboration Engineering, Pukyong national university, Busan city, South of Korea
Email: shaoxiaorui1994@gmail.com

Lester L Boey and Yifei Luo
Master Degree in Predictive Analytics, Curtin University, Perth, Australia
Email: {Lesterboeyail, yifeilyf}@gmail.com

***Abstract*—Intelligent transportation is an important part of the smart city. Predict the traffic accidents accurately which contributes to the scientific management of the city and utilizes the public spaces more efficiently. In this paper, construct a combination forecasting model by using the reciprocal variance method based on Autoregressive Integrated Moving Average Model(ARIMA). Using the constructed combination model to predict traffic events related index. Firstly, ARIMA and BP, ARIMA and Support Vector Machine(SVM) models are established, Through comparing, The SVM model is better than a BP neural network model, So, establish the ARIMA (2, 2, 2) and SVM combination model. Also establish the ARIMA (2, 2, 2) and SVM, BP neutral network combination model. The experimental results show that we can improve the accuracy of predicting traffic events related index time series through combination model generally. The ARIMA (2, 2, 2) and SVM, BP neural network combination model, is more accurate than each of single model, also than ARIMA (2, 2, 2) and SVM combination model. We can adopt ARIMA and SVM, BP neural network to predict traffic events index accurately.**

***Index Terms*—ARIMA, BP, SVM, combination prediction model**

I. Introduction

Prediction of Road Traffic Accident related indicators is a method of studying the law of accident changes and predicting the development trend of accidents based on accident data statistics, analysis and excavation. The commonly used traffic prediction methods include statistical regression method [1], [2], time series method [3], [4], Markov chain method [5], gray prediction method [6], support vector machine method, neural network method, and other Nonlinear prediction methods. However, most traffic accident prediction methods use a single model for prediction. To improve the prediction accuracy, Bates and Granger [7] proposed the idea of a combination prediction in 1969. Many scholars have applied the combination prediction to traffic accident prediction zones.

In this paper, we adopt a time series of traffic accidents in Beijing city from 1980 to 2016, first established the ARIMA model, and analyzed the linear change trend of traffic accident time series. Based on this, the BP time series prediction model and SVM time series prediction model are constructed. Using the reciprocal variance method to determine the weight of each model. The ARIMA and BP combined traffic accident time series prediction model, the ARIMA and SVM combined traffic accident time series prediction model, and the BP and SVM combined traffic accident time series prediction model was constructed. Finally, the weights of ARIMA, BP, and SVM models are determined by the same method. The ARIMA, BP, and SVM traffic accident time series prediction models were constructed. The experimental results show that the combined model of the three models is more accurate than the single model and the combined of two models. Therefore, the combined model of ARIMA, BP and SVM can be used to predict the traffic accident time series. And having the highest accuracy.

II. Application of Combination Forecasting Model

A. Data Description

In this paper, we collected time series data set about the number of injured because of a traffic accident in Beijing city from 1980 to 2016 as shown in Table I. And the change trend as shown in Fig. 1. From 1980 to 1994, with a downward trend. However, from 1994 to 1999, it is a rising trend, and then it has a downward trend.

TABLE I. Beijing City Traffic Accident Injured Numbers Statics Data from 1980 to 2016

Date	1980	1981	1982	1983	1984	1985	1986
Injured	7939	7287	6813	6837	6670	4917	5820
Date	1987	1988	1989	1990	1991	1992	1993
Injured	4579	4136	4110	4315	4724	3015	2878
Date	1994	1995	1996	1997	1998	1999	2000

Manuscript received March 5, 2019; revised December 11, 2019.

doi: 10.18178/jtle.7.2.41-46

Injured	3645	3834	4237	5674	8468	10607	10583
Date	2001	2002	2003	2004	2005	2006	2007
Injured	10424	10456	9877	8248	6888	6681	6088
Date	2008	2009	2010	2011	2012	2013	2014
Injured	4474	4420	4703	4506	3615	3359	3362
Date	2016						
injured	2781						

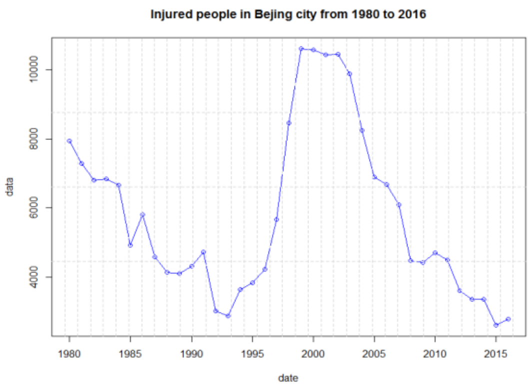


Figure 1. The changing trend

B. Autoregressive Integrated Moving Average Model

Firstly, the detection of white noise is necessary to do. The results show that the p-value is 1.797x10^ (-13), much less than 0.05. Therefore, this time series data is not a white noise sequence. Secondly, Through the correlation analysis of the time series data set, the data is non-stationary, so the data need to be processed with the 2 order difference. The results of the processed data are shown in Fig. 2.

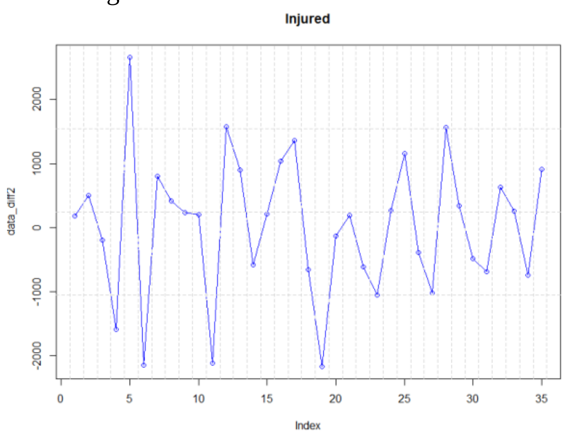


Figure 2. Stationary time series after 2 order difference

The auto-correlation coefficient of the difference sequence dataset Changed within two times of standard deviation after delayed 2 orders. It can be regarded as a 2 orders truncation. When the partial auto-correlation coefficient is delayed to 2 orders. It changes within two times of standard deviation, which can be regarded as 2 order truncation. So the ARIMA model can be defined as ARIMA (0, 2, 2), ARIMA (2, 2, 2), ARIMA (0, 2, 0). According to the minimum information criterion of AIC and BIC, As shown in Table II. The best model is an ARIMA (2, 2, 2). Finally, the P correlation of the residual is detected and the p-value is greater than 0.05, indicating that the defined model is effective. The prediction results are shown in Fig. 3. The blue line is true values, and the red line is forecasting values(fitted values), the green line is for residuals.

TABLE II. THE DETECTION OF AIC AND BIC

Model	AIC	BIC
ARIMA(0,2,2)	585.079	589.7639
ARIMA(2,2,2)	584.3634	589.1401
ARIMA(0,2,0)	590.2183	591.7737

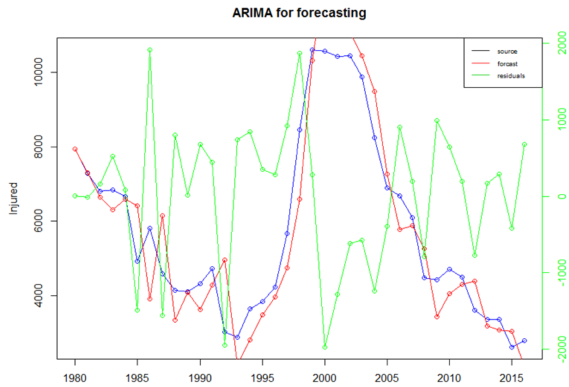


Figure 3. ARIMA (2, 2, 2) prediction results

C. BP Neural Networks

In this paper, the time series data set is one-dimensional data, but BP neural network requires the data be multidimensional, So we need to change the structure of the original data set, The procession is shown in table3. It has been verified normalization of time series is important and necessary to improve the accuracy of predicting by Wang Shuhua [8]. In our paper, The following formula is used to normalize.

$$x'(t) = x(t) / x^n \tag{1}$$

The data sample is x (t), n refers to the number of the max data sample, where n=5. The three-level neural network can realize arbitrary nonlinear mapping from input to output. Therefore, the experiment uses three layers of BP neural network. The number of input layer nodes is determined by our demand, in this paper, we set it is equal to 5. The output layer is the predictor variable, so the number of input layer nodes is equal to 1.And to compute the number of hidden layer nodes, we used 0.618 methods to compute, as shown by the formula(2):

$$m = \begin{cases} n + 0.618(n-t), n \geq t \\ n - 0.618(t-n), n < t \end{cases} \tag{2}$$

There, m represents the number of nodes in the hidden layer and represents the number of input layer nodes, and the t represents the number of the output layer nodes. According to this, we can compute the hidden layer nodes is 7. The BP model input data samples are defined as

$$x = x[x'(t-5), x'(t-4), x'(t-3), x'(t-2), x'(t-1)] \quad (3)$$

Output as $x = x'(t)$. As shown in the following Table III. Finally, the prediction results are shown in as Fig. 4.

TABLE III. BP MODELE DATA SAMPLES PREDICTION

Samples	Input					Output
	$x = x[x'(t-5), x'(t-4), x'(t-3), x'(t-2), x'(t-1)]$					
Array	$x'(t-5)$	$x'(t-5)$	$x'(t-5)$	$x'(t-5)$	$x'(t-5)$	$x'(t)$
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0794
2	0.0000	0.0000	0.0000	0.0000	0.0794	0.0729
3	0.0000	0.0000	0.0000	0.0794	0.0729	0.0681
4	0.0000	0.0000	0.0794	0.0729	0.0681	0.0684
5	0.0000	0.0794	0.0729	0.0681	0.0684	0.0667
6	0.0794	0.0729	0.0681	0.0684	0.0667	0.0492
7	0.0729	0.0681	0.0684	0.0667	0.0492	0.0582
...	...	...	...	...	...	...
33	0.0609	0.0447	0.0442	0.0470	0.0450	0.0362
34	0.0447	0.0442	0.0470	0.0450	0.0362	0.0336
35	0.0442	0.0470	0.0450	0.0362	0.0336	0.0336
36	0.0470	0.0450	0.0362	0.0336	0.0336	0.0262
37	0.0450	0.0362	0.0336	0.0336	0.0262	0.0278

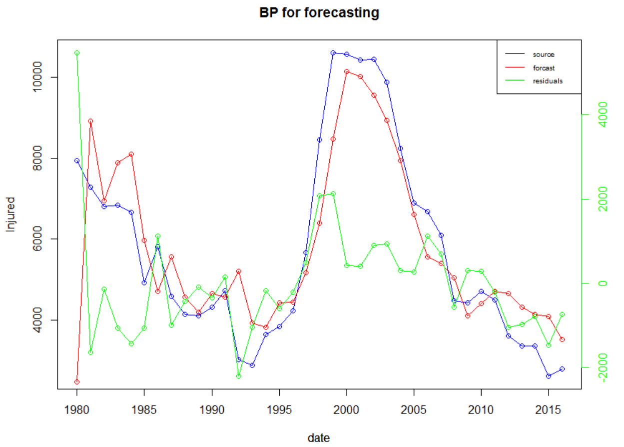


Figure 4. BP prediction results

D. Support Vector Machine Model

The SVM algorithm overcomes the disadvantages of slow convergence rate, small local point, difficult network structure, and a large number of data samples in the training of neural network, which make it more effective in small samples, nonlinear, high dimension. The SVM model is constructed in this paper. First, the following formulas are used to normalize the time series.

$$\bar{x}_i = \frac{x_i - \frac{1}{2}(x_{max} + x_{min})}{\frac{1}{2}(x_{max} + x_{min})} \quad (4)$$

$\bar{x}_i$ is the normalized time series, the same method of BP neural network is used to construct the SVM model. In this paper, the kernel function is selected as a radial, and the cost is set to 10. The prediction results are shown in Fig. 5.

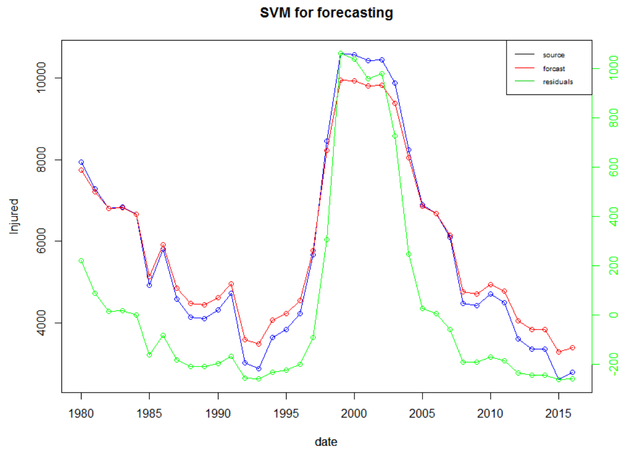


Figure 5. SVM prediction results

E. Comparative Analysis

The prediction accuracy is determined by the relative errors of the above models. The results of predicted and relative errors of the above single prediction models are shown in the Table IV. Ai is the real values of the injured people, Pr is the number of injured people, Fe is the fraction error for the specified model. All of the fraction errors we adopt the absolute values.

TABLE IV. BP MODEL THE RESULTS OF PREDICTED AND THE RELATIVE ERRORS

Date	Ai	ARIMA(2,2,2)		BP		SVM	
		Pr	Fe	Pr	Fe	Pr	Fe
1980	7939	7935	0.0004	2478	0.6879	7757	0.023
1981	7287	7299	0.0017	8928	0.2251	7208	0.0108
1982	6813	6654	0.0233	6945	0.0194	6799	0.0021
1983	6837	6316	0.0761	7889	0.1539	6820	0.0024
1984	6670	6587	0.0123	8103	0.2148	6670	0
1985	4917	6412	0.3041	5978	0.2157	5135	0.0443
1986	5820	3905	0.3291	4705	0.1917	5913	0.016
1987	4579	6141	0.3412	5560	0.2142	4844	0.0579
1988	4136	3332	0.1944	4569	0.1048	4472	0.0813
1989	4110	4095	0.0037	4196	0.021	4449	0.0825
1990	4315	3631	0.1586	4653	0.0783	4618	0.0703
1991	4724	4278	0.0945	4573	0.032	4962	0.0504
1992	3015	4966	0.6472	5204	0.7262	3581	0.1877
1993	2878	2136	0.2577	3922	0.3629	3474	0.2072
1994	3645	2804	0.2306	3816	0.0469	4068	0.1161
1995	3834	3484	0.0913	4424	0.154	4221	0.1009
1996	4237	3956	0.0662	4439	0.0477	4553	0.0745
1997	5674	4749	0.1631	5177	0.0876	5783	0.0192
1998	8468	6597	0.2209	6394	0.245	8230	0.0281
1999	10607	10327	0.0264	8471	0.2014	9946	0.0623

2000	10583	12556	0.1864	10149	0.041	9934	0.0613
2001	10424	11705	0.1229	10019	0.0389	9817	0.0583
2002	10456	11070	0.0587	9561	0.0856	9837	0.0592
2003	9877	10453	0.0584	8938	0.095	9391	0.0492
2004	8284	9492	0.1508	7936	0.0378	8051	0.0239
2005	6888	7277	0.0564	6614	0.0398	6863	0.0036
2006	6681	5777	0.1354	5559	0.168	6676	0.0007
2007	6088	5891	0.0324	5396	0.1137	6152	0.0104
2008	4474	5261	0.1758	5043	0.1271	4758	0.0634
2009	4420	3428	0.2245	4109	0.0703	4708	0.0652
2010	4703	4055	0.1377	4409	0.0625	4945	0.0514
2011	4503	4303	0.0443	4703	0.0443	4776	0.0607
2012	3615	4392	0.2149	4652	0.2868	4049	0.1199
2013	3359	3187	0.0513	4327	0.2881	3845	0.1446
2014	3362	3072	0.0863	4142	0.232	3846	0.1438
2015	2619	3036	0.1594	4083	0.5591	3280	0.2525
2016	2781	2104	0.2435	3513	0.2633	3399	0.2223

It can be seen from the table that the results of the three models are not the same, but we can see that the SVM model is better than the BP model, so the next step, we use the ARIMA (2, 2, 2) model and SVM model.

III. Application of Combination Forecasting Model

The combination model generally adopts the weighted average of every model. So the weighted average is the key point of the combination forecasting model. In this paper, we use the reciprocal method of variance proposed by Bates and Granger [7]. The basic principle of this method is to calculate the square sum of the error of each single prediction model, and then assign the weight of each single prediction model, which according to the minimum sum principle of the square sum of errors. This method provides the academic circle for the study of the combined prediction model. The calculation formula is as follows:

$$w_j = \frac{e_j^{-1}}{\sum_{j=1}^{m} e_j^{-1}} \tag{5}$$

The combination model can be established as follows the formula.

$$X = \sum_{j=1}^{m} w_j \hat{x}_j \tag{6}$$

According to the above comparative analysis, we will build the ARIMA (2, 2, 2) model and the SVM model.

A. ARIMA and SVM Combination Model

According to the principle described above. The weight of the ARIMA model is 0.669489, the weight of the BP model is 0.330511. We create the ARIMA and SVM combination model, Some details are shown in Table V.

Through the Table V, Most of the relative errors of ARIMA (2, 2, 2) model and the SVM combination model is less than the relative error of every single model. It shows that the combined prediction effect of ARIMA and SVM is better than that of a single model.

B. ARIMA and SVM, BP Combination Model

In the above section, we know the combination model can improve the accuracy of predicting model. So we continue to build a combination model of ARIMA and SVM, BP, The weight of ARIMA, SVM, and BP are c (0.1348021, 0.7986492, 0.06654866). The results are shown in Table VI.

The results show that ARIMA (2, 2, 2) and SVM, the BP combination model most of the fractional error are less than each of single model, Also less than ARIMA and SVM combination model. So, we can conclude that in the traffic events, forecasting model, we can adopt the ARIMA and SVM, BP combination model to improve the accuracy of predicting results.

IV. Conclusion

Due to their conditions of single prediction models, it cannot mine the completed data information in time series [9], so the accuracy of predicting results will be affected. Through the combination model, we can improve the accuracy of predicting results in a certain combination method.

TABLE V. ARIMA(2,2,2) and SVM Combination Model

Date	Actual injured (person)	ARIMA (2,2,2) and SVM combined model		
		Predict	Errors	Fractional error
1980	7939	7782.449285	35.016	0.0197
1981	7287	7221.508876	89.2503	0.009
1982	6813	6777.98398	11.8638	0.0051
1983	6837	6747.749713	-402.4426	0.0131
1984	6670	6658.136206	196.686	0.0018
1985	4917	5319.442632	-452.4981	0.0818
1986	5820	5623.313966	-171.5295	0.0338
1987	4579	5031.498105	-287.9221	0.0988
1988	4136	4307.52948	-160.6243	0.0415
1989	4110	4397.922143	-139.3446	0.0701
1990	4315	4475.624346	-766.0719	0.0372
1991	4724	4863.344552	-403.0646	0.0295
1992	3015	3781.071882	-240.6843	0.2541
1993	2878	3281.064634	-280.6239	0.1401
1994	3645	3885.684272	-229.6244	0.066
1995	3834	4114.623925	40.3674	0.0732
1996	4237	4466.624424	474.0752	0.0542
1997	5674	5633.632624	606.1746	0.0071
1998	8468	7993.924773	270.0682	0.056
1999	10607	10000.82536	334.5809	0.0571

2000	10583	10312.93179	440.9471	0.0255
2001	10424	10089.41911	332.1853	0.0321
2002	10456	10015.05293	-10.8729	0.0422
2003	9877	9544.814679	-35.1144	0.0336
2004	8284	8258.872883	134.8822	0.0013
2005	6888	6923.114396	-25.9176	0.0051
2006	6681	6546.117751	-356.4215	0.0202
2007	6088	6113.917606	-103.4385	0.0043
2008	4474	4830.421534	-113.3927	0.0797
2009	4420	4523.438533	-205.0118	0.0234
2010	4703	4816.392691	-483.1632	0.0241
2011	4503	4708.01179	-390.7017	0.0455
2012	3615	4098.16323	-371.7971	0.1337
2013	3359	3749.701656	-626.0766	0.1163
2014	3362	3733.797139	-431.0495	0.1106
2015	2619	3245.076559	35.016	0.2391
2016	2781	3212.049454	89.2503	0.155

TABLE VI. ARIMA (2, 2, 2) AND SVM, BP COMBINATION MODEL.

Date	Actual injured (person)	ARIMA (2,2,2) and SVM, BP combined model		
		Predict	Errors	Fractional error
1980	7939	7429.434567	509.5654328	0.0642
1981	7287	7335.043423	-48.04342321	0.0066
1982	6813	6789.12539	23.87460992	0.0035
1983	6837	6823.697773	13.30222669	0.0019
1984	6670	6754.261679	-84.26167903	0.0126
1985	4917	5363.24927	-446.24927	0.0908
1986	5820	5562.169288	257.8307122	0.0443
1987	4579	5066.644985	-487.6449848	0.1065
1988	4136	4324.957688	-188.9576879	0.0457
1989	4110	4384.515296	-274.5152962	0.0668
1990	4315	4487.405326	-172.405326	0.0400
1991	4724	4844.000661	-120.0006607	0.0254
1992	3015	3875.79772	-860.7977201	0.2855
1993	2878	3323.748801	-445.7488007	0.1549
1994	3645	3881.035567	-236.0355667	0.0648
1995	3834	4135.243908	-301.2439077	0.0786
1996	4237	4464.77955	-227.7795498	0.0538
1997	5674	5603.246194	70.7538061	0 .0125
1998	8468	7887.431786	580.5682144	0.0686
1999	10607	9899.025806	707.9741941	0.0667
2000	10583	10302.04554	280.9544561	0.0265
2001	10424	10084.7021	339.2978997	0.0325
2002	10456	9984.817205	471.1827954	0.0451
2003	9877	9504.461833	372.5381671	0.0377
2004	8284	8237.38701	10.6129895	0.0013
2005	6888	6902.553075	-14.55307518	0.0021
2006	6681	6480.413638	200.5863616	0.0300
2007	6088	6066.123296	21.87670385	0.0036
2008	4474	4844.551964	-370.5519643	0.0828
2009	4420	4495.886057	-75.88605734	0.0172
2010	4703	4789.285893	-86.2858926	0.0183
2011	4503	4707.658826	-204.6588262	0.0454
2012	3615	4135.001283	-520.0012825	0.1438
2013	3359	3788.098942	-429.098942	0.1277
2014	3362	3760.950717	-398.9507168	0.1187
2015	2619	3300.856465	-681.8564647	0.2603
2016	2781	3232.086782	-451.0867817	0.1622

In this paper, first, we established the ARIMA, BP neural network. Through comparing, find SVM is better than BP. Then we established two combination model using the reciprocal variance method. One is an ARIMA (2, 2, 2) and SVM model, another is an ARIMA (2, 2, 2) and SVM, BP neural network model. The results show that we can improve the accuracy of predicting traffic events time series through combination model generally. The ARIMA (2, 2, 2) and SVM, BP neural network combination model is more accurate than each of single model, also than ARIMA (2, 2, 2) and SVM combination model. We can adopt ARIMA and SVM, BP neural network to predict traffic events index accurately.

In the futures, we will combine more than three models to improve the accuracy of traffic events index prediction.

REFERENCES

[1] I. Guyon and D. G. Stork, "Linear discriminant and support vector classifiers," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, and B. Scholkopf, Eds., Cambridge, MA: MIT Press, 2000.

[2] C. Burge, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[3] Q. Zhao, "Principe Support vector machines for SAR automatic target recognition," *IEEE Transom Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 634 654, 2001.

[4] K. I. Kim and K. Jung, "Support vector machines for texture classification," *IEEE Transom Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1542-1550, 2002.

[5] Ei-NaqaI and Y. Y. Yang, "A support vector machine approach for detection of microcalcifications," *IEEE Transom Medical Image*, vol. 21, no. 12, pp. 1552 1563, 2002.

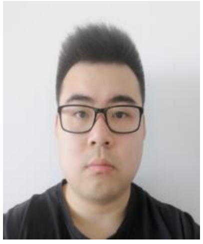
[6] G. M. Fung and O. L. Mangasarian, "Breast tumor susceptibility to chemotherapy via support vector machines," Technical Report 03 06, Data Mining Institute, 2003.

[7] J. M. Bates and C. W. J. Granger, "The combination of forecasts," in *Essays in Econometrics*, Cambridge University Press, 2001, pp. 451-468.

[8] W. Shuhua and G. Donglian, *et al*., *Journal of Yanshan University*, vol. 36, no. 1, pp. 79-83, 2012.

[9] W. Yan. Application time series analysis - Second R. J. Vidmar. (August 1992). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21(3). pp. 876-880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

Xiaorui Shao is an M.S. candidate in the Information collaboration Engineering department, Pukyong national university. His current research interests include big data, smart factory, machine learning.



Yifei Luo is a Master of Predictive Analytics student in School of Electrical Engineering, Computing and Mathematical Sciences (EECMS), Curtin University, Perth, WA, Australia.

Lester L Boey is a Master of Predictive Analytics student in School of Electrical Engineering, Computing and Mathematical Sciences (EECMS), Curtin University, Perth, WA, Australia