Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on the Prediction of Metropolitan Traffic Density in China using ARIMA Model

Shuai Hu[1] and Zhongluo Yu[2, *]

[1]School of Physics and Astronomy, University of Sheffield, Sheffield, S10 2TN, South Yorkshire

[2]Civil and Transportation Engineering, Southeast University Chengxian College, Nanjing, 210088, China

*Corresponding author: 220522135@cxxy.seu.edu.cn

Abstract. This investigation targets to research on construction of prediction model of traffic density and discuss suitable methods to decrease heavy traffic. The data is obtained from the public website of Government of Shenzhen which provides updates of traffic monitoring, time sequence analysis is put in practical manipulation, and it is verified that traffic undergoes regular alternations responsible for specific time periods every day. Data sets of traffic density of routes are analysed using ACF and PACF testing procedures and ARIMA, to obtain an output of the prediction model of tendency of traffic congestion of related area. The prediction model does not have apparent invalidity in statistical perspectives, and illustrate approximate results about probable time portions that produces proliferating heavy-traffic phenomena. Relevant analysis does encounter limitation that the model may become unpredictable to sudden situations like temporary route construction, and unpleasant weather conditions which may result in chaotic circumstances to fluctuation of traffic for citizens.

Keywords: Metropolitan traffic; prediction; ARIMA model.

1. Introduction

Contemporary urban transportation holds high demand of systematic control due to proliferation of traffic density. Traffic jams results several consequences, such as detrimental effects on environments and severely lowering efficiencies of activities of citizens. Consequently such effect would cause unnecessary inconveniences [1]. Sequence analysis is a common methodology of assessing traffic fluctuations especially at busy time periods. The investigation aims to excavate on understanding the main factors resulting heavy traffic and predicting their occurrences.

Considering specific characteristics of urban planning of the main cities in China, there are several noticeable differences in comparison of Western countries such as the United Kingdom and the United States: private cars occupy a lower percentage in the entire composition of transportation [2]. This is an inevitable tendency for development of modern Chinese transportation system because of challenges of population and requirements of environmentally friendly situation of living. Consequently, making predictions of urban traffic in China should involve in multidimensional considerations such as the positions and fluctuations of public transportation, such as subway, railway and share bikes [3, 4]. Furthermore, traffic density varies vigorously throughout special time periods, such as public holidays and alternatively change between weekdays and weekends as well. On weekdays traffic undergoes most pressure in the morning and before sunset, which respectively represents when citizens begin and leave work of the day [5]. All perspectives are crucial factors that should be considered in modelling and data analysis.

As aforementioned, researches on control of traffic congestion of cities in China is more perplexing than other modern global cities [6]. Notwithstanding the challenging factors and multidisciplinary difficulties, Chinese metropolitan systems obtain apparent strengths in comparison to western of European countries. For instance, China had made amazing progress for urban design over the past decade, which enables multiple selections that satisfy conveniences for citizens [7]. Especially high-efficiency development and constructions of public transportation such as subway and shared bikes,

which not only promotes environmentally friendly targets for smart city transportation, but loosening the pressure for each route on the transportation system [8, 9]. Appropriate urban design also enables sufficient monitoring in a more convenient way, lowering potential risks of accidents and increasing the probability of prediction of subsequent situation of flow of traffic [10].

2. Methods

2.1. Data Source

Time sequence analysis is the main methodology of this investigation. This is achieved by collecting historical traffic density datasets from official metropolitan transportation management websites. Required fields of assessment include daily flow of and average velocity of vehicles of a certain street line, statistics of weather conditions and information of weekdays and holidays with other probable factors that influence traffic environments [3]. Acquired data is filtered and outliers are eliminated to ensure data is complete and accurate. Along with preparations, the distance of traffic congestion of street lines in Guang Ming, Shenzhen, is chosen as the datasets of this investigation. The data set is an ideal selection because it is suitable for accessing chronological sequential alternatives during data collection. Therefore, it is convenient to modulate parameters of model output and test expected efficiency for model assessment. Furthermore, there would be external characteristics for targeted completion of the model. Combining all compulsory conditions of required field may eventually result in an optimized precision for the output model [4].

2.2. Method Introduction

The target of the investigation is to obtain a suitable time sequence prediction model for the selected data sets. It is crucial that the model is statistically acceptable for further prediction of any traffic circumstances. Therefore, for the TIMESLICE parameter given by the data set, which indicate historical in-time situations of traffic congestion over the range of the routes that the data set had been monitored, a timing diagram is firstly created to show overall tendency of TIMESLICE. Hence the samples of the data sets undergoes statistical testing procedures for to be ensured that predictions are valid. This is achieved by listing a ACF and PACF diagram and tables of lagging order of samples. The prediction model is eventually being completed by ARIMA together with model tables.

3. Results and Discussion

3.1. Original Time Series

The procedure of data analysis mainly focuses on chronological tendencies of transportation of a given street line of the chosen data set. This is achieved by constructing a timing diagram. The diagram effectively displays time dependence of relevant data (Figure 1).

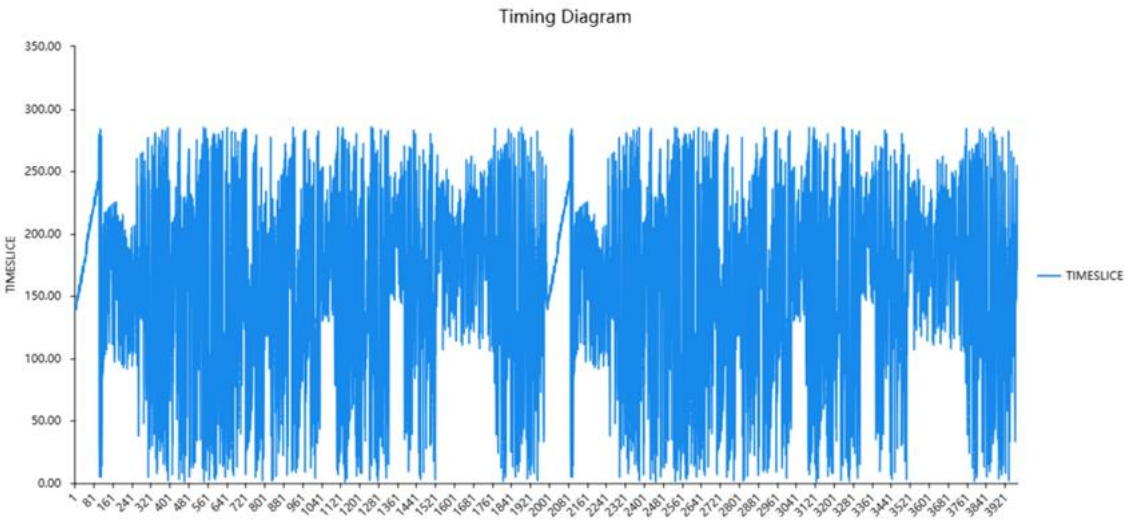

Fig. 1 Timing Diagram

The timing diagram that represents variation of the TIMESLICE parameter of the data set. For the horizontal axis of the diagram, if time is added the label of the axis would be time, if not, it would be shown as sequence numbers. If multiple sequence parameters are added, the diagram would display them separately by default.

3.2. ACF and PACF Test

Experimentally the modelling procedure is first accomplished by manipulating the testing datasets, which is designed to calculate the difference between predicted and realistic results (Table 1). Such uncertainty contains Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Focusing on TIMESLICE, the time sequence data has obtained a testing result of a statistical value of -9.065.

Table 1. ADF Testing Table

Differential Order	t	p	Critical Value		
			1%	5%	10%
0	-9.065	0.000	-3.432	-2.862	-2.567

Except for the statistical t value, the p value equals to 0.000(<0.01), and the 1%, 5%,10% critical values are respectively -3.432, -2.862, -2.567, which means that there is a probability of larger than 99% to refuse the original assumption, proving stability of orders simultaneously (Table 1). Further information of other relevant factors of TIMESLICE, including lags values that have been used by internal algorithm of data analysis, and information criterion (AIC and BIC). AIC and BIC values are used for comparison of qualities of two testing (Table 2).

Table 2. TIMESLICE-Other Paramaters

Differential Order	Number of samples	Number of being built based on lags	aic	bic
0	3951	20	45282.916	45421.114

To acquire prediction ARIMA transportation model is used for further analysis. Hence, ACF and PACF diagrams are applied to judge auto regressive order (p) and moving average (q):

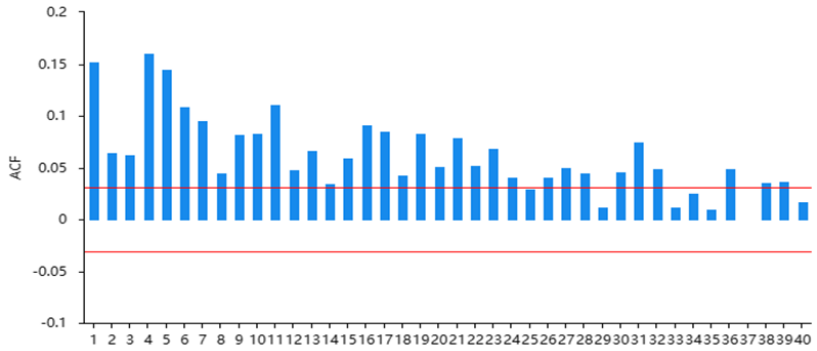

Fig. 2 ACF Diagram

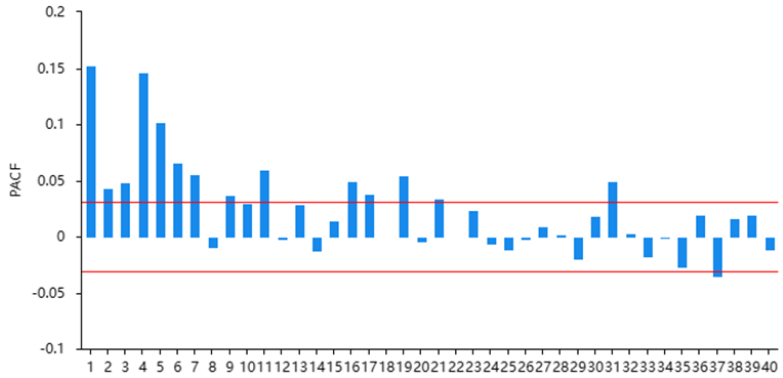

Fig. 3 PACF Diagram

Figure 2 and 3 display ACF and PACF diagrams that are used for judgement of orders. The horizontal axes represent lagging order. There are three principles to create appropriate predictions: (1) If ACF = 0 under a certain lagging order at q, and the related PACF diagram does not, the ARIMA model can be simplified to a MA (q) model. (2) If PACF = 0 under a certain lagging order at p, and the related ACF diagram does not, the ARIMA model can be simplified to a AR (p) model. (3) If both diagrams cannot satisfy (1) and (2), it is necessary to select suitable ARIMA orders. It is more available to select the most apparent number of orders at ACF diagram to create a p value and the same number at PACF diagram for a q value. Based on the tendencies of Figure 2 and 3 the results of ACF and PACF diagram of ARIMA model is shown by the table below.

3.3. ARIMA Prediction Model

From the table above it is able to illustrate that the model is verified after the ACF test. Hence it is available to create the prediction model including statistical errors. The prediction model is based on the model table 3 below:

Table 3. ARMA(1,4) model results

Term	Symbol	Coefficient	SD	z	p	95% CI
Constant	c	19.320	3.253	5.939	0.000	12.944 ~ 25.696
AR	α1	0.874	0.021	41.003	0.000	0.832 ~ 0.916
MA	β1	-0.761	0.024	-32.229	0.000	-0.808 ~ -0.715
	β2	-0.071	0.016	-4.331	0.000	-0.103 ~ -0.039
	β3	-0.010	0.017	-0.607	0.544	-0.044 ~ 0.023
	β4	0.096	0.014	6.710	0.000	0.068 ~ 0.124

The AIC value is 45536.024 while the BIC value is 45580.033. It displays the situation where the model is built and the aforementioned information criteria are used for comparisons of analyzed models. If model analysis is commenced multiple times these values would be used for contrast of variations to demonstrate improvements of model construction in general. According to Table 4, the model is obtained by the formula below:

$$y(t) = 19.320 + 0.874 * y(t-1) - 0.761 * \varepsilon(t-1) - 0.071 * \varepsilon(t-2) - 0.010 * \varepsilon(t-3) + 0.096 * \varepsilon(t-4) \quad (1)$$

The prediction model statistics is displayed by the table 4 below. The table shows information of Statistics of Model Q (Ljung-Box Statistics Testing), the ARIMA model requires white noise residuals (residuals have no existence of auto correlation). This is achieved by white noise testing by statistics of Q.

Table 4. Table of Model Q Statistics

Term	Statistics	p
Q1	0.047	0.829
Q2	0.131	0.936
Q3	0.132	0.988
Q4	1.157	0.885
Q5	1.801	0.876
Q6	2.179	0.903

From the results of statistical analysis of model Q, Q6 has p value larger than 0.1, which means that the null hypothesis cannot be rejected at the significance level of 0.1, and the residual of the model is white noise, and the model basically meets the requirements. Eventually the TIMESLICE model prediction (Figure 4) interprets the entire data analysis. The ARIMA model shows the data fitting and the data prediction for the next 12 periods, only the actual value and fitted value of the last 1000 periods of the original time series are displayed in the figure (if the original time series is less than 1000 periods, the actual series data is displayed).

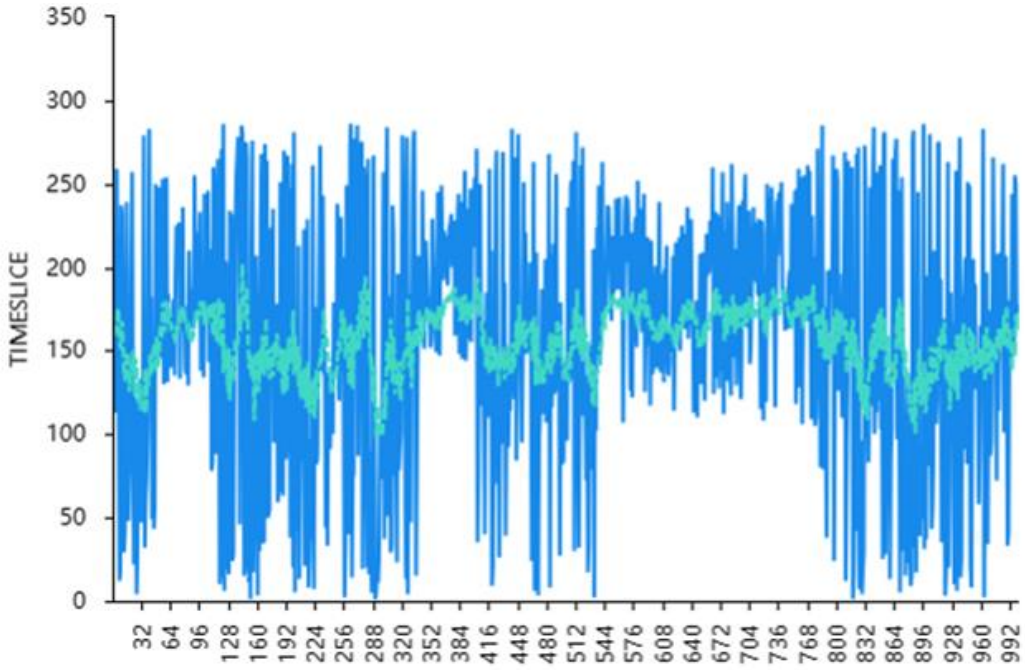

Fig. 4 Model fitting and prediction

Figure 4 shows model fitting and prediction. The blue line represents realistic values; the dotted green line represents fitted value, the green line represents prediction value. The dotted yellow line represents the 95% upper limit of prediction value (may not be displayed apparently on the figure) and the red line represents the 95% lower limit of the prediction value (may not be displayed apparently on the figure.

4. Conclusion

Modelling gives a result of statistical demonstration of traffic congestion at different portion of time over the range covered throughout the data set. It is unprecedented that there would be inevitable heavy traffic at peak periods when citizens begin and leave work of the day, which is also an unproblematic phenomenon under high population density. During weekends and holidays fluctuation of population would diminish to a more ideal amount. It is also noticeable that temporary events cause some influences on traffic density as well, such as road construction. Apparently, traffic density is a parameter that is hypersensitive to any changes to the environment. The purpose of making prediction models is to provide on-time information to take relevant strategies to minimize continuing congestion.

There are also some limitations for the traffic density model. For instance, it is unavailable to precisely predict the influences of weather and seasonal changes to daily traffic density, as the dataset does not give clear guidance to the weather condition of given dates. Moreover, realistic conditions would be much more chaotic than ideal predicting results. There are more multidimensional factors that can deviate the tendency of traffic density from expected, such as moving population between different cities, which occurs most during long festivals like the National Day and the Spring Festival in China. The model is based on historical data and is susceptible to be influenced by uncertain factors in the future. Consequently, the model can be improved and produce more enlightenment to smart city traffic control.

5. Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

6. References

[1] Jin, Zijuan. Research on Probability Density Estimation and Prediction Methods of Road Traffic Flow Density. Hunan University, 2019.

[2] Qian, Kun, Zhang, Jian, Lou, Huan, et al. Traffic Flow Density Prediction Based on Fuzzy Information Granulation and Support Vector Machine Combination Model. The 6th Yunnan Provincial Association for Science and Technology Academic Annual Meeting and Honghe River Basin Development Forum Proceedings, Hefei University of Technology, 2016.

[3] Yang, Senyan. Research on Urban Road Traffic Situation Analysis and Application Based on Spatiotemporal Data Mining. Tsinghua University, 2019.

[4] Zhang Fuqiang. Research on Traffic Congestion Prediction Methods of Urban Main Roads. Chang'an University, 2015.

[5] Wang Teng. Traffic Congestion Event Detection Based on Time Series. Tianjin University, 2012.

[6] Zhang Y, Van der Schaar M. Traffic Flow Prediction Using Deep Learning: A Survey. Computer Networks, 2018, 146: 248-260.

[7] Li X, et al. Deep Learning Based Traffic Flow Prediction Considering Spatio-Temporal Correlations. Transportation Research Part C: Emerging Technologies, 2019, 105: 548-564.

[8] Wu Y, et al. A Hybrid Model for Predicting Traffic Flow Based on Deep Learning and Time Series Analysis. Journal of Advanced Transportation, 2020.

[9] Chen Min, Gao Na, Wang Hao, et al. Research on Urban Traffic Congestion Prediction Model Based on Deep Learning. Traffic Standardization, 2019, 8: 121-124.

[10] Chen Tao, Zhang Yan, Xiao Tong, et al. Research on Urban Road Traffic Flow Prediction Method Based on ARIMA Model. Highway Traffic Science and Technology, 2017, 34(8): 88-93.