Journal of Southeast University (English Edition) Vol. 29, No. 4, pp. 445 - 449 Dec. 2013 ISSN 1003—7985

Effect of aggregation interval on vehicular traffic flow heteroscedasticity

Shi Guogang[1] Xiang Qiaojun[1] Guo Jianhua[2] Zhang Hongxin[2]

([1] School of Transportation, Southeast University, Nanjing 210096, China)

([2] Intelligent Transportation System Research Center, Southeast University, Nanjing 210096, China)

Abstract: The effect of the aggregation interval on vehicular traffic flow heteroscedasticity is investigated using real-world traffic flow data collected from the motorway system in the United Kingdom. 30 traffic flow series are generated using 30 aggregation intervals ranging from 1 to 30 min at 1 min increment, and autoregressive integrated moving average (ARIMA) models are constructed and applied in these series, generating 30 residual series. Through applying the portmanteau Q-test and the Lagrange multiplier (LM) test in the residual series from the ARIMA models, the heteroscedasticity in traffic flow series is investigated. Empirical results show that traffic flow is heteroscedastic across these selected aggregation intervals, and longer aggregation intervals tend to cancel out the noise in the traffic flow data and hence reduce the heteroscedasticity in traffic flow series. The above findings can be utilized in the development of reliable and robust traffic management and control systems.

Key words: heteroscedasticity; traffic flow; autoregressive integrated moving average (ARIMA); residual

doi: 10.3969/j.issn.1003-7985.2013.04.017

With the rapid economic and social development, the number of vehicles is increasing dramatically in China, causing serious congestion and safety issues for the transportation systems, in particular, for big cities. For alleviating the negative impacts associated with traffic congestion and safety issues, the conventional approach of building more roads is limited due to land use constraints. The applications of advanced technologies, such as computer technology, communication technology, database technology, statistical data mining technology, etc., into the conventional transportation systems have been receiving increasing attention from both research and industry sectors, stimulating the rapid development of the intelligent transportation systems (ITS).

Many applications have been developed under the broad umbrella of ITS and reliability-related applications can provide more robust solutions for battling the congestion and safety issues, where the uncertainty information concerning transportation systems is modeled and utilized to impart the reliable treatment into traffic management and control. In this direction, two major approaches are used to model the second-order moment of transportation information, i.e., the generalized autoregressive conditional heteroscedasticity (GARCH) model [1-4] and the stochastic volatility model[5]. In addition, Guo et al.[6] established the heteroscedastic nature of traffic information, providing a foundation for performing the abovementioned uncertainty modeling analysis.

Note that all these studies are conducted for a single time interval; however, as shown in Refs. [7-9], the aggregation interval is an essential component for transportation system applications. Therefore, in this paper, we investigate the effect of aggregation intervals on the traffic condition heteroscedasticity.

1 Theoretical Background

1.1 Aggregation interval

The aggregation interval is a critical factor for characterizing or defining traffic condition data. For example, in *Highway Capacity Manual* 2000, a 15 min aggregation interval is usually used to compute the volume of traffic, i.e., the number of vehicles passing a road section within a single 15 min aggregation interval [10]. Similarly, other traffic characteristics such as occupancy, speed, etc. can also be defined over a certain aggregation interval. Intuitively, aggregation intervals will affect the characteristics of traffic variables series. As shown in Ref. [8], longer aggregation intervals will help to cancel out the noise and hence create a smoother traffic flow series. The effects of aggregation intervals have been investigated in other fields such as short term traffic flow forecasting and single loop speed estimation. In this paper, the effect of aggregation intervals on the traffic flow heteroscedas-

Received 2013-08-06.

Biographies: Shi Guogang (1975—), male, graduate; Xiang Qiaojun (corresponding author), male, doctor, professor, xqj@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 71101025), the National Key Technology R&D Program of China during the 12th Five-Year Plan Period (No. 2011BAK21B01), the Doctoral Programs Foundation of the Ministry of Education of China (No. 20100092110037), the Fundamental Research Funds for the Central Universities.

Citation: Shi Guogang, Xiang Qiaojun, Guo Jianhua, et al. Effect of aggregation interval on vehicular traffic flow heteroscedasticity[J]. Journal of Southeast University (English Edition), 2013, 29(4): 445 - 449. [doi: 10.3969/j.issn.1003-7985.2013.04.017]

ticity is investigated.

1.2 ARIMA modeling

For a selected aggregation interval, the collection of traffic flow data formulates a traffic flow time series, and for this data, the conventional autoregressive integrated moving average (ARIMA) model has been proven to be an effective modeling tool[11]. Given a traffic flow time series process denoted by $\{X_t\}$, the ARIMA(p, d, q) model is defined as

$$\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t$$

where t is the time index; p is the order of the short-term autoregressive polynomial; q is the order of the short-term moving average (MA) polynomial; d is the order of short-term differencing; B is the backshift operator such that $BX_t = X_{t-1}$; ε_t is the random error at time t; $\phi(B)$ is the short-term AR polynomial defined as $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$; $\theta(B)$ is the short-term MA polynomial defined as $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$.

In the above definition, the roots of $\phi(B)$ and $\theta(B)$ are assumed to be outside of the unit circle and have no common factors. ε_t is the residual series, and for traffic flow series, it has been proven to be heteroscedastic; i.e., it has zero mean and time-varying conditional variance [6].

The ARIMA model can be processed using the conventional Box-Cox framework, which includes three major steps, i.e., model identification, model estimation, and model diagnostic check[11]. In the model identification step, the orders of the model will be selected based on the minimum information criterion. In the model estimation step, the model parameters will be estimated using primarily the maximum likelihood estimation approach. In the model diagnostic check step, the estimated model will be tested to make sure that the residuals after applying the estimated model on the data containing no autocorrelation structure. It is worthwhile to mention that the above three steps can be conducted iteratively so that the final model will meet the requirements. Since these three steps are generally complex for carrying out manually, commercial software packages have been developed to facilitate ARIMA modeling, e.g., SAS PROC ARIMA[12].

1.3 Traffic heteroscedasticity

In conventional statistical analysis models such as analysis of variance, regression, etc., the data are generally assumed to be having constant variance, or the data are called homoscedastic. However, in some cases, e.g., for the traffic flow series, the process variance is time-variant; i.e., traffic flow series is called heteroscedastic. Using traffic flow series collected in the United Kingdom and aggregated at a 15 min aggregation interval, the heteroscedastic phenomenon can be demonstrated in Fig. 1.

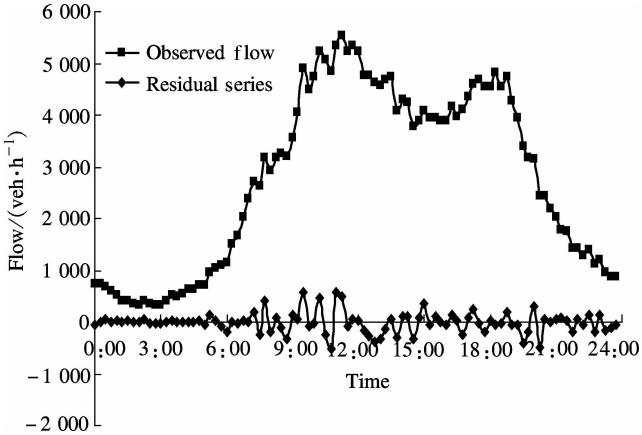

Fig. 1 Traffic flow heteroscedasticity demonstration

From Fig. 1, we can see that the residual series is scattered around the x-axis while the ranges of the scattering are different for different times of the day; i.e., the residuals scatter more widely for high level traffic in the daytime. This information indicates that certain measures should be taken to handle the added uncertainty for peak hour traffic when developing advanced traffic management and control systems. As mentioned previously, the heteroscedasticity is important for developing ITS applications, and Guo et al. [6] showed that this heteroscedasticity is universal across many sites for a single 15 min aggregation interval. In this paper, we will show that this heteroscedasticity is also significant across different aggregation intervals.

1.4 Heteroscedasticity test

The heteroscedasticity test will be performed on the residual series after the autocorrelation structure is removed from the traffic flow series. Based on the residual series, the tests used in this paper include the portmanteau Q-test and the Lagrange multiplier (LM) test, which have the ability of testing the presence of nonlinear effects (such as GARCH effects) in the residuals. Note that the Q-test and the LM test are among many tests that can be selected for performing the heteroscedasticity test, and these two tests are selected in this paper due to their ready implementation and application through commercial software, i.e., SAS PROC AUTOREG. For ARIMA modeling, many off-the-shelf commercial software packages have been developed and in this paper, this SAS PROC AUTOREG is applied to test the heteroscedasticity in the residual series.

2 Data Description

2.1 Data collection

The traffic flow data used in this paper was collected

from the MIDAS system installed for the motorway of M25 around London, the United Kingdom. M25 was initially built for servicing the through traffic across London, while over the years, M25 gradually merged into the urban transportation system, servicing many local trips and causing serious congestion issues on M25. Under this circumstance, the MIDAS system was developed for battling this situation. Using the MIDAS, traffic data including traffic volume, speed, occupancy, etc. was continuously registered over traffic detectors installed along the motorway. In this paper, traffic flow data collected from the station 4762a is used. The time range of the data is from Jan 1, 2002, to Dec 31, 2002; i. e., the whole year traffic flow data is used in this paper.

2.2 Data aggregation

The purpose of this paper is to investigate the effect of aggregation intervals on the traffic flow heteroscedasticity. Therefore, an important step is to aggregate the traffic flow data over multiple aggregation intervals. In this paper, altogether 30 aggregation intervals are used, i. e., aggregation intervals starting from 1 to 30 min with 1 min increment. Note that the aggregation of traffic flow data follows the rule proposed by Edie[13] and the aggregation operation is carried out using SAS PROC EXPAND, through which the aggregation rule can be easily implemented. The traffic flow data aggregated at 1 min interval are shown in Fig. 2. As can be seen that the seasonal pattern can be identified for this 1 min traffic flow data series, which is an important phenomenon that should be handled. In addition, this seasonal pattern also exits in traffic flow series aggregated over other aggregation intervals.

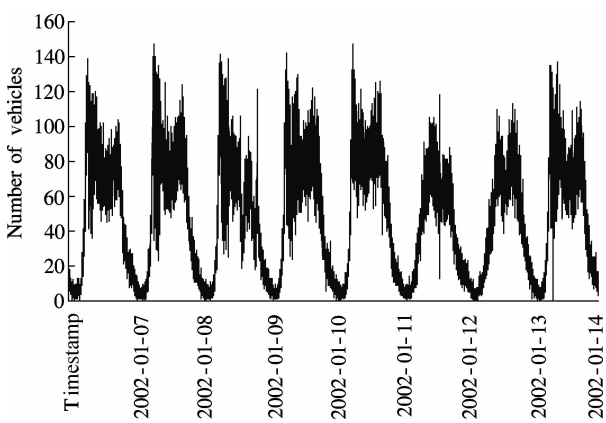

Fig. 2 Traffic flow series demonstration (partial data aggregated at 1 min interval)

3 Empirical Results

In this section, the empirical results will be shown, including the ARIMA modeling results and the results of the heteroscedasticity test over multiple traffic flow series aggregated at different aggregation intervals.

3.1 ARIMA modeling results

The purpose of ARIMA modeling is to capture and remove the first-order moment of the traffic flow series and hence generate the residuals for the heteroscedasticity test. Note that we have 30 traffic flow series corresponding to 30 aggregation intervals; therefore, we will have 30 identified and estimated ARIMA models.

As mentioned previously, the ARIMA model can be processed through the steps of identification, estimation, and diagnostic check. In this paper, in the model identification step, the seasonal pattern is first handled by seasonal differencing. As identified in Ref. [14], a weekly pattern can be used. Taking the 15 min aggregation interval as an example, a weekly pattern will have 24 (h/d) $\times$4(data point/h) $\times$7(d) = 672 data points. Therefore, for removing the seasonal effects, two traffic flow data points at a distance of 672 aggregation intervals are differenced; i. e., the seasonal differencing order is 672. After seasonal differencing, the differenced series are used to identify the orders of ARIMA.

Using PROC ARIMA, the orders of the ARIMA model for all the 30 traffic flow series are shown in Tab. 1.

Tab. 1 ARIMA modeling and heteroscedasticity test results

Aggregation interval/min	ARIMA model			Monthly test result/%
	p	d	q	
1	3	0	2	100
2	4	0	4	100
3	3	1	2	100
4	2	0	3	100
5	2	1	3	100
6	2	1	3	100
7	2	1	3	100
8	2	1	3	100
9	1	1	4	100
10	1	1	4	100
11	1	1	4	100
12	1	1	4	100
13	2	0	3	100
14	2	0	3	100
15	2	0	3	100
16	2	0	3	100
17	2	0	3	100
18	2	0	3	100
19	2	0	3	100
20	2	0	3	100
21	2	0	3	100
22	2	1	3	100
23	2	1	3	100
24	2	1	3	100
25	2	1	3	100
26	2	1	3	100
27	2	1	3	100
28	2	1	3	91.67
29	2	1	3	91.67
30	2	1	3	91.67

After selecting these orders, the ARIMA models are estimated using the maximum likelihood method and the residuals are computed using the estimated parameters. Then, in the final model diagnostic check step, the selected models and estimated parameters are validated by checking the characteristics of the residual series. According to the ARIMA modeling theory, the residual series should be white noise or the autocorrelations in the residuals are trivial, indicating that the autocorrelation structure has been adequately removed. In this paper, the autocorrelations in all the 30 residual series are trivial, indicating the adequacy of selecting the identified ARIMA models.

3.2 Heteroscedasticity test results

Based on the obtained residual series, the heteroscedasticity test results are presented as follows. In the whole series test, for all the aggregation intervals, the two heteroscedasticity tests are applied on the entire residual series with the *p*-values of the two test statistics less than 0.000 1, showing that for all the 30 residual series the traffic flow series are heteroscedastic.

For the monthly test, the results are also shown in Tab. 1. For each aggregation interval, the entire series is first broken into monthly series, and each monthly series is tested separately. Then the percentage of the heteroscedastic monthly residual series is computed as the test results. We can see that except for the aggregation intervals of 28, 29, and 30 min that have 91.67% heteroscedastic monthly residual series, and the remaining aggregation intervals have 100% heteroscedastic monthly residual series, indicating that traffic flow series are heteroscedastic at the monthly level for these aggregation intervals. The test results indicate that the longer aggregation interval can cancel out the noise in traffic flow data and hence reduce the heteroscedasticity in traffic flow.

4 Conclusions

Considering the importance of reliability in many ITS applications, the uncertainty analysis has received increasing attention from transportation research communities. In this direction, the traffic flow heteroscedasticity test and modeling play an essential role. Previous studies have shown that traffic flow is heteroscedastic across stations for a certain data aggregation interval, and in this paper we investigate the effects of multiple aggregation intervals on traffic heteroscedasticity. Using real-world traffic flow data, the following two conclusions can be drawn as follows:

1) Traffic flow is heteroscedastic across multiple intervals ranging from 1 to 30 min at 1 min increment.

2) Longer aggregation intervals can cancel out the noise in the traffic flow data and hence reduce the heteroscedasticity in traffic flow series.

Based on the above conclusions together with the previous investigations, heteroscedasticity can be claimed to be universal for traffic flow data both temporally and spatially, and considerations should be taken to improve the reliability or robustness of ITS-related traffic management and control applications.

References

[1] Karlaftis M G, Vlahogianni E I. Memory properties and fractional integration in transportation time series [J]. *Transportation Research Part C*, 2009, **17**(4): 444 − 453.

[2] Guo J H, Williams B M. Real time short term traffic speed level forecasting and uncertainty quantification using layered Kalman filters [J]. *Transportation Research Record*, 2010, **2175**: 28 − 37.

[3] Sohn K, Kim D. Statistical model for forecasting link travel time variability [J]. *ASCE Journal of Transportation Engineering*, 2009, **135**(7): 440 − 453.

[4] Yang M L, Liu Y G, You Z S. The reliability of travel time forecasting [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2010, **11**(1): 162 − 171.

[5] Tsekeris T, Stathopoulos A. Short-term prediction of urban traffic variability: stochastic volatility modeling approach [J]. *ASCE Journal of Transportation Engineering*, 2010, **136**(7): 606 − 613.

[6] Guo J H, Huang W, Williams B M. Integrated heteroscedasticity test for vehicular traffic condition series [J]. *ASCE Journal of Transportation Engineering*, 2012, **138**(9): 1161 − 1170.

[7] Guo J H, Williams B M, Smith B L. Data collection time interval for stochastic short-term traffic flow forecasting [J]. *Transportation Research Record*, 2008, **2024**: 18 − 26.

[8] Smith B L, Ulmer J M. Freeway traffic flow rate measurement: investigation into impact of measurement time interval [J]. *ASCE Journal of Transportation Engineering*, 2003, **129**(3): 223 − 229.

[9] Guo J H, Huang W, Wei Y, et al. Effect of time interval on speed estimation using single loop detectors [J]. *KSCE Journal of Civil Engineering*, 2013, **17**(5): 1130 − 1138.

[10] TRB. Highway capacity manual 2000 [R]. Washington: Transportation Research Board, 2000.

[11] Box G E P, Jenkins G M, Reinsel G C. *Time series analysis*: *forecasting and control* [M]. 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.

[12] SAS Institute, Inc. *SAS OnlineDoc version* 8 [M]. Cary, NC, USA: SAS Institute, 2000.

[13] Edie L C. Discussion of traffic stream measurements and definitions [C]//*Proc of* 2*nd International Symposium on the Theory of Traffic*. Paris, France, 1963: 139 − 154.

[14] Williams B M, Hoel L A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results [J]. *ASCE Journal of Transportation Engineering*, 2003, **129**(6): 664 − 672.

时间汇集间隔对交通流异方差性的影响

史国刚[1] 项乔君[1] 郭建华[2] 张宏新[2]

([1] 东南大学交通学院,南京 210096)
([2] 东南大学智能运输系统研究中心,南京 210096)

摘要:依托从英国快速路系统中采集到的实际交通流数据,研究了时间汇集间隔对交通流异方差性的影响. 使用 1 ~ 30 min 共 30 种时间汇集间隔,生成了 30 个实际交通流数据序列,确定并估计了相应的 ARIMA 模型,计算后得到 30 个交通流量残差序列. 针对不同汇集间隔下的 ARIMA 模型残差序列,应用 portmanteau Q 检验和 LM(Lagrange multiplier)检验,分析了交通流量序列的异方差性. 实证结果表明:交通流量序列在选定的 30 个汇集间隔都具有显著的异方差性;较长的时间汇集间隔可以消减交通流量序列中的噪声,从而减弱交通流异方差性的程度. 研究结果有助于开发具有较高可靠性和鲁棒性的交通管理和控制系统.

关键词:异方差性;交通流;ARIMA;残差

中图分类号:U491