Qingxuan Kong

Prof. Guangjian Zhang

Time Series Analysis 60143

May 4, 2019

Short-term prediction for traffic volume

The univariate time series model for short-term traffic volume forecasting has been proved to be efficient. This study applies the univariate ARMA model on the traffic volume from seven sensors on national roads in Dublin separately and then identifies the common orders of AR and MA terms for future use.

1. Introduction

Traffic congestion has been a serious problem in the metropolis for many years. An accurate prediction of the traffic volume in the future is required for both the traffic regulators and the road users to respond more actively to the potential congestion.

Several pure time series model for short-term traffic flow forecasting has been proposed at the very beginning. An autoregressive integrated moving average (ARIMA) (0,1,3) model has been proved to represent all the datasets from three surveillance systems in Los Angeles, Minneapolis and Detroit (MS Ahmed and AR Cook, 1979). The effectiveness of pure time series model has been proved and then improved by some more complicated time series models, like a dynamic prediction through Kalman filtering theory (Iwao Okutani and Yorgos J.Stephanedes, 1984), the statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of

traffic volume (Eleni I. Vlahogianni, Matthew G. Karlaftis and John C.Golias, 2006), and a Bayesian time series model (Bidisha Ghosh, Biswajit Basu, M.ASCE and Margaret O'Mahony, 2007).

Based on the previous discussion, the contributions of this paper are 2-fold: (1) a univariate ARMA model of the traffic volume at a specific spot; (2) a representative ARMA model to fit the traffic volume on different roads.

2. Data set description and empirical research questions

2.1 Dataset description

The ARMA model is built on the hourly data of the traffic volume from seven sensors in Dublin. The seven sensors' location is shown in Table 1 and only one direction is considered for each sensor for simplicity. The traffic volume is defined as the counts of traffic passing through the specific sensor. We first tried to build a model for both weekdays and weekends, but the model was not fully adequate because the residuals were not white noise and the model was not even causal and invertible. Thus, we only focus on the weekdays in this project. The dataset used in this project is collected every hour from January 2nd,2017 to September 1st, 2017, which contains 4,200 hourly data, 175 daily data and 35 weeks.

2.2 Empirical research questions

The time series data has to be decomposed into the trend, seasonal and residuals components first. The trend and seasonal components should illustrate the reasonable changes in the traffic volume per hour, per day or per week, while the ideal residuals

should be stationary, which allow building a causal and invertible ARMA to make accurate predictions. Thus, how clear the trend and seasonal components explain the traffic volume data and how strong the ARMA model is are two big questions need to be answered.

Table 1

Dataset description of the sensors' name, location and channel

Sensor	Location	Channel
Sensor1	M01 Between Jn01 M50/M01 and Jn02 Dublin Airport	Northbound
Sensor2	M02 Between Jn01 M50/N02 and Jn02, Coldwinters, Co Dublin	Northbound
Sensor3	N03 Between Jn02 Blanchardstown and Jn03 Clonsilla, Blanchardstown, Co. Dublin	Northbound
Sensor4	N04 Between Jn01 N4/M50 and Jn02 Liffey Valley, Liffey Valley, Co. Dublin	Eastbound
Sensor5	N07 Between Jn01 M50 and Jn1a Newlands Cross (R113), Dublin	Eastbound
Sensor6	N81 Between Tallaght Village and M50, Tallaght, Co. Dublin	Eastbound
Sensor7	M50 Between Jn11 Tallaght and Jn12 Firhouse, Co. Dublin	Northbound

3. Analysis plan and justifications

3.1 Analysis and justifications of the trend and seasonal components

The dataset of the traffic volume used in this project contains the hourly data. We can tell the hourly, daily or weekly trend and seasonal components. Since the dataset

contains 4200 hourly data, we can extract part of the continuous hourly data for one month (20 weekdays) to better illustrate the trend and seasonal components on the hourly and daily basis, and also extract the data of a specific hour from January 2nd,2017 to September 1st, 2017 to better illustrate the trend and seasonal components on the daily and weekly basis. We expect we can tell the busiest hours and days from the trend components, and the variation of the hours and days from the seasonal components

3.2 Analysis and justifications of the ARMA model

A separate ARMA model will be built for each data of the traffic volume from different sensors. The residuals of the decomposition will be used to make the model. Before the modeling, we need to test whether the residuals of the decomposition are normally distributed and stationary. A normally distributed data follows a straight diagonal line in the quantile-quantile (Q-Q) plot and a stationary process has the property that the mean, variance, and autocorrelation structure do not change over time. The residuals need to be stationary but not necessarily normally distributed so that we can fit them with an ARMA model later. When choosing an appropriate ARMA model, we will test possible ARMA models using AIC and select one with the smallest value of AIC. If that model is causal and invertible, and the residuals of the ARMA model is white noise, we will assume it is the most ideal ARMA model for the data of the traffic volume.

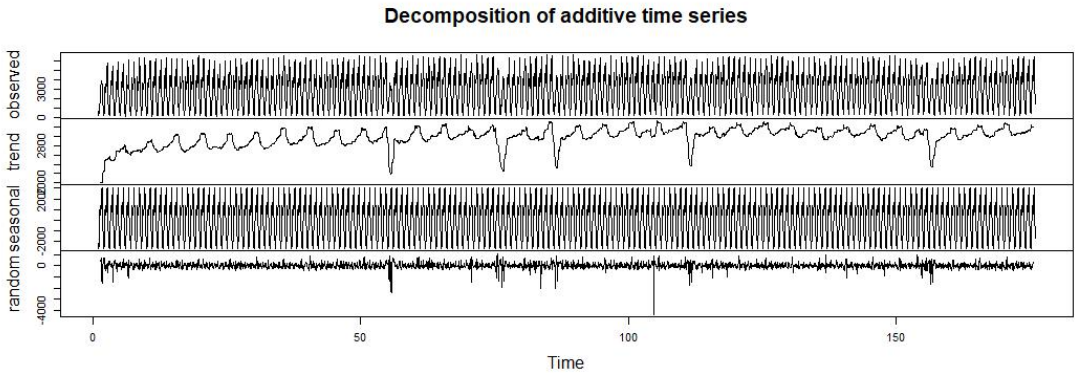
4. Summary statistics

We decompose the data of traffic volume and test the stationarity and normality of the residuals of the decomposition on the traffic volume data from seven sensors. We

apply an augmented Dickey-Fuller (ADF) test to test the stationarity, and the Q-Q plot and Shapiro-Wilk test to test the normality.

Figure 1

Decomposition of the data of the traffic volume from Sensor 1



4.1 Augmented Dickey-Fuller (ADF) test

An ADF test tests the null hypothesis that the time series can be represented by a unit root, which means it is not stationary. The p-values here as shown in Table 2 are all smaller than 0.05, so we reject the null hypothesis to say the residuals of the decomposition of the traffic volume data from seven sensors are stationary.

4.2 Q-Q plot and Shapiro-Wilk test

A Q-Q plot shows the distribution of the data against the expected normal distribution. For normally distributed data, observations should lie approximately on a straight line. If the data is non-normal, the points form a curve that deviates markedly from a straight line. Possible outliers are points at the ends of the line, distanced from the bulk of the observations. On the Q-Q plots shown in Figure 2, the residuals of the decomposition of the traffic volume data from seven sensors appear as roughly a straight line but the ends of the Q-Q plots start to deviate from the straight lines.

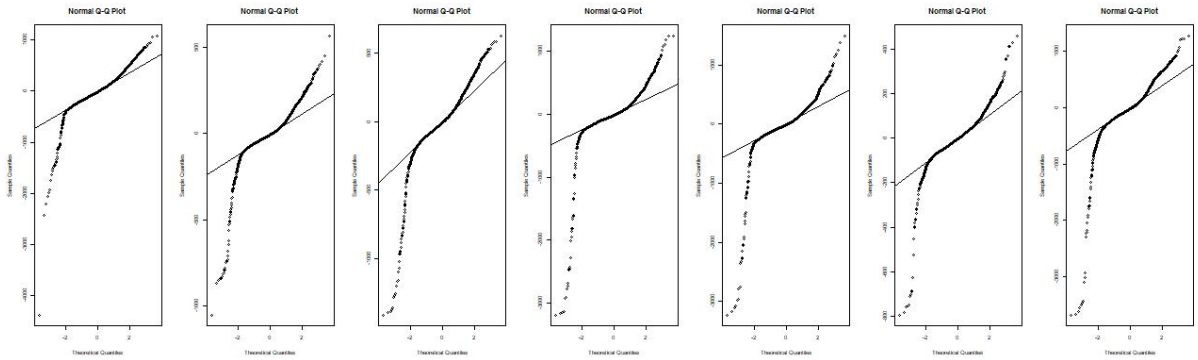
Table 2

Statistical test for stationarity and normality

Sensor	Augmented Dickey-Fuller test		Shapiro-Wilk normality test	
	P-value	Dickey-Fuller	P-value	W
Sensor 1	<0.01	-25.891	$<2.2e^{-16}$	0.8432
Sensor 2	<0.01	-25.121	$<2.2e^{-16}$	0.80366
Sensor 3	<0.01	-23.307	$<2.2e^{-16}$	0.84724
Sensor 4	<0.01	-27.797	$<2.2e^{-16}$	0.63344
Sensor 5	<0.01	-28.201	$<2.2e^{-16}$	0.6901
Sensor 6	<0.01	-25.648	$<2.2e^{-16}$	0.83786
Sensor 7	<0.01	-23.813	$<2.2e^{-16}$	0.77205

Figure 2

Q-Q plots of the residuals of the decomposition



A Shapiro-Wilk test also tests the normality but from the statistical side. It tests the null hypothesis that a sample comes from a normally distributed population. The p-values here as shown in Table 2 are all smaller than 0.05, so we reject the null hypothesis to say the residuals of the decomposition of the traffic volume data from

seven sensors are not normally distributed.

4.3 Stationarity and normality of the data

As the residuals are stationary, we are allowed to fit ARMA models on them. But the residuals are not normally distributed, which means the residuals are somehow still correlated with each other. This affects the ARMA models and is one reason why the residuals in the ARMA models show statistically significant autocorrelation at some lags.

Table 3

ARMA models of each dataset

Sensor	ARMA model
Sensor 1	$X_t = 0.5608X_{t-1} - 0.0889X_{t-2} - 0.0743X_{t-3} - 0.0415X_{t-4} - 0.0762X_{t-5} - 0.0884X_{t-6} + Z_t$
Sensor 2	$X_t = 0.7285X_{t-1} - 0.2031X_{t-2} - 0.0662X_{t-3} - 0.0378X_{t-4} - 0.0453X_{t-5} - 0.0502X_{t-6} + Z_t$
Sensor 3	$X_t = 0.6207X_{t-1} - 0.0697X_{t-2} - 0.1426X_{t-3} - 0.0442X_{t-4} - 0.0504X_{t-5} - 0.0613X_{t-6} + Z_t$
Sensor 4	$X_t = 0.7828X_{t-1} - 0.1967X_{t-2} - 0.0571X_{t-3} - 0.1314X_{t-4} - 0.0329X_{t-5} - 0.0598X_{t-6} + Z_t$
Sensor 5	$X_t = 0.7811X_{t-1} - 0.1663X_{t-2} - 0.0213X_{t-3} - 0.0753X_{t-4} - 0.0442X_{t-5} - 0.0316X_{t-6} + Z_t$
Sensor 6	$X_t = 0.6309X_{t-1} - 0.1025X_{t-2} - 0.0716X_{t-3} - 0.0127X_{t-4} - 0.0616X_{t-5} - 0.0465X_{t-6} + Z_t$
Sensor 7	$X_t = 0.5050X_{t-1} - 0.0674X_{t-2} - 0.1162X_{t-3} - 0.0446X_{t-4} - 0.0725X_{t-5} - 0.0857X_{t-6} + Z_t$

5. Stating the model and checking the assumptions

In order to build a causal and invertible ARMA model, we select a most appropriate ARMA model for each dataset using AIC. Based on this criterion, we have surprisingly found out that ARMA(6,0) is the most ideal model for all datasets. The ARMA models for each data of traffic volume from different sensors are listed in Table 3.

Before we finally accept ARMA(6,0) as the univariate time series model for the short-term traffic volume prediction, we need to make sure the residuals of the ARMA models are white noise. As shown in Figure 3, the ACF and PACF values of the residuals are outside the boundaries, which makes the ARMA(6,0) models look like incorrect.

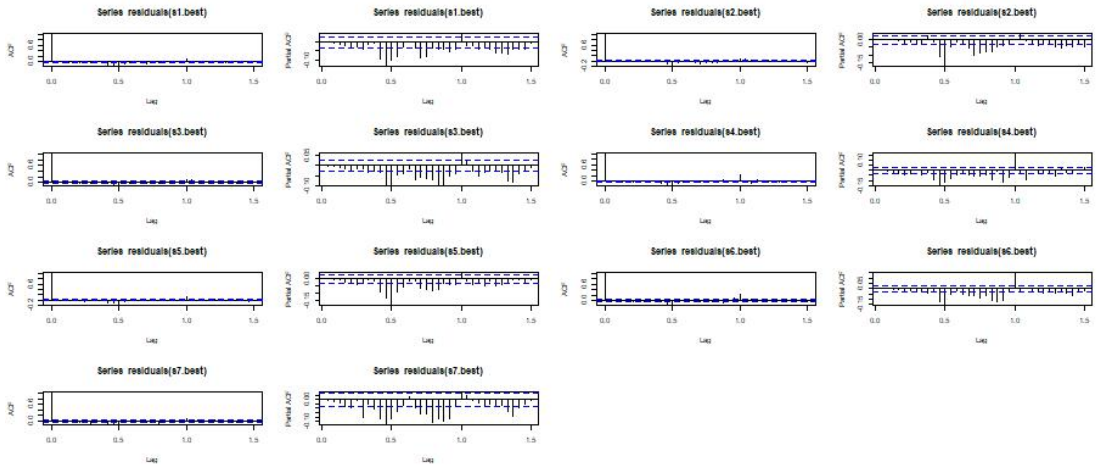
Table 4

Statistical test for independent distribution

Sensor		Sensor1	Sensor2	Sensor3	Sensor4	Sensor5	Sensor6	Sensor7
Ljung-Box test	P-value	0.6834	0.9942	0.9597	0.8893	0.9895	0.9766	0.6833
	W	0.1663	0.0001	0.0026	0.0194	0.0002	0.0009	0.1664

Figure 3

ACF and PACF of the residuals in ARMA models



We also conduct the Ljung-Box test to test the residuals from the statistical aspect. The Ljung-Box test tests the null hypothesis that the data are independently distributed, which means the correlation in the population from which the sample is taken are zero. The p-values here as shown in Table 4 are all greater than 0.05, so we

cannot reject the null hypothesis. Thus, the residuals of the ARMA models are white noise. Therefore, the ARMA(6,0) model is an appropriate model for the short-term traffic volume prediction.

6. Executing the analysis and presenting the results

In this part, we will analyze the results of the decomposition and the ARMA models. We have decomposed seven datasets of the traffic volume from different sensors. Each dataset has the similar trend and seasonal components. We will explain the trend and seasonal components of the data from Sensor 1. Also, seven ARMA models will be included.

Figure 4

Decomposition of a five-weekday data of the traffic volume from Sensor 1

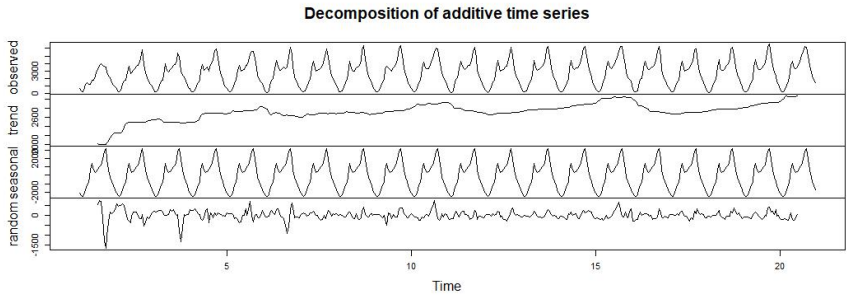
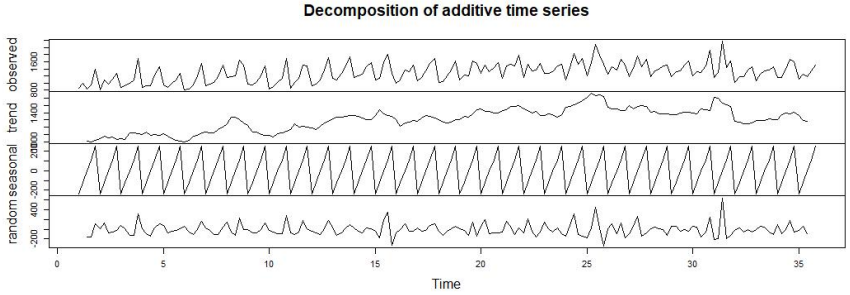


Figure 5

Decomposition of the data of a specific hour of the traffic volume from Sensor 1



s

6.1 The trend component

The data of the traffic volume has an upward hourly trend from 0:00 a.m. to 23 p.m. and a slightly increasing daily trend from the first observation day to the last observation day but with unexpected drops at 0:00 a.m. on some particular days, which should have some relation with the exogenous variables, for example the weather, the road conditions and the holidays.

6.2 The seasonal component

The data of the traffic volume repeats an hourly pattern, which shows the traffic is busier during the daytime than the nighttime and is even much busier during two peak hours. It also repeats a daily pattern, which shows the traffic volume is heavier in the midweek.

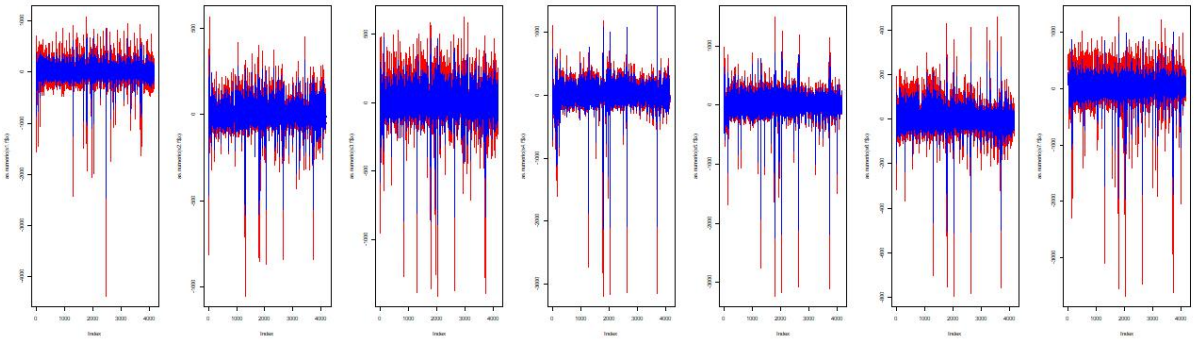
6.3 The ARMA model

Based on the ARMA models, we will interpret the coefficients of the AR terms in this part. The AR(1) coefficients in all seven ARMA(6,0) models are positive, while all the rest coefficients are negative. The coefficient of AR(1) stands for the impact of the traffic volume one hour ago on the current hour's traffic volume. The positive value of this coefficient means the impact is positive. A high traffic volume one hour ago makes the current traffic much heavier. But the traffic volume two to six hours ago negatively affects the current traffic volume. A high traffic volume at the early times only decreases the burden of the traffic. But the extent of the impact of the traffic volume two to six hours ago on the current hour of a single spot varies a lot and also the seven ARMA(6,0) models do not share a common changing pattern from AR(2) to AR(6). Although the

orders of the ARMA models can be determined, the coefficients of each ARMA model can only be decided after a deeper insight into the data.

Figure 6

The predicted values(blue) vs. the true values(red)



The goal of our project is to predict the short-term traffic value. We calculated the error between the observed values in the ARMA models and the true values in our dataset. Here we use the root-mean-square error(RMSE), which represents the square root of the second sample moment of the differences between predicted values and the observed values. It compares forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent. The values of the RMSE for each ARMA model on different datasets are shown in Table 5. For example, we can interpret the RMSE value of Sensor 1 as the deviations of the predicted values from the true values is 212.6783.

Table 5

RMSE of the prediction errors

Sensor	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6	Sensor 7
RMSE	212.6783	77.9012	133.0920	179.0912	194.4596	62.1235	286.4554

7. Discussion and future studies

We have made a simple prediction based on the univariate time series model. But the traffic volume could also be affected by many exogenous variables, which could be better interpreted from a multivariate time series model (i.e. a multivariate short-term traffic flow forecasting considering the trend, seasonal, cyclical and calendar variations (B Ghosh, B Basu and M O'Mahony, 2006)). Moreover, a large amount of literature has been concerned with predictions from a single spot to multiple spots based on state-space model (Yao Zhi-Sheng, and Chun-Fu Shao, 2007, and Anthony Stathopoulos and Matthew G. Karlaftis, 2002).

Due to the stochastic nature of traffic flow and highly nonlinear characteristics for short-term prediction, artificial intelligence techniques have received much attention and are considered as an alternative for traffic flow prediction model. A central issue is to find a model structure that contains more statistically significant features and improves the prediction. One previous study (R. Yasdi, 1999) has considered the past values to predict the future value using the Recurrent Neural Network. Now more studies consider exogenous factors like the average speed of all vehicles (Theja and Vanajakskshi, 2010; Celikoglu and Cigizoglu, 2007), and build a more stable and efficient Long Short-Term Memory Neural Network for traffic speed prediction by determining the optimal time window for time series in an automatic manner (X Ma, Z Tao, Y Wang, H Yu and Y Wang, 2015).

Our future work will first focus on a more complicated time series model which takes exogenous variables into account and expands from the univariate to multivariate.

Then, based on the clear insight got from the complicated time series model, we will try to build an artificial neural network to determine the optimal time window for time series in an automatic manner.

8. References

Ahmed, Mohammed S., and Allen R. Cook. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. No. 722. 1979.

Okutani, Iwao, and Yorgos J. Stephanedes. "Dynamic prediction of traffic volume through Kalman filtering theory." Transportation Research Part B: Methodological 18.1 (1984): 1-11.

Vlahogianni, Eleni I., Matthew G. Karlaftis, and John C. Golias. "Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume." Transportation Research Part C: Emerging Technologies 14.5 (2006): 351-367.

Ghosh, Bidisha, Biswajit Basu, and Margaret O'Mahony. "Bayesian time-series model for short-term traffic flow forecasting." Journal of transportation engineering 133.3 (2007): 180-189.

Ghosh, Bidisha, Biswajit Basu, and Margaret O'Mahony. "Analysis of trend in vehicular traffic flow data by wavelets." (2006): 415-419.

Yao, Zhi-Sheng, and Chun-Fu Shao. "Road traffic state multi-spot time series forecasting based on state space model." Zhongguo Gonglu Xuebao(China Journal of Highway and Transport)20.4 (2007): 113-117.

Stathopoulos, Antony, and Matthew G. Karlaftis. "Modeling duration of urban traffic congestion." Journal of Transportation Engineering 128.6 (2002): 587-590.

Yasdi, Ramin. "Prediction of road traffic using a neural network approach." Neural computing & applications 8.2 (1999): 135-142.

Kumar, Kranti, M. Parida, and V. K. Katiyar. "Short term traffic flow prediction for a non urban highway using artificial neural network." Procedia-Social and Behavioral Sciences 104 (2013): 755-764.

Celikoglu, Hilmi Berk, and Hikmet Kerem Cigizoglu. "Public transportation trip flow modeling with generalized regression neural networks." Advances in Engineering Software 38.2 (2007): 71-79.

Ma, Xiaolei, et al. "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data." Transportation Research Part C: Emerging Technologies 54 (2015): 187-197.