Forecasting Traffic Flow in Bellevue, Washington

Theron Spencer, 2019

**Abstract:**

The traffic flow at a location on Interstate-90, one-half of a mile east of the Interstate-405 junction, is modelled in this paper. Four different time series models are implemented: SSA (Singular Spectrum Analysis), SARIMA (Seasonal ARIMA), TBATS (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components), DHR (Dynamic Harmonic Regression). The data was gathered by the Washington State Department of Transportation (WSDOT) using a permanent traffic recorder (site ID: S203).

**Background and Research Objective**

Road infrastructure contributes greatly to society by stimulating economic and social development. These road networks provide access to varying employment, social, health and educational services. For these reasons, it’s important we maximize transportation efficiency. This can be achieved in part by modeling traffic flow as a time series to gain the insight needed to make the necessary changes to reduce congestion and increase efficiency.

The Washington State Department of Transportation has permanent traffic recorders strategically set up all throughout the state. The traffic recorder of particular interest in this paper is one on I-90, just a half a mile away from the I-405 junction (site ID: S203). (Data, 2019) The recorder records the total traffic flow each hour in both the East and West direction.

The goal of this paper is to uncover any underlying patterns in the traffic flow and identify which of the models preforms the best at this particular location. We will focus primarily on East flowing traffic from the 6/1/2019 to 8/31/2019. (Fig. 1a)

**2 Methods**

The subset of data just described was chosen as it didn’t have any missing data so no data imputation was necessary. Additionally no transformation was necessary as shown by the Box-Cox transformation, it was close enough to 1 (Fig. 1b). The following are the models implemented:

**2.1 Singular Spectrum Analysis (SSA)**

Singular Spectrum Analysis is a nonparametric spectral estimation method. SSA aims to decompose the time series into the sum of interpretable components including trend, periodicities and noise. SSA consists of two stages: decomposition and reconstruction. The first stage of decomposition involves embedding the original time series into a sequence of lagged vectors. These vectors are then used to create a trajectory matrix. This trajectory matrix is then decomposed using Singular Value Decomposition (SVD). During the second stage, eigentriple grouping and diagonal averaging is done to reconstruct a series. That is, the initial series is decomposed into the sum of reconstructed series (Golyandina & Korobeynikov, 2013). In order to construct the proper decomposition, parameters must be chosen correctly. This can be done by viewing the eigenvectors and eigenvector pairs. Trend can be identified by a single high-magnitude component norm (Fig. 2a/b). Periodicities can be identified in the graphs of paired eigenvectors that form a regular T-vertex polygon (Fig. 2c). After SSA is preformed, if correlation is present in the residuals, the residuals can be modeled using some (S)ARIMA process.

**2.2 SARIMA**

ARIMA stands for autoregressive moving average and is essentially an ARMA of the differenced time series. SARIMA is a generalization of ARIMA(p,d,q) , where the seasonal component with period m is also modeled by an ARIMA(P,D,Q) process. That is, SARIMA can be specified by SARIMA(p,d,q)x(P,D,Q)m.

If the series can be modeled as a SARIMA process it takes on the following form:

where B is the backshift operator and (Kimihiro, 2019).

**2.3 Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components (TBATS)**

TBATS is an innovations state space modeling framework that is used for forecasting complex seasonal time series such as those with multiple seasonal periods, high-frequency seasonality, non-integer seasonality, and dual-calendar effects. TBATS compares various models including those with/without the following: Box-Cox transformation, trend, trend damping, ARMA(p,q) process used to model residuals, non-seasonal model, various amount of harmonics to model seasonal effects (De Livera and Hyndman and Snyder, 2011). Here the harmonics are modeled using Fourier terms explained in the next model description.

**2.4 Dynamic Harmonic Regression (DHR)**

Dynamic Harmonic Regression is a regression method where the periodic seasonality is handled using pairs of Fourier terms and where the error terms are modeled by a non-seasonal ARIMA process. This model is of particular interest when there are long seasonal periods (over a year, for example) or when there are multiple seasonalities. This model assumes the seasonal pattern is unchanging and can be represented as the following:

Where *m* = seasonal period, and are regression coefficients, and is the y-intercept (Hyndman, 2018).

**3 Results**

(Fig. 4) shows an 18 day forecast for each model with associated prediction intervals. However, it should be noted that for or all the models, none of their residuals were normally distributed. Additionally, they did satisfy the assumption of homoscedasticity. As a result of this non-normality, these prediction intervals are unreliable.

From the SSA model, some periodicities were identified. The most significant periodicity was identified to be 24 hours which corresponds to an expected daily seasonality. Additionally, two other periodicities were identified to be 12 hours and 8 hours. These could correspond to sub daily patterns like rush hours occurring in the morning and afternoon or could just be overfitting. Surprisingly, SSA didn’t pick up the weekly seasonality (changing window length didn’t affect this). This could be because the weekly pattern changed over time which would make it not significant or there wasn’t enough training data used to capture this pattern.

After fitting SSA, residuals showed clear correlation and (Fig. 3) so using small sample corrected Akaike information criteria (AICc), a SARIMA(1,0,0)x(0,1,0)168 process with seasonality of 168 hours (7 days) was fit to the residuals. The second model was a SARIMA model fit to the original series. Without forcing the seven day seasonality, several SARIMA models were suggested including a ARIMA(0,0,4), ARIMA(5,1,2) model. However these failed to capture the weekend dips in traffic flow so after forcing the 168 hour seasonality, the final best model was found to be SARIMA(1,0,0)x(0,1,0)168, similar to that fit to the residuals of the SSA model. The next model was automatically determined by the TBATS framework. The model was Box-cox transformed with lambda equal to 0.292, with the errors modeled by a ARMA(3,5) process, a 24 hour seasonality using 7 Fourier terms and a 168 hour seasonality using 5 Fourier terms. No damping was required. See (Fig. 5) for TBATS visual decomposition. Lastly the DHR model was fit using k1=10 Fourier terms to model the daily seasonality and k2=30 Fourier terms to model the weekly seasonality. The optimum amount of Fourier terms were found by comparing various models with k1=1,…,12 and k2=1,…,30 using AICc.

The RMSE function in the “Metrics” package was used to assess model accuracy. The RMSE’s were determined based off an 18 day traffic flow forecast compared to the actual traffic flow. (Fig. 6) shows the RMSE of each model, with the lowest being 302.02 for the SARIMA model and 310.39 for the SSA and ARMA model. In comparison, the RMSE for the TBATS and DHR models were quite a bit higher at 465.49 and 945.80, respectively.

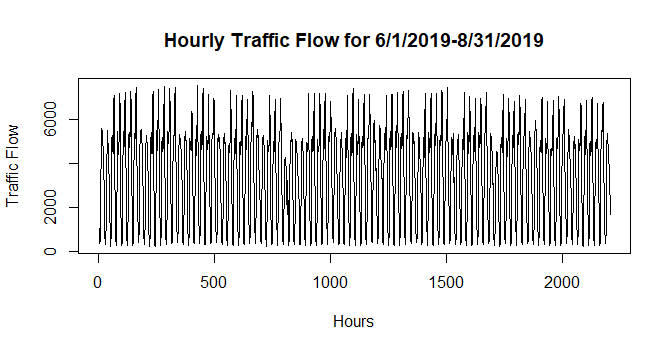
**4 Discussion**

For three months of hourly traffic data, it appears a forced seasonal SARIMA(1,0,0)x(0,1,0)168 performs the best compared to the other three models. SSA identified some seasonalities that weren’t immediately apparent and warrants investigation to see if they are indeed relevant or if they are just a result of overfitting. As already mentioned, the prediction intervals given are unreliable. This could be remedied by calculating the prediction intervals from bootstrapped residuals (Hyndman, 2018). Also, from visual inspection of the original time series (Fig. 1a), there appears to be a significant dip in traffic about one third of the way in, this corresponds to July 4th/5th. None of these models took that holiday into consideration so it might be a good idea to create dummy variables for holidays in the future. Also, It might be beneficial to (a) see how additional models perform compared to the four already evaluated here, (b) see how these models and others perform in terms of short-term forecasting when trained with only a week or less of higher resolution data, (c) see how these models and others perform in terms of long-term forecasting when trained with a year or more of data. Admittedly, these three month time series models probably don’t have much practical applications in the real world. However, short-term and long-term models might. Long-term models would be able to help identify trends and seasonalities that aren’t apparent at a weekly/monthly scale. These could be used to better predict when maintenance of roads will be needed or additional construction is necessary to reduce traffic congestion. Similarly, short term models using higher resolution data could yield insights to trends and seasonalities not apparent in the lower resolution weekly/montly data. Aside from these considerations, it might be beneficial to implement some kind of spatio-temporal model such as STARIMA or BVAR that takes into account the time series data of neighboring traffic recorders.

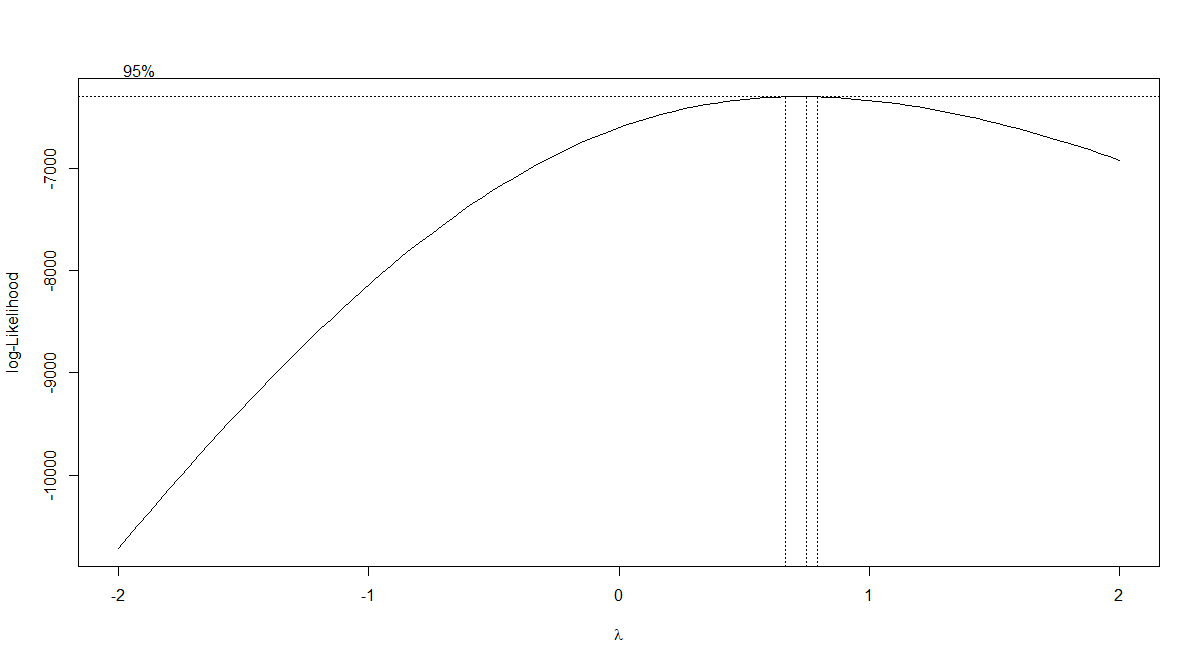
**References**

1. Data for Public Traffic Recorder S203. (Jan. 2007-Aug. 2019). Washington State Department of Transportation. Web. <https://www.wsdot.wa.gov/Traffic/API/PermanentTrafficRecorder/?siteId=S203&startYear=2007&endYear=2019&startMonth=1&endMonth=8&tables=TrafficSpeedByHour>
2. Golyandina, N., & Korobeynikov, A. (2013). Basic Singular Spectrum Analysis and Forecasting with R. Computational Statistics & Data Analysis.
3. Noguchi, Kimihiro. (2019). Time Series Modeling in Practice.
4. De Livera, Alysha M., Hyndman, Rob J. & Snyder, Ralph D. (2011) Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing, Journal of the American Statistical Association, 106:496, 1513-1527, DOI: 10.1198/jasa.2011.tm09771
5. Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on <12/01/2019>.

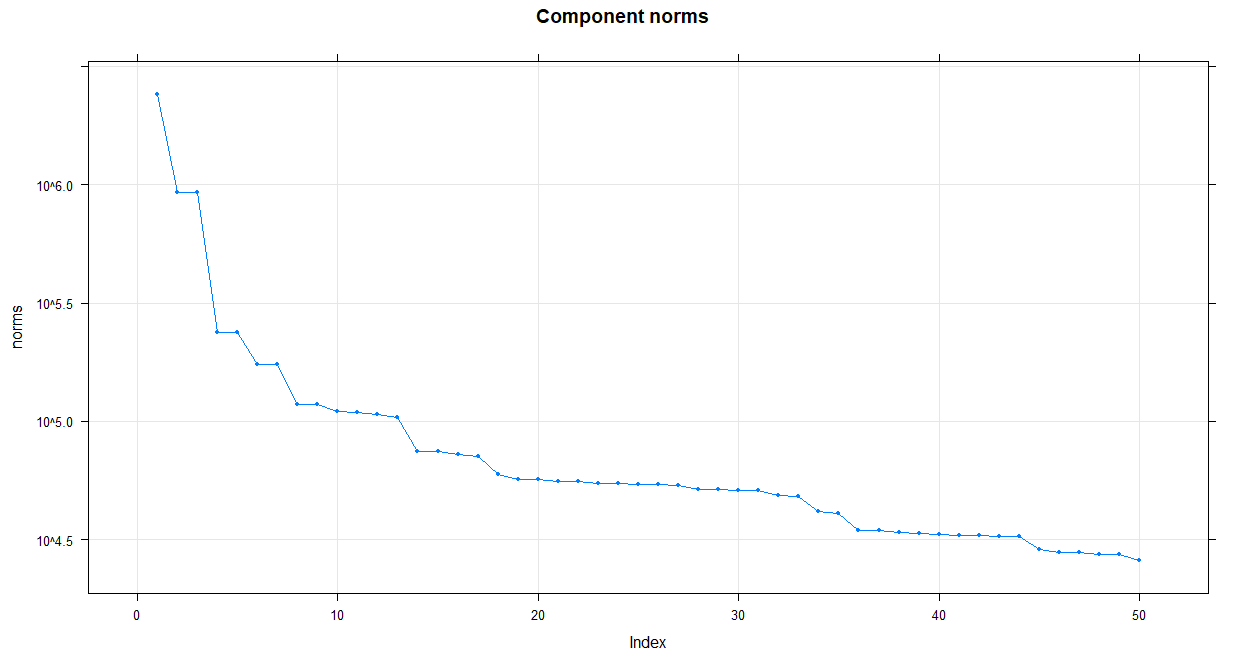
**Figure 1a.**



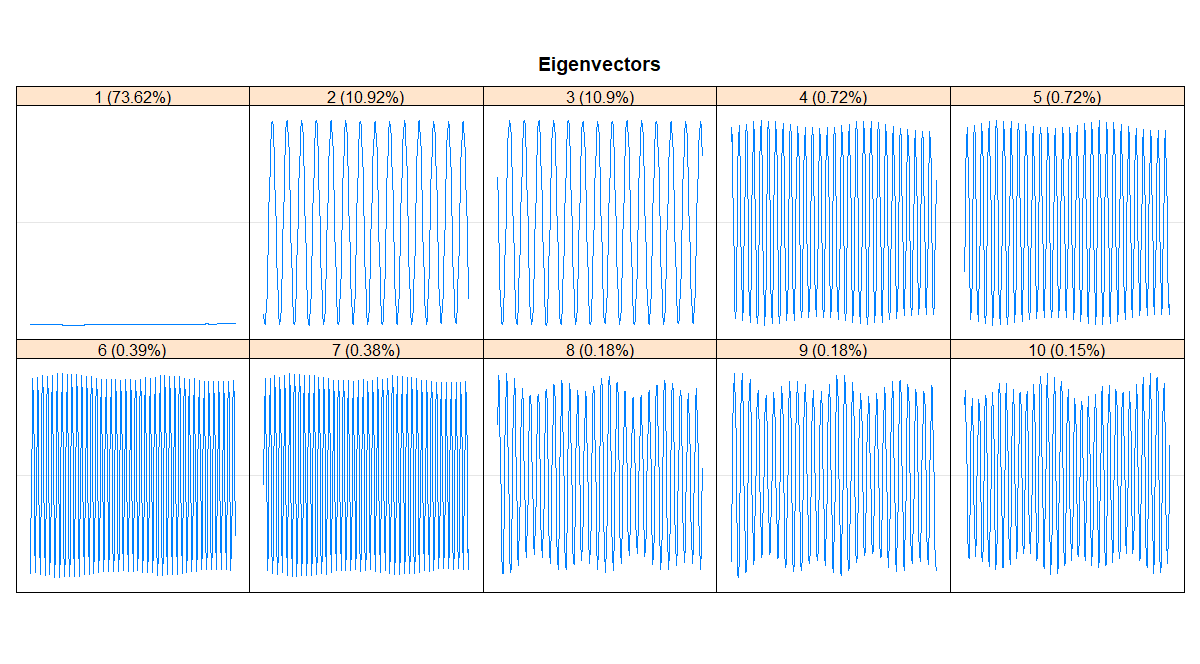
**Figure 1b.**



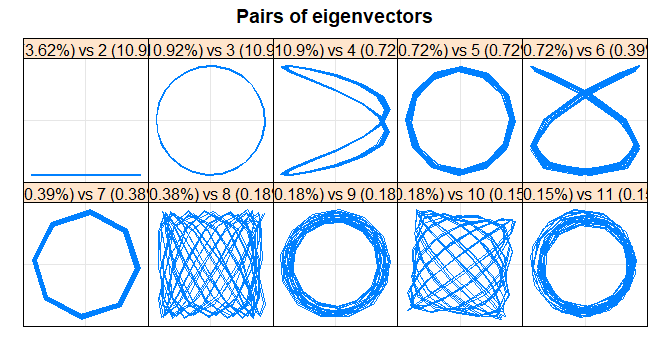
**Figure 2a.**



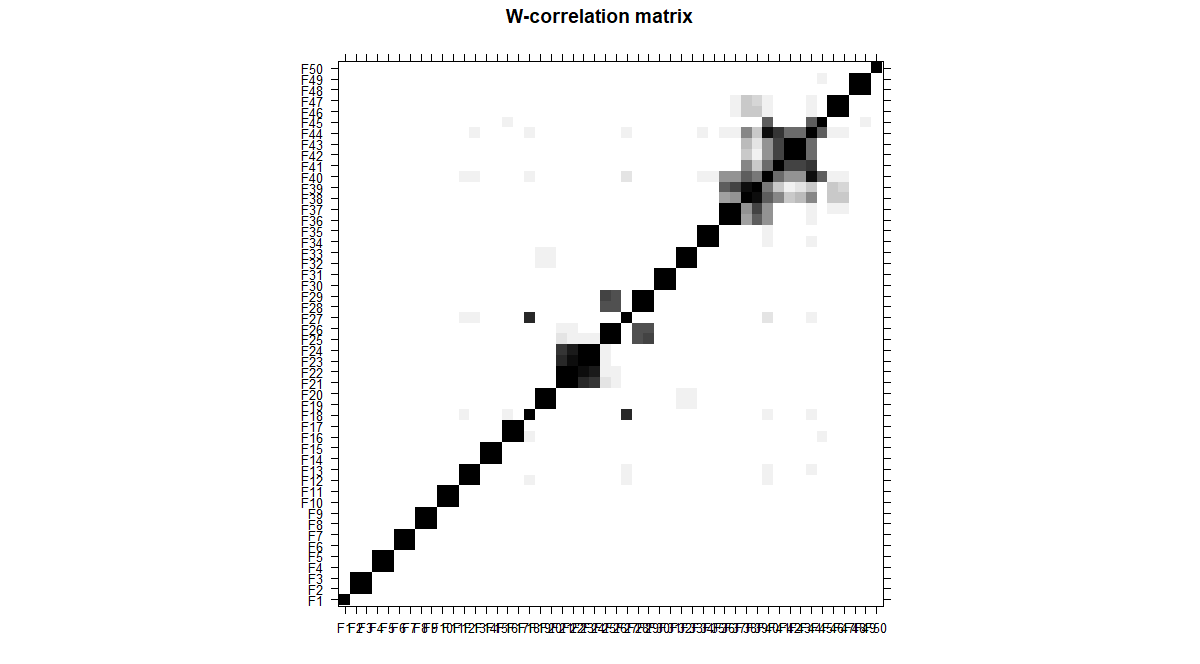
**Figure 2b.**



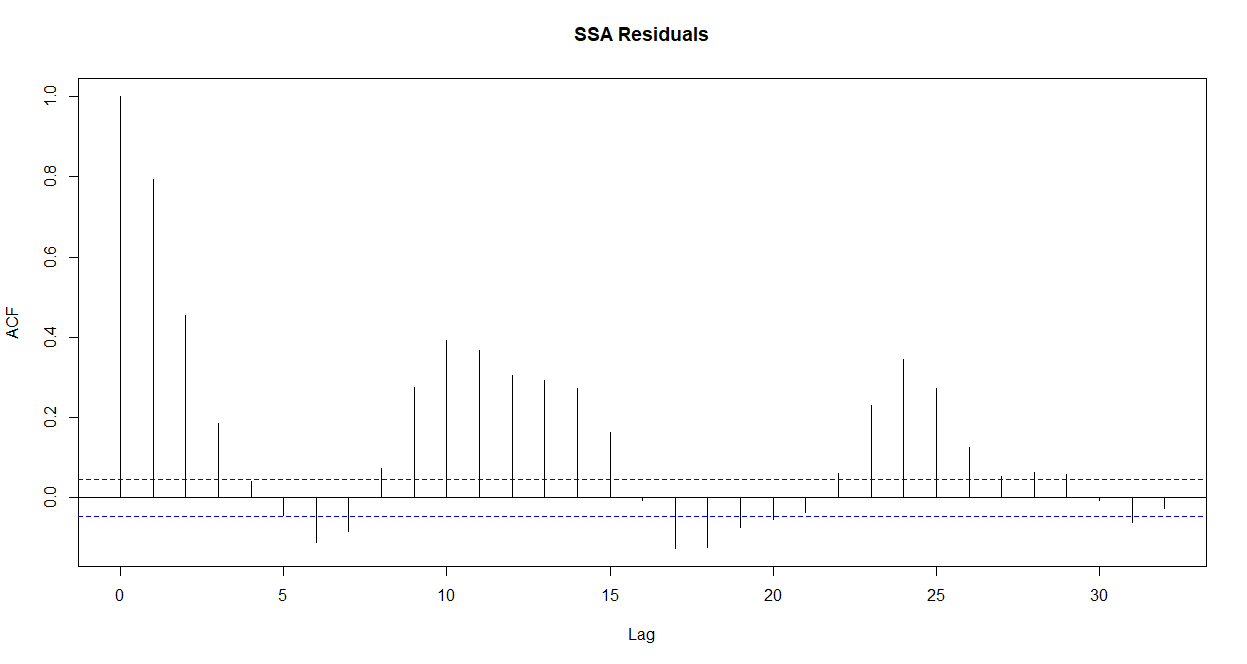
**Figure 2c.**



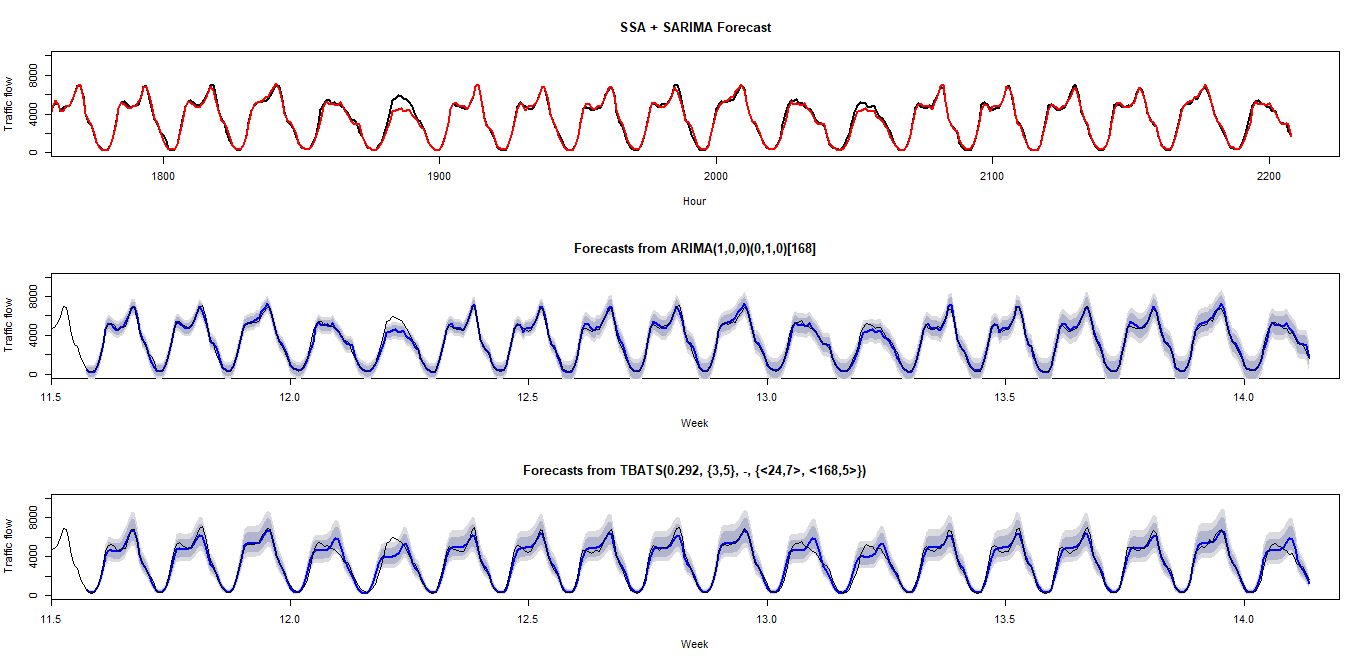
**Figure 2d.**

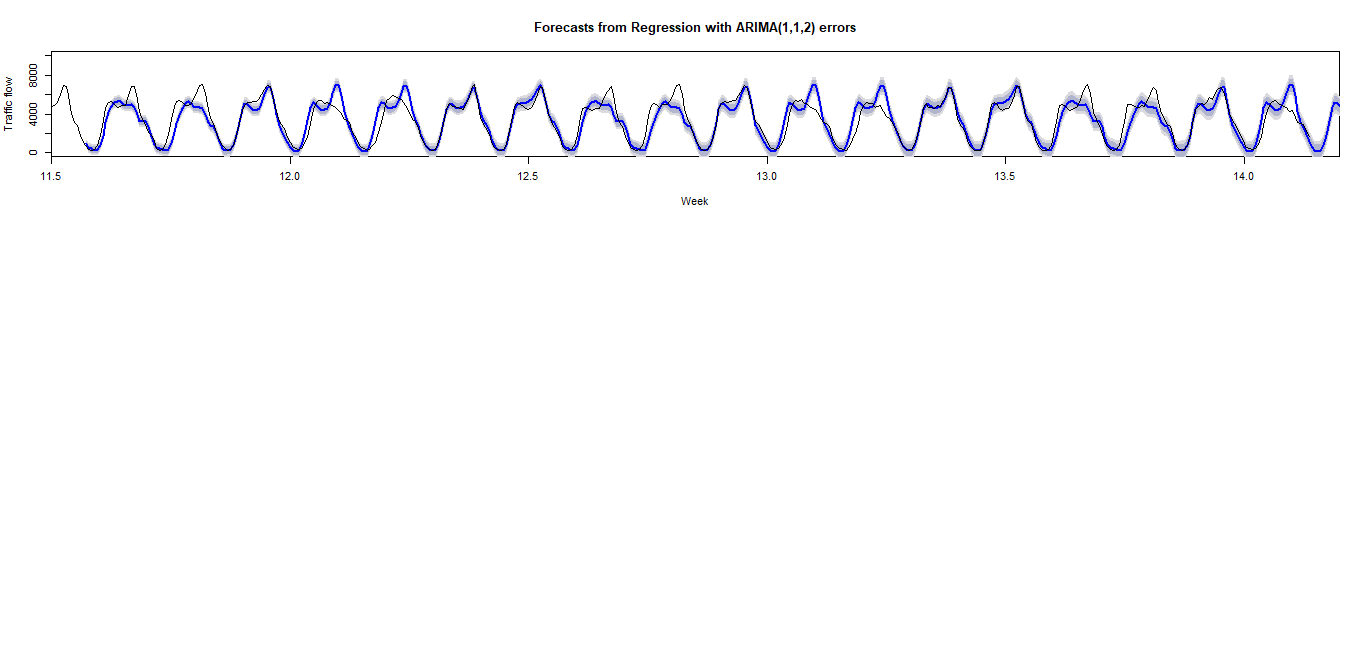


**Figure 3.**

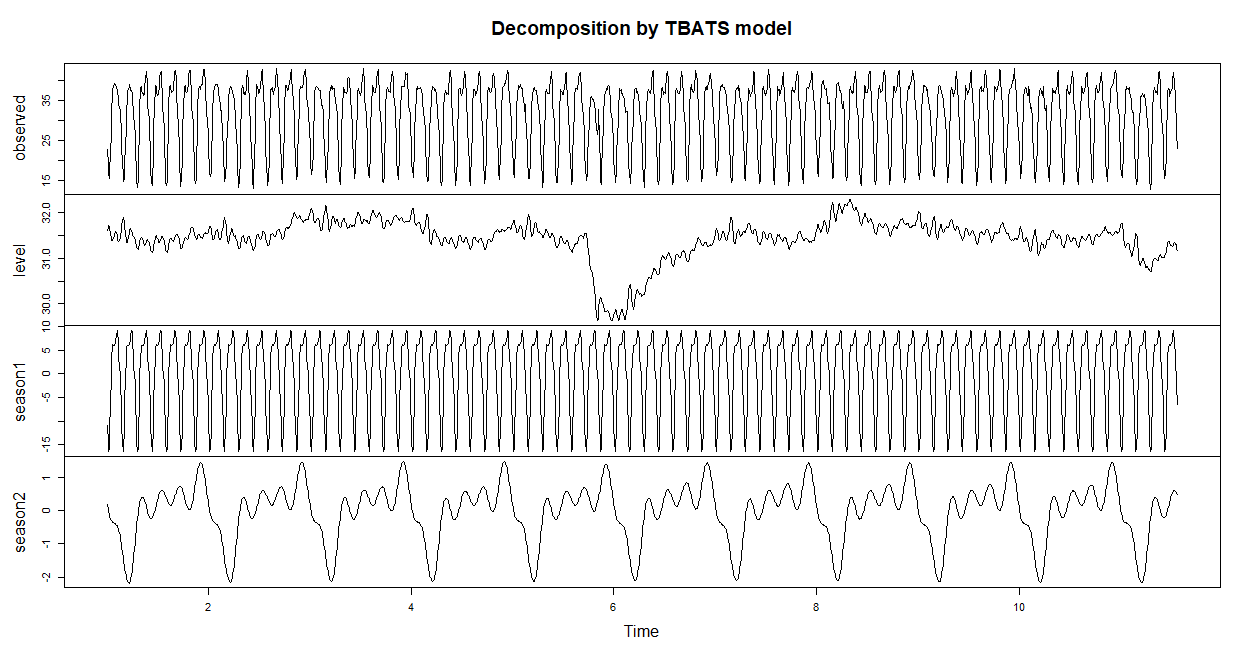


**Figure 4**





**Figure 5**



**Figure 6**

SSAandARMA\_rmse SARMA\_rmse tbats\_rmse dhr\_rmse

[1,] 310.3854 302.0173 465.4923 945.7953