****

**Assignment 11 – Draft Manuscript (Discussion and Conclusion)**

**Vehicle Traffic, Provided by City of Aarhus in Denmark**

**Presented by:** Amit Pandey

**Guided by:** Professor Sylvain Jaume

**Course:** DS670 - Capstone

**Manuscript**

The urban traffic congestion is transforming into an epidemic all over the world. The brisk increase in vehicle traffic has become one of the critical problems faced by cities all over the world. As a result of the constant traffic congestion, transportation cost has increased significantly due to all the time wasted on the road and the corresponding fuel cost. This proposal looks at how the Vehicle Traffic data collected can be used in making Aarhus a smart city in terms of traffic guidance and management. City Pulse project services smart city solution, by means of interpreting big data from Internet of Things and social networks. This project focus on Arhus city sensor data sources.

**Discussion:**

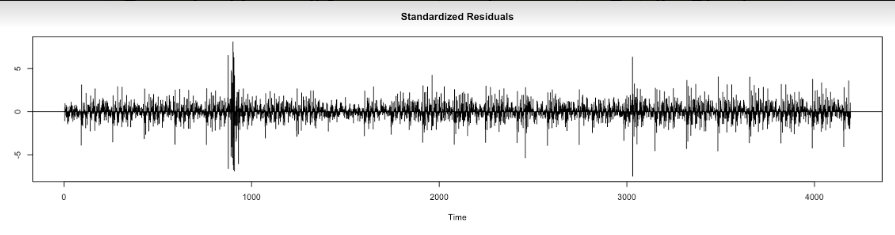
The dataset assigned for the capstone project is the “Vehicle Traffic, Provided by City of Aarhus in Denmark” from CityPulse dataset collection. A collection of datasets of vehicle traffic, observed between two points for a set duration of time over a period of 6 months (449 observation points in total). The data is available in raw (CSV) and semantically annotated format using the citypulse information model.

As we are going to do some prediction we split our combined dataset into two different datasets: Training (80%) and Testing (20%). The Training dataset was used to build and test different statistical mode and the Testing dataset was used to evaluate (cross-validate) our model and assumption. We did the splitting and create the Test partition to provide us with a fair assessment of our prediction model build.

The project here has used the smart city data collected for building up a model to predict the traffic for some certain roads that are covered on the data within that smart city. The data was collected on monthly periods and was captured as a brief record sowe need the meta data to unlock the data and make sense on it. We have loaded the data into the system then we have mapped the data into the meta-data and then we got the full information that we needed. The issue that we have faced is to get rid of the not applicable data fields that empty as well as the corrupted data that we faced. The corrupted data is more complicated than not applicable data, as we first need to identify what is corrupted and then decided whether to exclude or substitute. And an example of corrupted is when you see a speed of 600 miles per hour and these definitely means the data is corrupted as the is no way the speed could come to 600 miles per hour in certain roads, maybe all the roads. In order to handle the outliers, what we have done here is we have rejected the anomaly from the data, as this will influence the data as we didn't evacuate it or erase it, as this can be utilized on further analysis. Then we had to load the data into multiple stage then aggregated them tighter to get the full data set in one single frame. Therefore, based on that, we have built up a time series model for predicting the traffic on a giving road.

We did the time-series analysis using ARIMA (Autoregressive integrated moving average) model. ARIMA is a generalization of an autoregressive moving average (ARMA) model, both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

We predicted the next 5 hours (data points) for the time-series using ARIMA and compared them with the actual. Here’s the result of the model:



**Chart 1**:For the existing data, this is the residuals of the predicted fits.

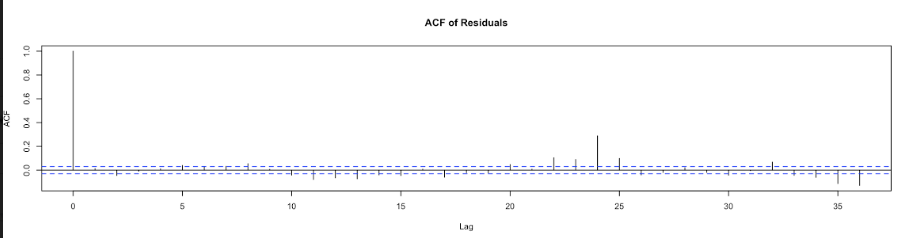
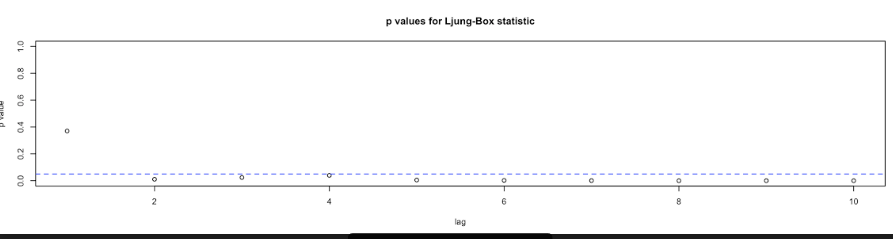


Chart 2: Autocorrelation function (ACF) of residuals shows the correlation of the residuals (as a time series) with its own lags.



**Chart 3:** p-value showing the fitting of the model

**Conclusion**:

The competitor have presented an extensive experimental review of many statistical and machine learning approaches to short-term traffic flow forecasting. Following the approach in SARIMA, they proposed two new SVR models: employing a seasonal kernel to measure similarity between time-series examples. They presented results confirm that seasonality is a key feature in achieving high accuracy; however, the most accurate models often require high computational resources both during the training phase and at prediction time. For this reason, they presented that seasonal kernel approach might be a reasonable compromise between forecasting accuracy and computational complexity issues. In particular, while SARIMA employed in combination with the Kalman filter ends up being the best model on average, the competitors proposed approach is particularly competitive when considering predictions during highly congested periods. The SARIMA version that does not include a Kalman filter and the ANNs perform consistently worse than SVR with an RBF kernel, which, in turn, is less accurate than the seasonal kernel variant. The competitor result show that the accuracy of the Seasonal Mean (SM) predictor starts degrading when the temporal distance between training and test set grows too much, and for the other predictors, no further improvement is observed when using larger training sets including past months, which are too distant from prediction time.

Even though the competitor have used seven different models, their MAPE is coming out to be much higher, as shown below in table 1:

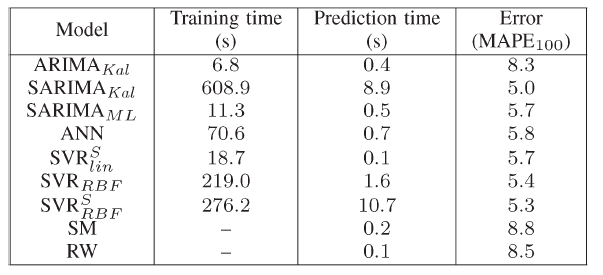


Table 1: TRAINING AND PREDICTION TIME FOR ALL THE ALGORITHMS,AVERAGED ON THE 16 NODES CONSIDERED IN THE TEST SET

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation.

When compared to our model, we are getting a much lower MAPE, which shows that our model is better fitting than the competitor.

