Time Series Forecasting in R for Business: A Comparison of Models

1.Introduction

This guide details the methodology and R code that can be used to evaluate the performance of a time series model and then use the model to forecast a time series. As an example, we use a transportation dataset, but the methdology can be used for any other dataset. The dataset consists of data collected on daily car traffic passing through the Sion-Lausanne tunnel in Switzerland from 2002-01-01 through 2006-02-10[1].

In this forecasting exercise, our main goal will be to predict the car traffic for the remainder of 2006 and then calculate the year over year change in car traffic relative to 2005.

To accomplish this, we first split our dataset into training and test datasets. Then, we build a number of forecasting models (TBATS, ARIMA, Neural Networks) using the training dataset and compare their performances against the test dataset. We show that the TBATS[2] model had the highest degree of accuracy. We then make a forecast using the TBATS model of car traffic to pass through the tunnel in the remainder of year 2006, and calculate the forecasted year over year change in car traffic.

All forecasts in this exercise are implemented using the 'forecast' package in R.

Forecasting Steps

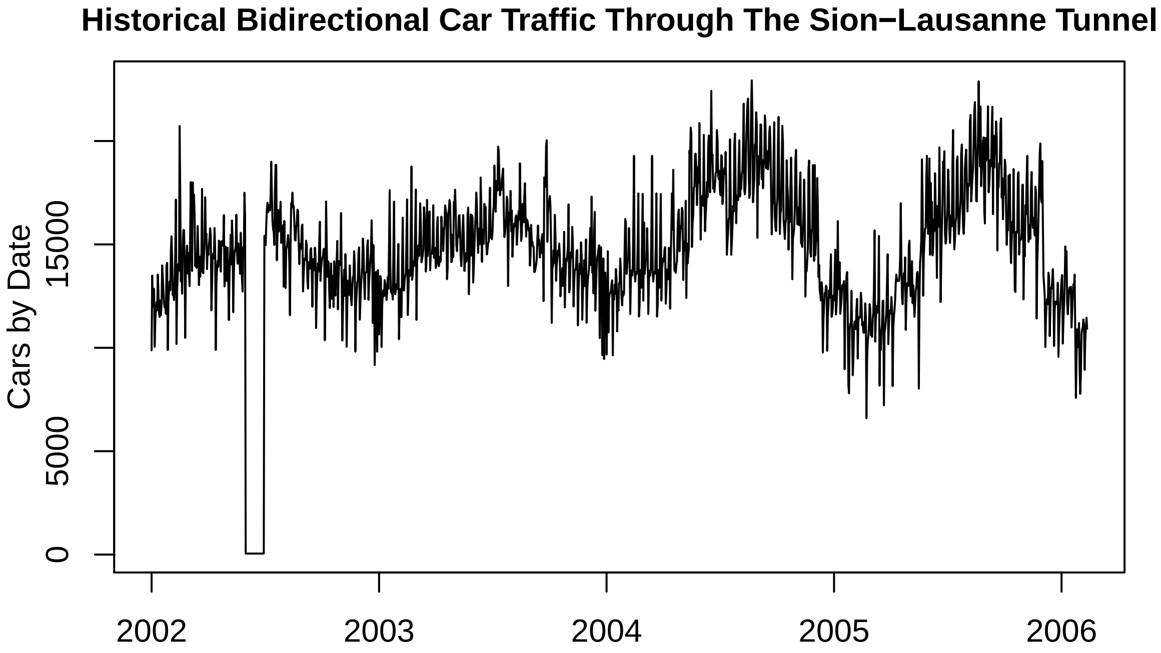
2.1 Dataset Preparation

First, we plot below the car traffic data to check for any obvious trends, seasonality patterns and any possible problems with the dataset. The data appears to show both annual and weekly level seasonalities, with big annual spikes during the summer months and periods of lower traffic during the winter months. There do not appear to be obvious trends, but it does appear that traffic had been gradually increasing since 2002, but then stabilized and showed a very slight decline in year 2005.

[1] This dataset was made available through the T-Competition, a forecasting competition on transporation data. The dataset is available here http://forecastingprinciples.com/index.php/data

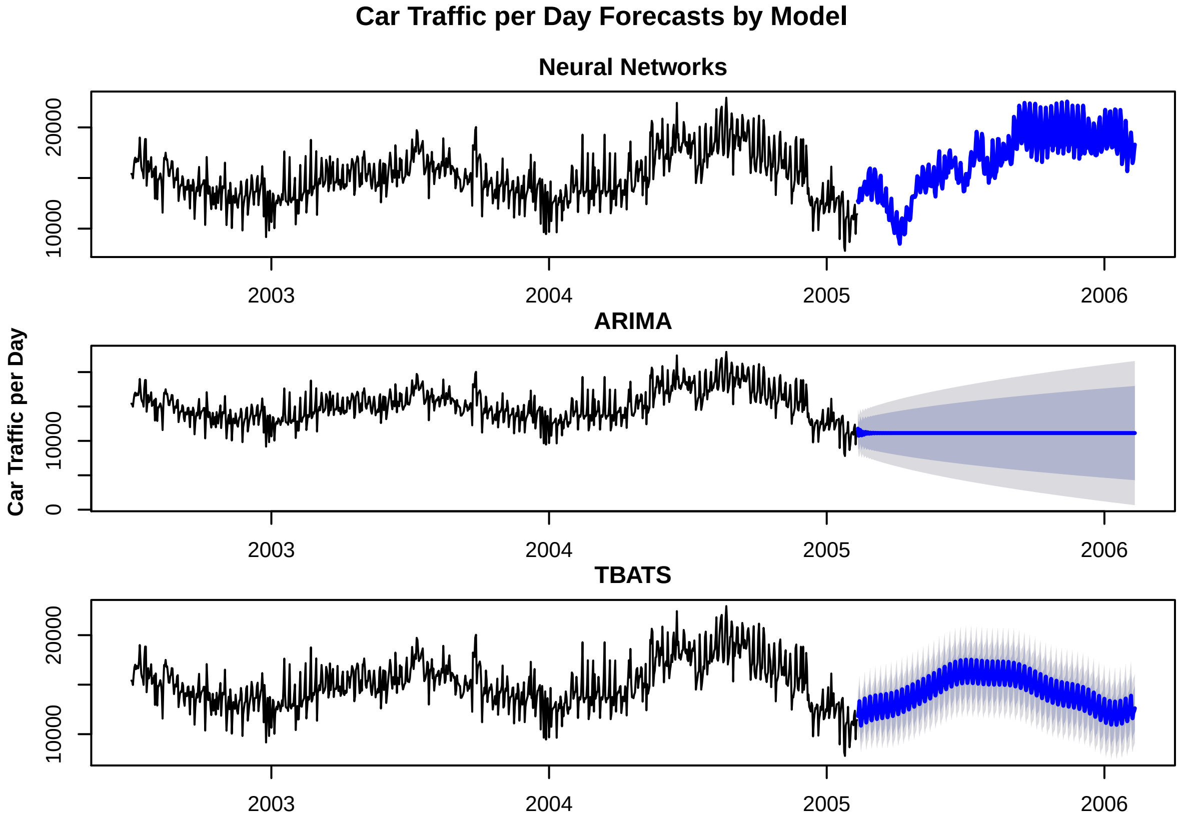
[2] TBATS, short for Trigonometric Box-Cox transform, ARMA errors, Trend, and Seasonal components. For more information: Forecasting time series with complex seasonal patterns using exponential smoothing; Alysha M De Livera, Rob J Hyndman and Ralph D Snyder

Furthermore, we observe that there an anomalous event in June 2002. This could have been caused by a data collection issue or a real-life event like a tunnel closure. In any case, the anomalous event would unnecessarily decrease the accuracy of our forecasts. The missing data can be dealt with in various ways. Here, for the sake of efficiency we simply adjust our data to use only data from after the anomalous incident to build our models. This still leaves use with over 3.5 years worth of data, which is more than sufficient to build a statistical model and perform model testing.



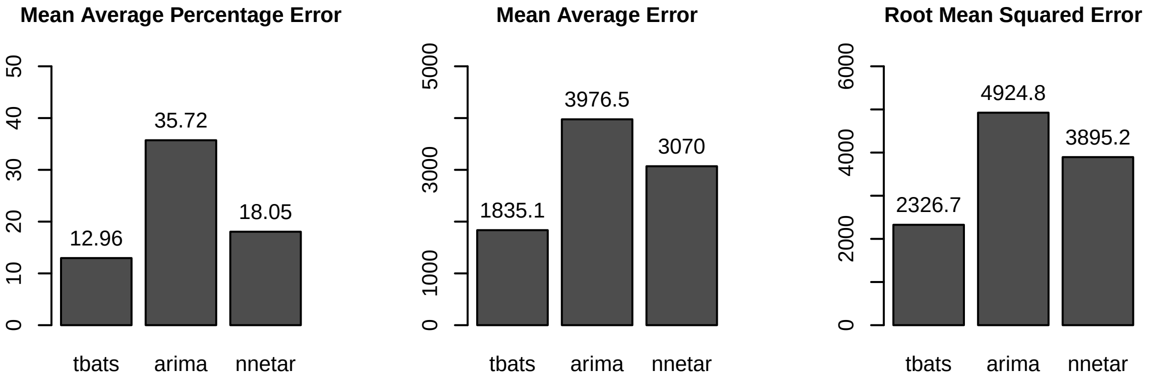
2.2 Model Training and Validation

To identify the best performing model, data from the latest full year (2005-02-11 to 2006-02-10) is excluded from model training as a hold-out set. The data for 2002-07-01 to 2005-02-10 is used to train three different time series prediction models – TBATS, ARIMA and Neural Networks – and the models are used to make predictions for 2005-02-11 to 2006-02-10. The chart below shows the predictions made by each model for that time period.



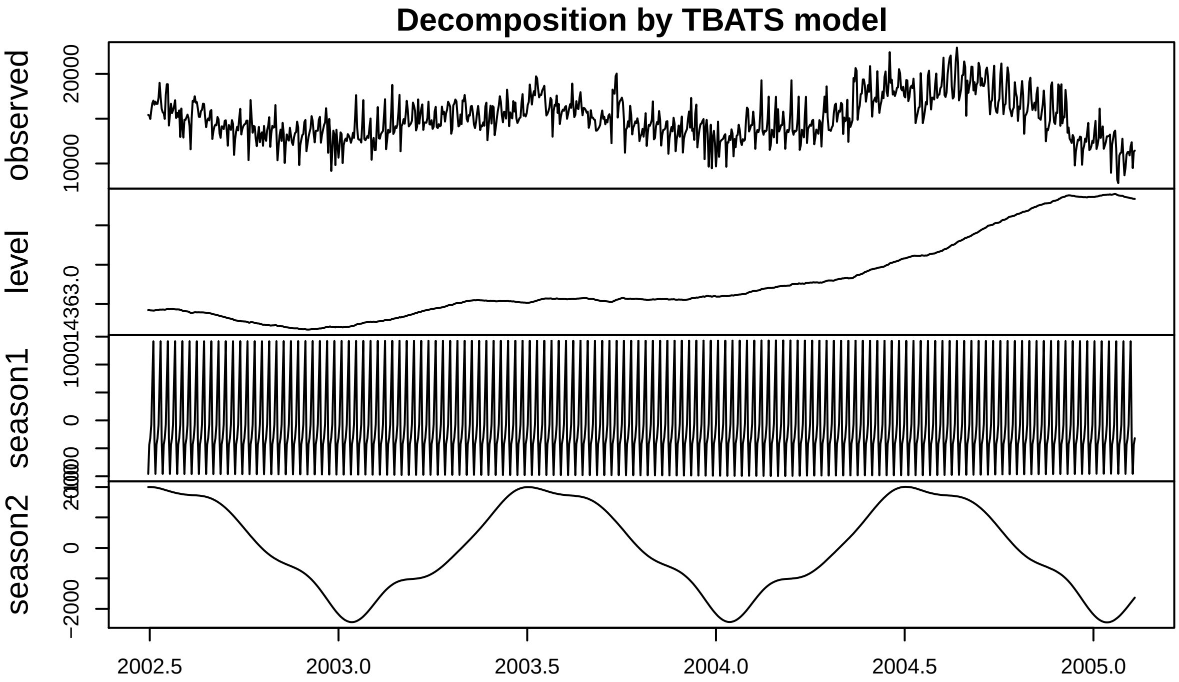
The Neural Network and TBATS models appear to have captured the overall growth trend and seasonality well. In addition, the TBATS model is showing relatively small confidence intervals, indicating high confidence in its predictions. The ARIMA model appears to have less accuracy and provides very broad confidence intervals.

To compare the accuracy of each model, we calculate the Mean Average Percentage Error (MAPE), Mean Average Error (MAE), and the Root Mean Squared Error (RMSE) to measure how much the predictions for the past year from each model diverged from the actual hold-out dataset. The ‘forecast’ package in R contains an accuracy function that will calculate various error metrics automatically. Here we calculate the error metrics manually for illustration purposes.



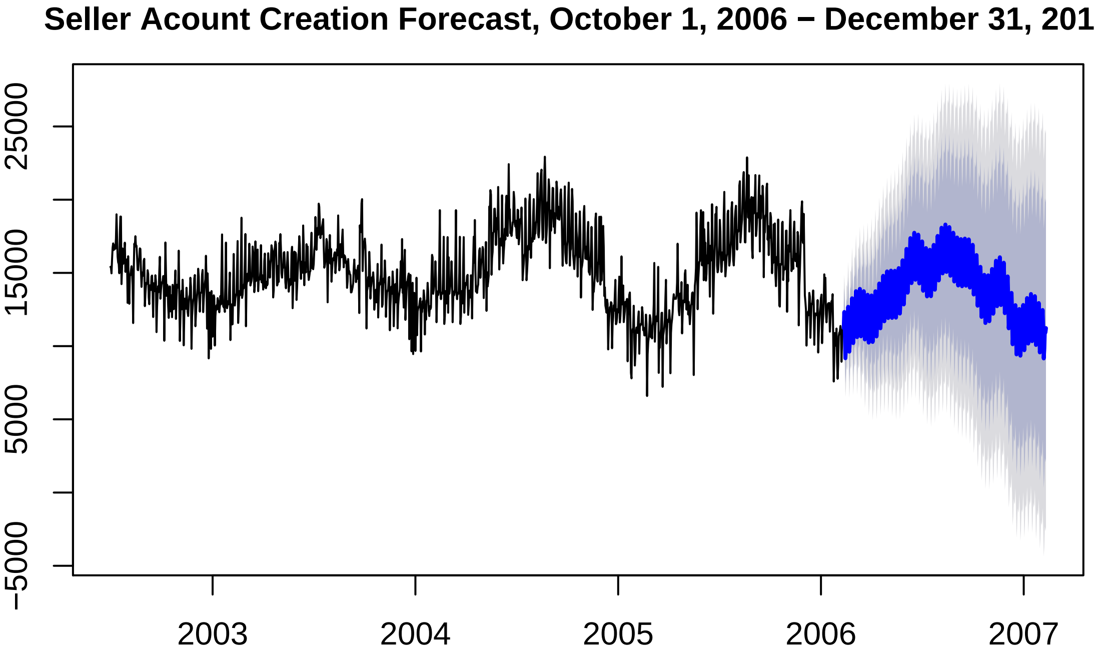
The MAPE for the TBATS model is 12.96%. For the ARIMA model, the MAPE is 35.72% and for the neural networks model, the MAPE is 18.05%. The TBATS model is the best performing according to all metrics. In addition to its performance, the current implementation of the TBATS model in the forecast package is superior to the Neural Networks model because it offers a model decomposition and Confidence Interval calculation functionalities, as well as deterministic performance.

The chart below shows the decomposition of the trends and seasonalities detected in the data by the TBATS model. The ‘level’ trendline shows the overall growth trend, while the season1 and season2 charts show the weekly and annual seasonalities, respectively.



2.3 Predicting Car Traffic For The Rest of 2006

Since the TBATS model showed the highest degree of accuracy and interpretability, it is used to make the forecast for car traffic for the rest of 2006. The TBATS model was retrained on all data (2002-07-01 – 2006-02-10), and a forecast was made for the remainder of 2006. The chart below shows the forecast.



The model predicts that 4, 559, 981 cars will pass through the Sion-Lausanne tunnel in the remainder of

2006, making for a total of $5,031,450$ cars passing through the tunnel in 2006. In 2005, 5423070 cars passed through the tunnel, meaning that the number of cars to pass through the Sion-Lausanne tunnel is expected to decrease by -7.22% in 2006 compared to 2005.

3. References

1. Forecasting time series with complex seasonal patterns using exponential smoothing; Alysha M De Livera, Rob J Hyndman and Ralph D Snyder
2. forecast package in R
3. Tunnel image credit: https://www.autobahnen.ch/images/pic00823.jpg

Links

Github repo link