Modeling Traffic Volume Esther Chen

Summary

The goal of this project was to model and predict traffic volume based on temperature, precipitation, and other weather conditions, and time.

I considered using multiple linear regression, polynomial regression, as well as time series to model, and a bunch of other models too. I tried a polynomial and tree regression, but felt that the predictions would not be great because the relationship between hour to traffic volume has a weird gap in it so the predictions for a polynomial would be quite off. It also ended up being high degrees so I thought at that point I should just swap to a different method since they have high variability. The tree regression also gave a simplistic answer but the reality was more complex, it only selected hour and day as important variables. I think I can agree with that since the effects of other variables were much more marginal when it came down to it.

I thought that a natural model to use would be time series since we are given data in regular intervals. I thought that time would be the most important variable because regardless of weather conditions, people necessarily have to go somewhere for their job or for school unless in extreme weather conditions.

For the multiple linear regression, I tried to use lm and tslm to model traffic volume, using time as the explanatory variable.

For the time series regression, I made exponential smoothing and arima models, as well as stl.

In the end, neither of the models I made were very good at forecasting. I think that the multiple linear regression looks more like the actual data though which was interesting. At the end of the day, I think that I did not create a useful model, but if I could work on the project longer I think that I would have done something.

Data

The dataset consists of 48204 observations of temperature in Kelvin, snowfall in inches, rainfall in inches, traffic volume, a main description of the weather, a more specific description of weather, and whether it was a holiday or not, taken hourly by an automatic traffic detector over the time period from October of 2012 to September of 2018. However, not all the years contained complete records of every hour. There is a gap in the data spanning the latter half of 2014 to the first half of 2015. There is also some missing dates in 2013.

Weather vs. Traffic Volume

There are noticeable differences in the median traffic volume, however the min and max for less severe weather is about the same. The only situation with very different distribution is Squall and perhaps Smoke. Therefore I feel that while weather did play a role, because people need to be places around the year regardless of the weather, I feel that it does not have as significant effect as time, unless it is absolutely dangerous for people to drive or go outside. But I also don't think it is negligible. But I had difficulty including this variable in the analysis because it was hard to do it. There are also very few cases of this weather occuring. The most common weather occurrence was clear and clouds.

Month vs. Traffic Volume

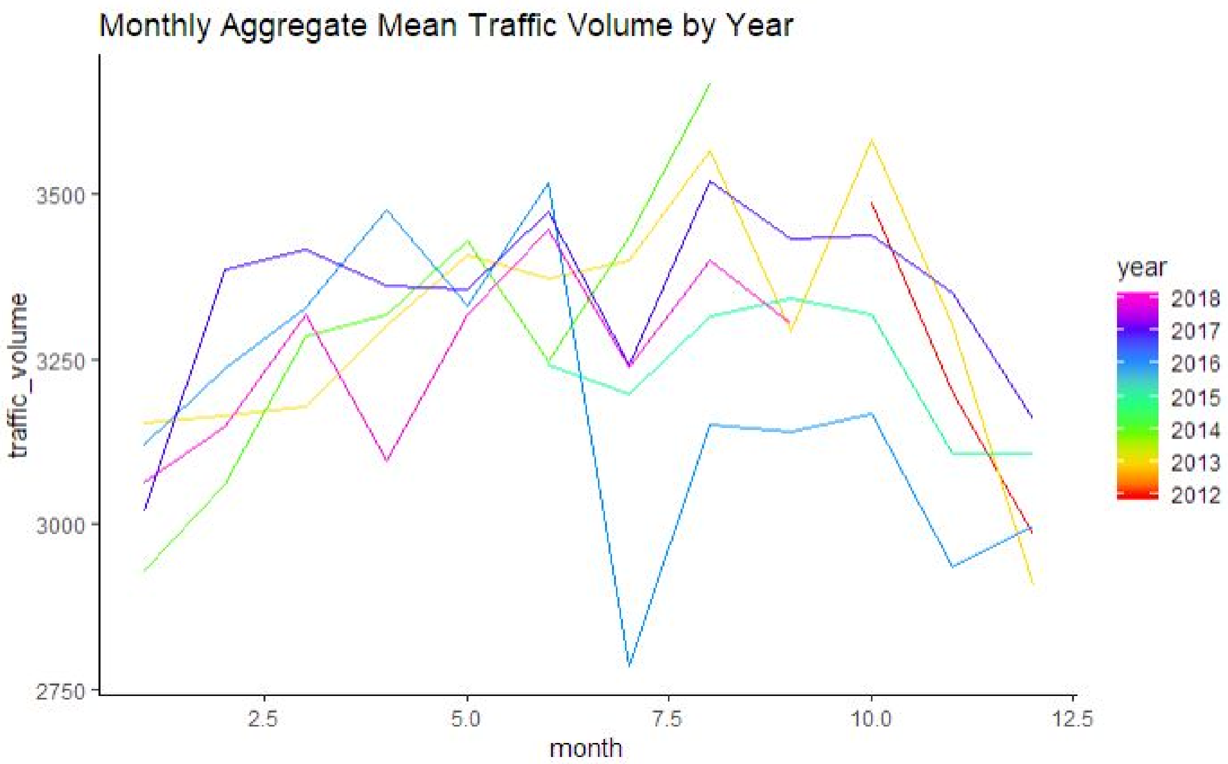
The months had varying medians but the spread was about the same across the months. There is less traffic during the winter months and the average amount of traffic was the highest during spring months.

Weekday vs. Traffic Volume

There is much less traffic over the weekend and it peaks in the middle of the week on Wednesday.

Hour vs. Traffic Volume

Traffic peaks at around 7:00 AM and then at 5:00 PM which makes sense because those are rush hour times. The traffic is lowest at around 3:00 AM. There is a split in the data so it appears that some time there must have been a lower amount of traffic. The spread of those hours is a lot compared to the other hours.



There was no clear pattern over the year except that the mean decreases in winter for all of them, each year had different behavior over the months.

Methods

I made multiple linear regression models using lm and using time as the explanatory variables. The models were assessed by mean absolute percentage error. This is calculated by taking the average of the absolute value of the percentage difference between the real value and the predicted value.

For the simple linear regression, I used all the time variables in the data and I took the aggregated mean of each month to predict. I then turned it into a time series object and forecasted with it.

For the multiple linear regression, I used day and week as explanatory variables with an interaction between day and week. I assessed the models by their adjusted R squared values, QQ plot, and residuals, as well as comparing the fitted values with the actual values from a test set.

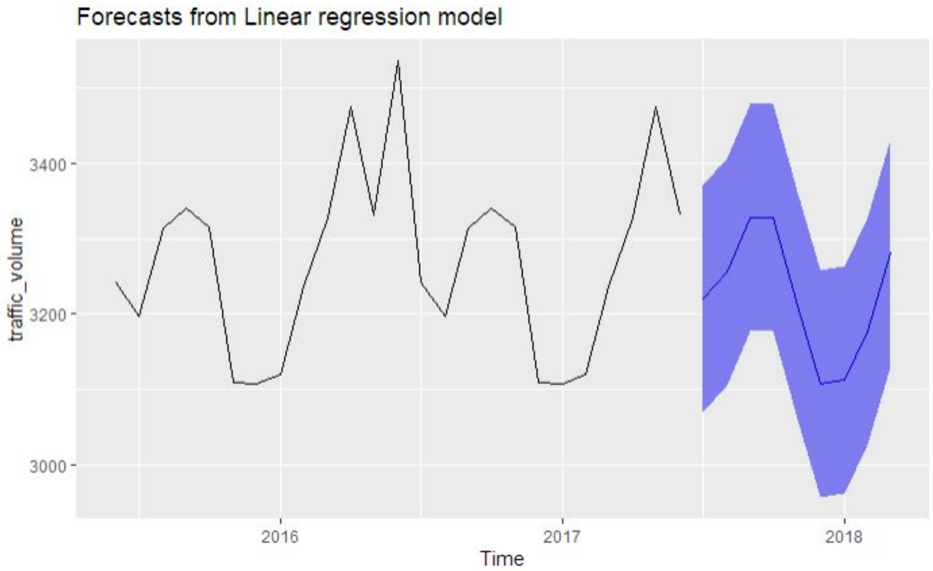
For the time series, I used the entire dataset because I was bored. To train the models, I used previous data and then used the post data to test the model’s forecasting abilities. The forecasts were assessed based on the mean average percentage error.

The first model I used was ets, which stands for error, trend, seasonality, and it is an exponential smoothing method. This means that observations which are further back are weighted less than more recent observations by an exponential factor. I aggregated the data and took the mean of each month. Then I made various ets models with different settings for error, trend and seasonality. For example, with ZZZ and MMM which means automatically selected and multiplicative respectively (the three letters refer to the error, trend, and seasonality modes). After that, I calculated various measures of accuracy using the accuracy function on each of the models' forecasts and compared them to the test data.

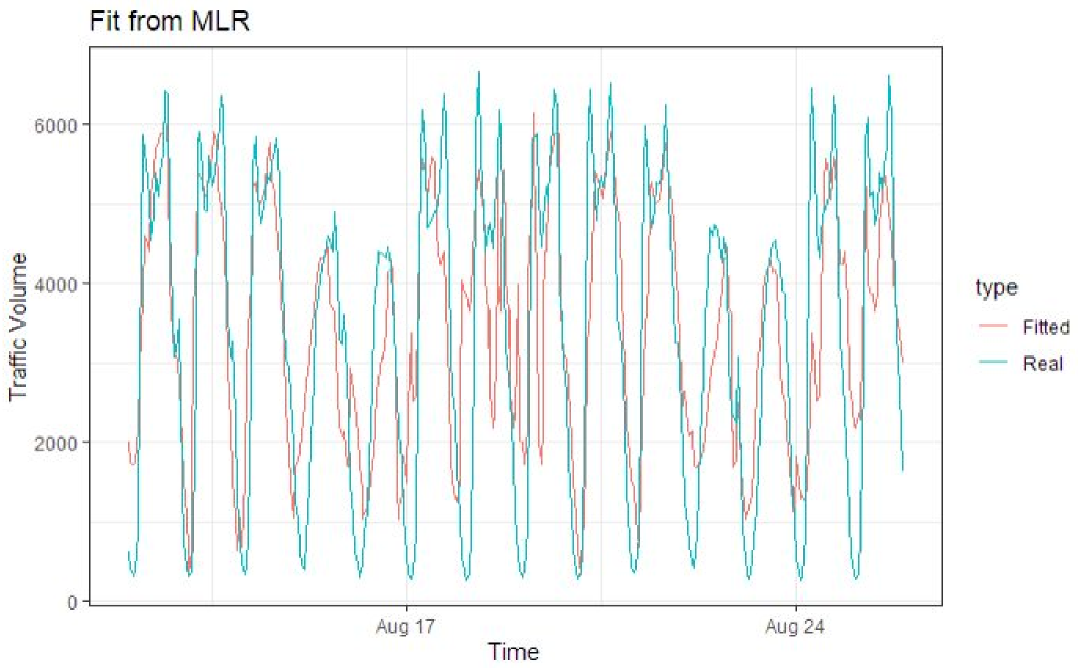
The next model I made was arima and another type of ets model which tried to forecast weekly activity rather than yearly, and made both of these by first making a time series object, plugging the time series into a seasonal decomposition of time (stl) function, and then finally putting it into the arima function. STL is a function that tries to decompose a time series into a trend, seasonality, and the remainder which is considered to be random variation that is not due to either aforementioned layers. A trend is how the mean of the data moves as time goes on. A seasonality is a repeating pattern which appears periodically. For the ets I did the same thing to it. Then I trained the model on two weeks of data to try to predict the following week. I assessed these models by their MAPE.

Results

The linear regression model with just all the time parameters came up with this.

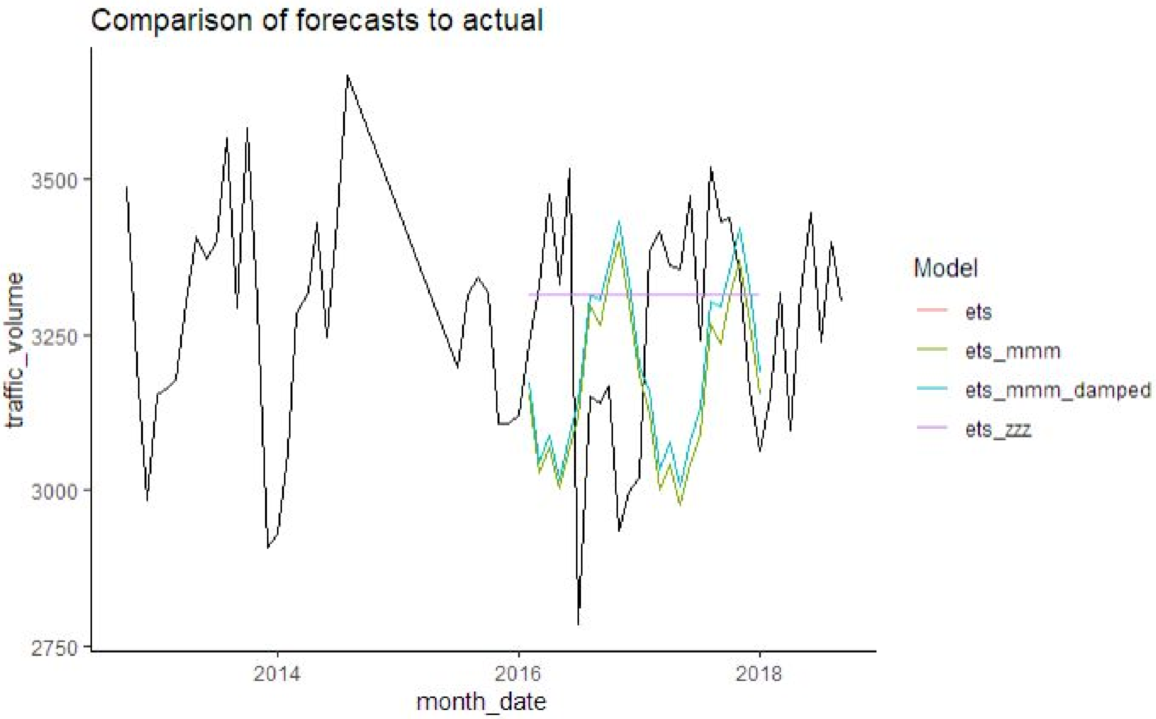


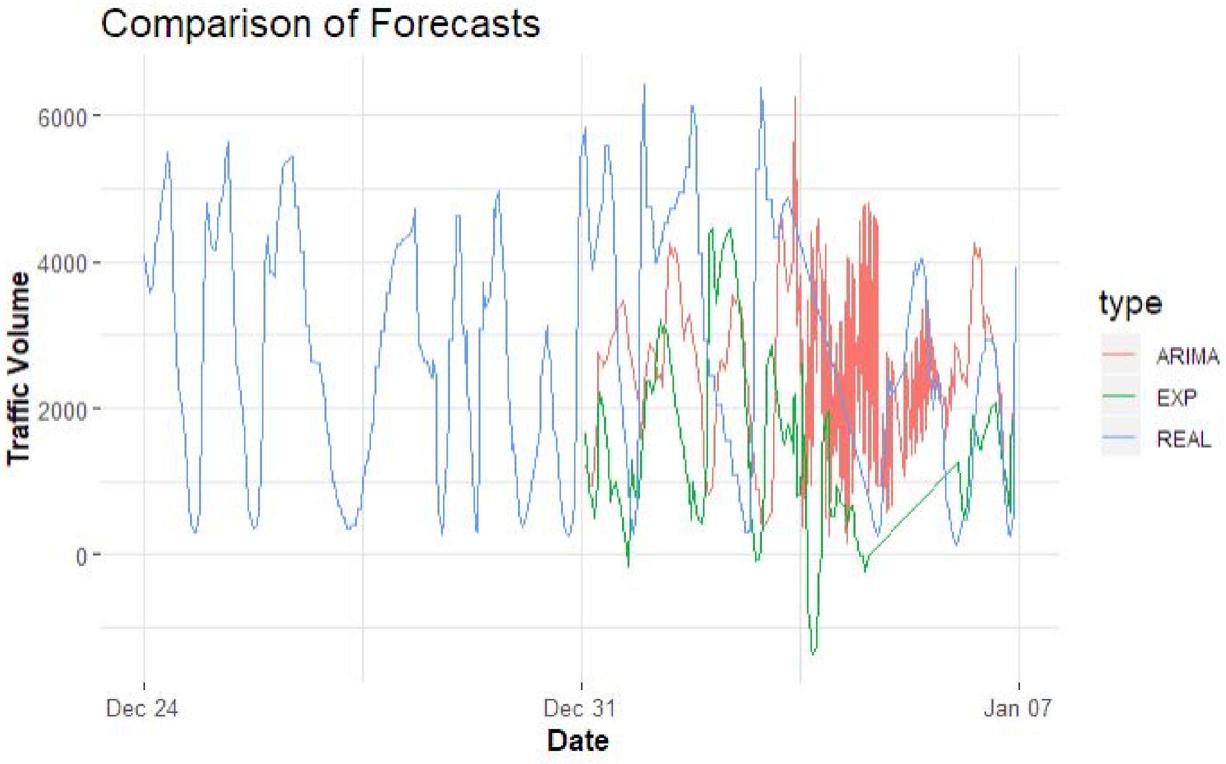
It was not very accurate, and looking at the residuals there was a pattern underneath that was not explained by this model. It was very simplified and was not entirely off as far as going up or down but it didn't go up or down with the right magnitude.



Surprisingly was able to get a decent fit on the data with a multiple linear regression so that is cool. It is nowhere near good but I think I could probably make it better in the future. It seemed to understand when to go up and down, and it has the right amount of days, as well as lowering over the weekend so I think it captured that aspect quite well. It also was slightly able to show some daily effect like peaking around morning and evening rush hour.

The ets model for the monthly averages did not perform well because it assumed too much of a constant pattern even though I had one that did automatic selection. so I think I would choose something more flexible or modify it so that it is less rigid. However this could also just be difficult because the average had an inconsistent pattern and the ets I chose which was multiplicative and automatic did not suit this process very well. The arima and stl composition as well as the ets and stl composition had varying success but it was mostly very off. Sometimes one would perform better than the others depending on where in the data, but overall the stl ets model performed better I think across different training and testing sets. It was weird because it went negative sometimes and I’m not really sure why it did that because the data was never negative. The arima model was also quite wonky and I think it is just because I didn’t adapt the code well to my data.





Discussion

I think that modeling this data was difficult because there was a lot of stuff in it and a lot of data to parse. During this project I had multiple issues, such as not knowing how to code, and also it is hard to understand, but I tried kind of. I would say that I was not very successful at all, which is kind of disappointing, but it’s okay because I feel like working on this project was slightly enriching. If I could do the project again, I would do more initial data exploration to understand the data more for the purposes of selecting a proper time series option. I did not approach the project in a very organized way and looking back I would have planned more. For example, I tried to do everything at the same time but it would have been much better if I did more conceptual research first rather than just plugging and not understanding what I am doing.

If I had more time on the project, I would probably invest more into both of the methods I tried already and see if I could make them better. I thought that the results of the multiple linear regression were interesting. I would make a model for predicting daily, weekly, monthly, and yearly behavior. But maybe that isn’t necessary. I think that the arima also has potential too if I tuned it better. It was also hard to consider the effects of the daily trend, weekly trend, and yearly trend on modeling the data. I think I would have used more of the zoo package to deal with the hourly trends because I think arima is more suitable towards less frequent observations. I would like to fix it more so it actually worked because I think it has a high chance of working but I could not figure out coding that well yet, also I did not really understand the math properly to work with it. However, I do think that I could perform some cv or other thing. But we did have a boatload of data so maybe it is not as important (?).

Forecasting is quite an interesting topic and there seems to be one dude who does a lot on it (Rob Hyndman), so I think I would like to read his textbook on my own. I am also interested in using time series further in economic contexts.

If I could pick another model to make this data I think I would try a neural network to deal with the various time dimensions and also include categorical variables.

References and Sources

Forecasting: Principles and Practice
Rob Hyndman-George Athanasopoulos - Otexts - 2018

Forecast Double Seasonal Time Series with Multiple Linear Regression in R – Peter Laurinec – Time Series Data Mining in R. Bratislava, Slovakia. https://petolau.github.io/Forecast-double-seasonal-time-series-with-multiple-linear-regression-in-R/

Similar Day Approach (TS)
https://petolau.github.io/Forecast-electricity-consumption-with-similar-day-approach-in-R/

Time Series Analysis

https://ourcodingclub.github.io/2017/04/26/time.html#format

https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/