PUI Extra Credit Paper: Analyzing NYC Traffic Datasets with Time Series

Po-Yang Kang[1]

[1]Affiliation not available

December 14, 2018

Abstract

This paper analyzes the data set of two NYC Open Data sets: One that is stated is gathered from 2011 to 2012; the other from 2012 to 2013, and both are data collected by Open Data on New York City Traffic. This paper has two goals to test: One is to test the null hypothesis that the datasets of these two time period distributions are the same. The other goal is to give a time series analysis of these two datasets with analysis of seasonality and event affects. The result is that the null hypothesis that these two time period distributions are the same has been rejected and that there has been several seasonality and event affects in the distributions which detract from the autoregression predictions.

Introduction

The interest with looking at Open Data traffic datasets came from my previous analysis of one specific Open Data traffic dataset where Average Yearly Traffic data was gathered for years from 1977 to 2015 in a different project. The process of data gathering for this is interesting, as the data gathered in 2011 and 2012, in comparison to 2013 and 2015, has only traffic for a couple of roads whereas in 2013-2015 it is for most of New York. For this reason, I am especially interested in datasets that helped determine the Average Yearly Traffic in 2011 and 2012, and therefore chose two datasets from NYC Open, the 2011-2012 and 2012-2013 dataset, to analyze their distributions and time series, in order to see if the limitations in the Average Yearly Traffic data has been the result of the source data as well.

Unlike the Average Yearly Traffic data, the data from the 2011 and 2012 datasets contain daily traffic instead, which might be interesting to see the effects of seasonality throughout the year and how traffic changes by day or even season.

Data

The datasets include two. One is stated from 2011 to 2012. The Other from 2012 to 2013. The original variables on both include the Segment ID or GIS ID (the street ID), the roadway name, where it's from, where it's going to, the direction, the date, and the hourly traffic. The dataset treats the hourly traffic as its own variable (so one variable is for traffic from 12:00-1:00 PM for example). Originally, I plan to group the data based on what street it is on. So, I wrangled it to create an average variable, that is the average of all traffic on that day on that street, and I then plotted all streets' averages. However, this caused an issue, as judging from the graph, not all traffic has been recorded equally with each street, and so a lot of data has been missing or just not recorded. Not only that, but for the 2011-2012 dataset, it only has the records from January 2012 to end of February 2012. Meanwhile, for the 2012-2013 dataset, it has records from October 2012, and from January to March 2013. Also, different traffic data on the same road has been

graphed twice on the same day, which creates an awkward relationship with the date. The 2011-2012 graph can be seen as figure 1 with a log10 relation with time.

So, to minimize the problems that differing data lengths for each road might cause, I instead averaged them based on the date. All analyses are therefore based on the relationship between datetime and the average of all roads on that day.

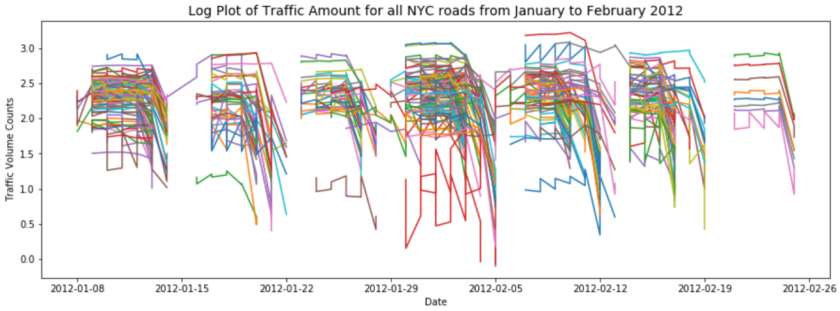

Figure 1: Log Plot of the 2011-2012 Dataset.

Methodology

The first goal is to find out whether the distributions of these two datasets are similar.

Null Hypothesis: The distribution of traffic in the 2011-2012 dataset is the same or like the distribution of traffic in the 2012-2013 dataset.

F 2011-2012 = F 2012-2013

Alternative Hypothesis: The distribution of traffic in the 2011-2012 dataset is not the same or like the distribution of traffic in the 2012-2013 dataset.

F 2011-2012 [?] F2012-2013

testing with significance level

$\alpha = 0.05$

Because two samples are compared through their distributions, a K-S test is appropriate here, as a K-S test checks whether the two samples come from the same distribution. If the K-S statistic is small or the p-value is high, then we cannot reject the hypothesis that the distributions of the two samples are the same. However, the K-S statistic is 0.8 with a p-value of 2.4481975790778211e-17, so therefore, since the p-value is extremely low, and the statistic is high, we reject the hypothesis that the distributions of the samples are the same.

The distribution can also be judged from the time series plots as well.

The second goal is to use time series analysis to check for seasonality, event detection, and to forecast the pattern using Autoregression.

Autoregression model has been used in order to forecast based on the current patterns in the model. It has been used in the short run to predict traffic not just on roads but also for networking and broadband connections. If the patterns drawn by autoregressions are not like the last thirty days to the time series, there is a concern on whether forecasting using this data may be accurate.

Figures 2 includes the normalized graphs for the 2011-2012 and 2012-2013 datasets. In the 2011-2012 dataset there is a bit of "seasonality" going on, in a sense that the traffic dips down almost once per week, but minimizing seasonality affects here may not guarantee the data to be accurate, as the dips do not happen every seven days but sometimes every six or eight days. Graphing by minimizing differences between weeks would therefore give unneeded peaks and troughs in the data.

In the 2012-2013 datasets, unlike the 2011-2012 dataset which follows a pattern, the 2012-2013 one does not, it instead has several events, with the highest ones peaking at February 8th and March 21st. There is no data collected for November and December 2012, so a large chunk of data has been missing in the records for 2012-2013.

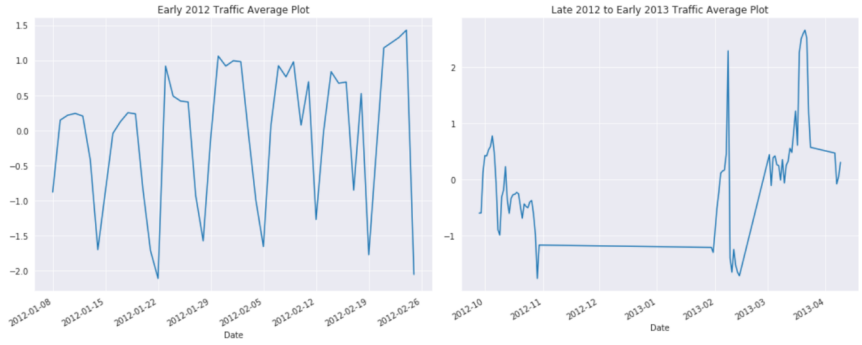

Figure 2: Normalized Plots of the Traffic Averages

Figure 3 include lag plots. Lag plotting is a way to check for autocorrelation in a time series data set. Random data should not exhibit any structure in the lag plot. Non-random structure implies that the underlying data are not random. For the 2011-2012 dataset, the larger the traffic, the more it falls into an observable linear pattern. In the 2012-2013 dataset, the smaller the traffic values are the more it falls into a linear pattern.

Figure 4 shows the autoregressive models of Early 2012 Traffic Averages and Late 2012-Early 2013 Traffic Averages over 30 days. The AR predictor predicts the overall trend of where the data is going. However, at some areas the predictions do not entirely match up to the data: While the test data from Early 2012's last 30 days seem to be flat, the AR prediction seems to be much lower around the beginning and varies tremendously over time. Early 2013, meanwhile, has varying test data from its last 30 days that doesn't match with its typical pattern judging from its AR predictions. But regardless, what seems to be interesting

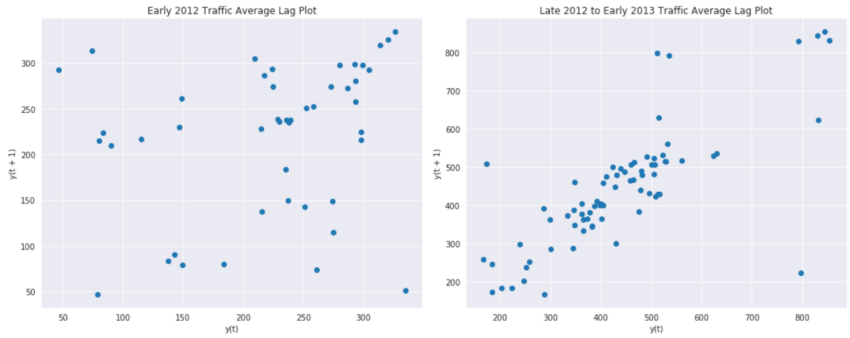

Figure 3: Lag Plots of Traffic Averages

is that the AR prediction patterns seem to be both cyclical, and that the dataset may lead in to each other if more data has been included for the time period described.

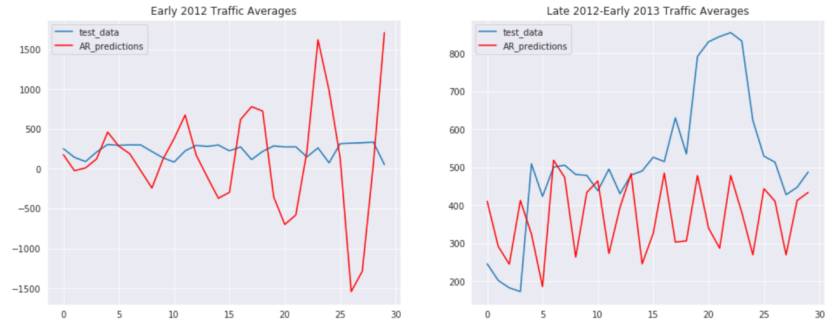

Figure 4: AR plots of Traffic Averages

Conclusion

Originally, these two datasets have been analyzed to see whether the limitations in the Average Yearly Traffic dataset is reflective with the source data as well. What I found out was although the two datasets analyzed do lack data for various roads, they also lack data for traffic across consistent days as well, which ends up the distributions of the two sets being unequal and result in several major peaks and troughs. However, through Autoregression, both sets exhibit a certain cyclical pattern that are like each other and regardless exhibits an accuracy appropriate for predicting and determining traffic in New York City in 2012 and 2013, and therefore the pattern might be found in later and previous years as well.

Future Works

In the future, more data should be gathered over time. Not just for all roads, but for all days as well. Other data that might be interesting to further the analysis is to check what caused these large events in data: The repeated dips during the weekends can be attributed to less rush hour, but other factors can and should be considered as well, especially since sometimes it's either Saturday or Sunday that has low traffic and not both. This can be concerning, as getting rid of seasonality through week differences would not be possible if the recording of data is not consistent on certain days. Also, with more data, perhaps a time series analysis should be run across several years as well.

Bibliography

Adas, Abdelnaser: https://ieeexplore.ieee.org/abstract/document/601746

Brownlee, Jason: https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

Pavlyuk, Dmitry: https://www.sciencedirect.com/science/article/pii/S1877705817300620

Open Data 2011-2012 Traffic Counts: https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2011-2012-/wng2-85mv

Open Data 2012-2013 Traffic Counts: https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2012-2013-/p424-amsu

Po-Yang Kang's Github Link