Waste Tonnage Analysis: ARIMA Modeling for NYC Waste Tonnage

Daniel Lupercio

Professor Chun Pan

STAT 790 – Case Seminar

City University of New York – Hunter College

Abstract

This paper deals with analyzing solid waste tonnage numbers collected in NYC within 2005 – 2020. The original data set includes tonnage numbers on Refuse, Paper, metal, glass, and plastic (MGP), residential organics, etc.… This project takes a time series approach to analyze and forecast tonnage values at the borough level for all five boroughs in NYC. The three waste streams were totaled to create one single waste stream. In an attempt at a dynamic regression model at the NYC level, a single waste stream variable was regressed onto external variables.

Keywords – refuse, paper, metal, glass, and plastic (MGP), ARIMA, acf, pacf.

1. Introduction

The NYC Department of Sanitation (DSNY) is the world's largest sanitation department. The DSNY collects more than 10,500 tons of residential and institutional garbage and 1,760 tons of the recyclables – each day. In 2019, approximately 3.25 million tons of refuse was disposed from residential and institutional buildings. Along with approximately 682,000 tons of recycling material (mgp, paper, organics). During the winter, the DSNY is responsible for clearing snow and ice from New York City's more than 19,000 lane-miles of roadways in a prompt and reliable manner.

This dataset used in this project can be found on the "NYC Open Data" website [**Error! Reference source not found.**]. It is updated on the last day of every month by the DSNY. The data includes tonnage numbers for the variables written above, for each of the 59 community districts in New York City. The Bronx has 12 community districts, Brooklyn has 18 districts, Queens has 12 districts, Staten Island has 3 districts and Manhattan has 12 districts. This public data has been tracked for the last 3 decades. With refuse tonnage beginning to be reported in 1991, and the recycling streams beginning to be reported in 1993.

In NYC, not all properties are serviced the same. For example, an apartment complex in MN01 (Manhattan Community District 1) will most likely have containerized service. DSNY personnel would drive a specific vehicle that can collect the container and drive it to the district's designated waste dump. This is very different compared to a residential house in QN07 (Queens Community District 07), where DSNY personnel collect garbage that is placed on the curb and load the garbage onto a collection truck. Most private businesses are not serviced by the DSNY, so data regarding their garbage tonnage is not included in the data set.

2. Background

There have been numerous studies that attempted to predict waste generation using machine learning techniques with waste data from NYC. Similar techniques have been used with other countries. C.E. Kontokosta et al. used daily waste collection data to apply gradient boosting regression trees and neural network models to estimate daily and weekly refuse and recycling tonnages for each of the more than 750,000 residential properties in the NYC. Similarly, N.E. Johnson et al used DSNY data, in conjunction with other datasets related to New York City to forecast municipal solid waste generation across the city to produce a gradient boosting model for short-term waste prediction. J. Navarro-Esbrí et al. used a prediction technique based on non-linear dynamics, to compare its performance with a seasonal AutoRegressive and Moving Average (sARIMA) methodology, dealing with short and medium term forecasting for two cities in Spain and one city in Greece.

3. Methods

This study aims to analyze and forecast monthly municipal waste generation at the borough level using a time-series and dynamic regression approach. The refuse, mgp and paper tonnage were summed to represent a “total_waste” variable. The total waste was then calculated for every borough, and month in the dataset. This allowed for us to have 192 time points of data, beginning with January of 2005, and ending through December 2020.

3.1 ARIMA Modeling

A stationary time series is one whose statistical properties do not depend on the time at which the series is observed. Those statistical properties are the mean, variance and auto-correlation, and these properties should be constant over time. In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behavior is possible).

The DSNY tonnage values exhibit seasonal behavior and/or cycles. To model and forecast the time series with low order ARMA arguments, it is convenient to eliminate this behavior. To identify possible seasonal and/or cyclical components present in the time series, the autocorrelation function (ACF) is used. Alongside that, a guide written by Professor Robert Nau of Duke University was used to help analyze the plots of partial autocorrelation function (PACF) of the time series. This guide allowed us to identify the proper number of ARMA arguments needed to return low statistical metrics.

3.1.1 Bronx Total Tonnage

We begin by isolating the Bronx tonnage values and plotting the time series.

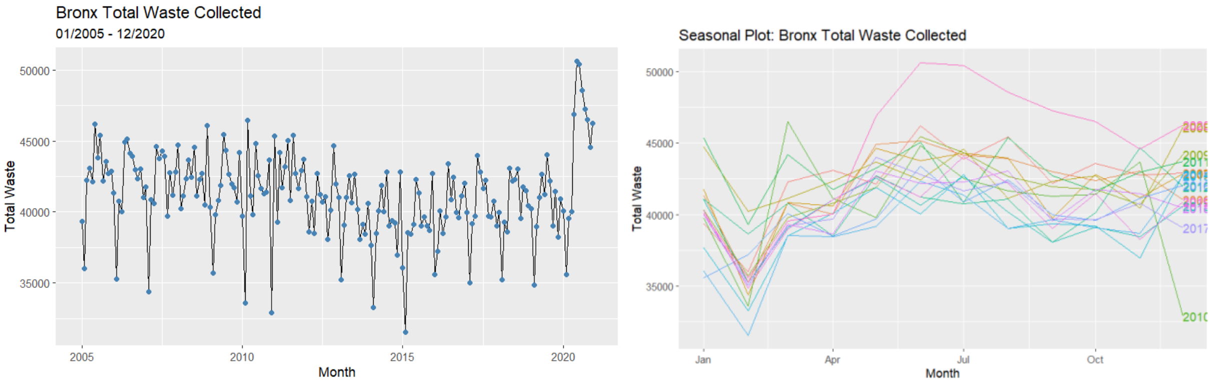

Figure 1 & 2: Bronx time series, and seasonal plot over time.

At a first glance, we see a seasonal trend in the first plot. Every calendar year begins with a low amount of waste tonnage collected, with an even lower total waste tonnage collected in the February months. The waste tonnage then ramps up during the summer months.

To get a sense as whether we can use the total waste values for modelling, we will use the KPSS test for stationarity. The H_0= The time series is trend stationary, vs H_a= The time series is not trend stationary. If the p-value of the test is less than some significance level (e.g. $\alpha = 0.05$), then we reject the null hypothesis and conclude that the time series is not trend stationary. For the Bronx, we performed a KPSS test both on the total waste and differenced time series. Both tests returned a p-value greater than 0.5, indicating that we fail to reject the null-hypothesis. We assume both time series are trend stationary.

Following a sequence of models, along with the guide from Professor Nau, we have found two Arima models that can approximate the Bronx time series well enough. Using the total waste time series, we believe that a seasonal Arima model, $ARIMA(0,0,4)(1,0,0)[12]$, approximates the time series the best. This model returns a RMSE = 2304.198.

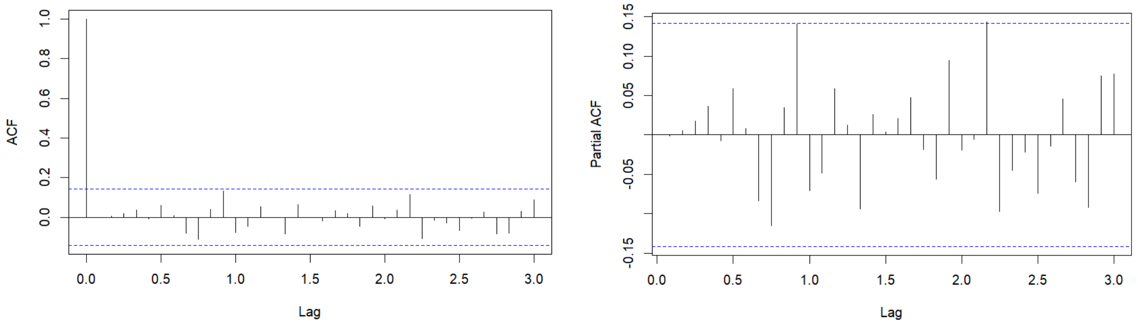

Figure 3 & 4: ACF and PACF plots of $ARIMA(0,0,4)(1,0,0)[12]$.

Similarly, we are also able to find a seasonal Arima model using the differenced values of the total waste time series. Using the differenced time series, we believe that $ARIMA(0,1,1)(1,0,0)[12]$ approximates the seasonality of the time series the best. This model returns a RMSE = 2341.145.

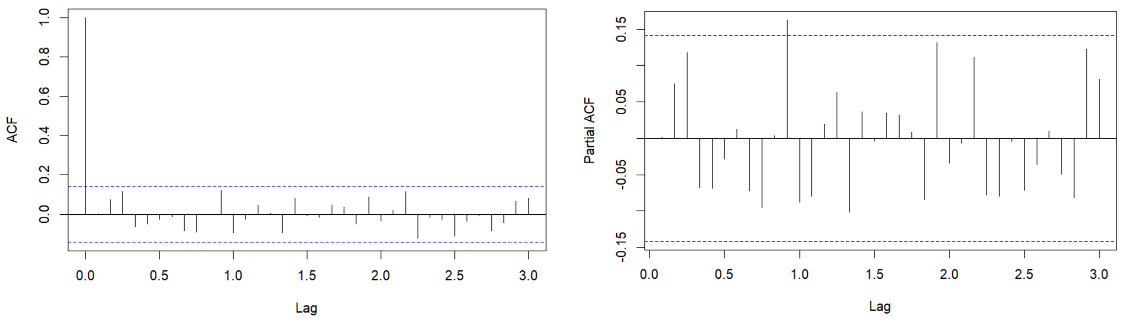

Figure: 5 & 6: ACF and PACF plots of $ARIMA(0,1,1)(1,0,0)[12]$.

For our final Arima model, we look and compare the auto-Arima results from the fable package with the previous models that were selected. The function returns $ARIMA(0,0,3)$ with mean, as the best model that approximates the total waste time series. This model also returns a RMSE = 2688.7.

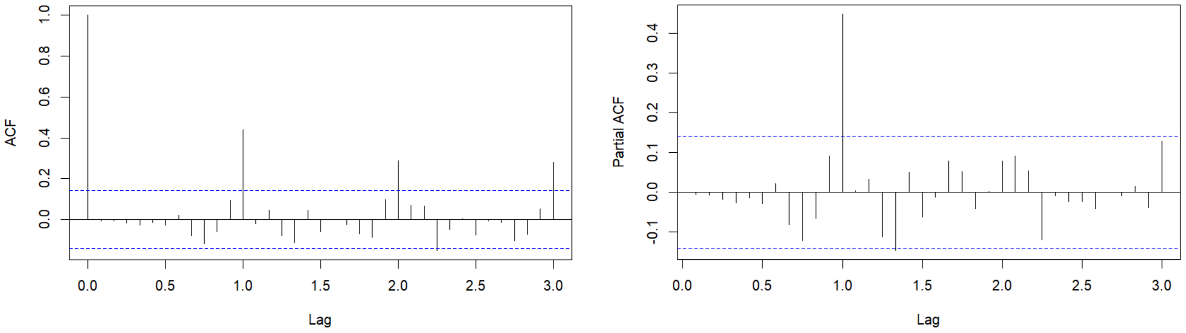

Figure 7 & 8: ACF and PACF plots of $ARIMA(0,0,3)$.

Looking at these plots further, we see that the function does not use the differenced series. We do see that the first 11 lags have little autocorrelation and are not significant. The seasonal lags are not addressed in the model, which is why we see lags = (12, 24, 36) positively auto correlated and significant in the ACF plot.

3.1.1 Manhattan Total Tonnage

We begin with two preliminary plots.

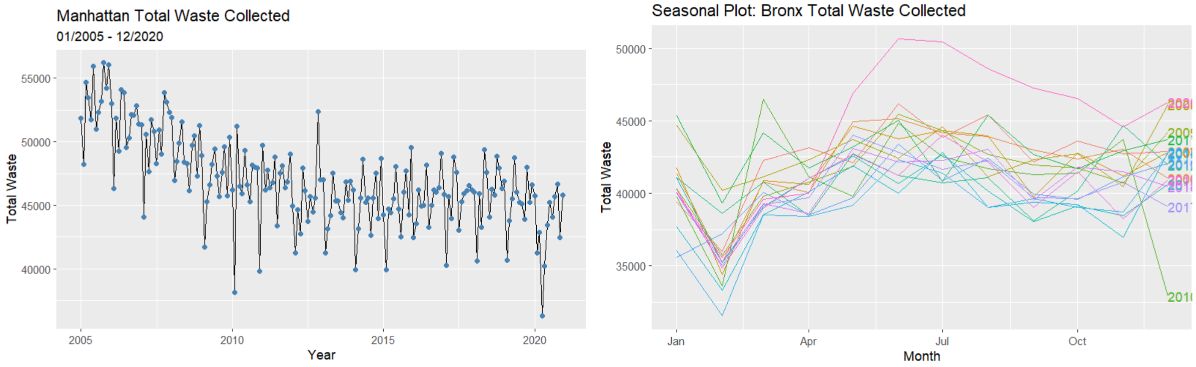

Figure 9 & 10: Total waste plot and seasonal plot of Manhattan.

When compared to the Bronx, more tonnage is collected in Manhattan overall. We see a pattern where the tonnage values decrease in February each year. When performing the KPSS test on the total waste values, we are returned a p-value = 0.01. We reject the null hypothesis that this series is trend stationary. Performing the test on the differenced values, we are returned a p-value = 0.1, which fails to reject the null hypothesis.

Following along with the guide from Professor Nau, we have found two Arima models that can approximate the Manhattan time series well enough. Using the differenced time series, we believe that $ARIMA(0,1,2)(1,0,0)[12]$ can be a suitable model that captures the Manhattan tonnage values. This model returns a RMSE = 2215.12.

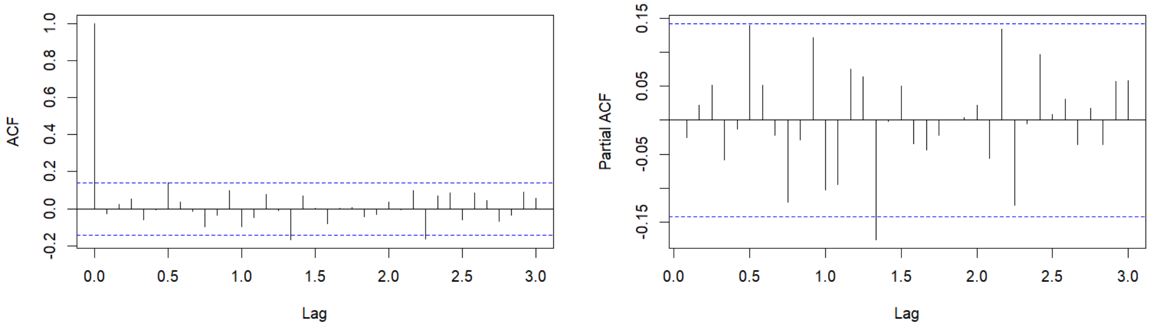

Figure 11 & 12: ACF and PACF plots of $ARIMA(0,1,2)(1,0,0)[12]$.

From the previous figure, the majority of the ACF and PACF lag values are contained within bounds. However, lag 16 in both plots are significant. There wouldn't be a direct way to address this lag without adding a high MA argument and potentially overfitting this model. In the PACF plot, the autocorrelation values are bounded between (-0.15, 0.15).

The second model that can approximate the tonnage values the best also uses the differenced time series. The model is an $ARIMA(0,1,4)(1,0,0)[12]$ returns a RMSE = 2208.185.

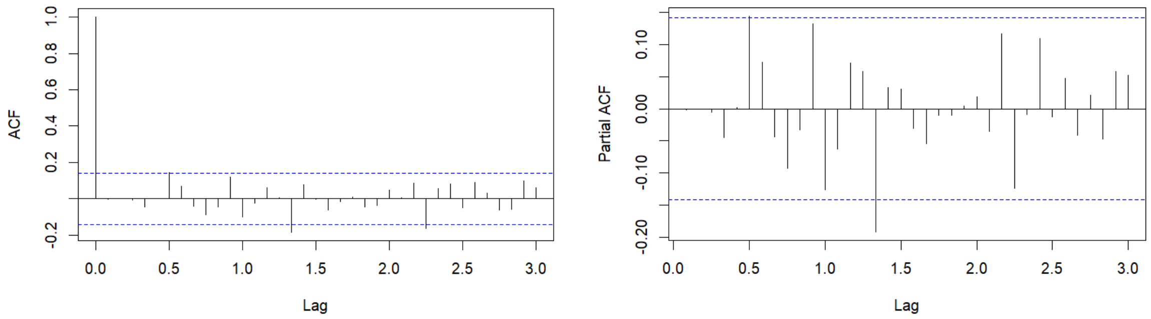

Figure 13 & 14: ACF and PACF plots of $ARIMA(0,1,4)(1,0,0)[12]$.

The RMSE has decreased, when compared to the previous Arima model. Again, most of the lags are contained within the significant bounds. However, we are not able to make the lag 16 insignificant in both plots.

The final model that will be compared is the auto-Arima model. The fable() package returns a $ARIMA(0,1,5)$ model, with RMSE = 2449.647.

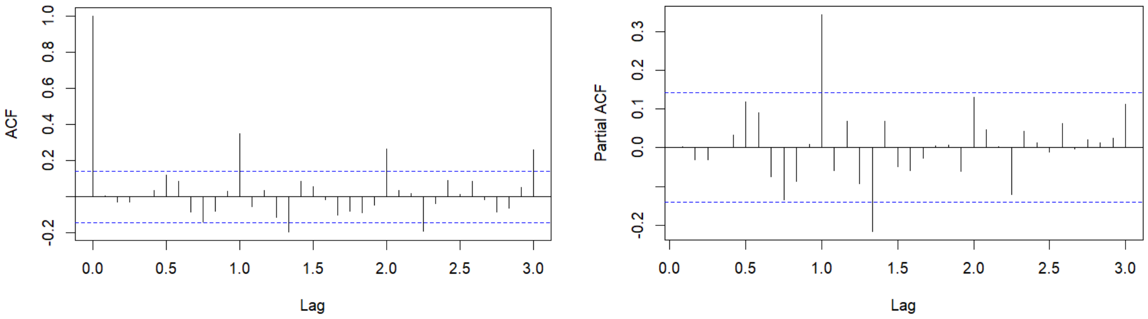

Figure 15 & 16: ACF and PACF plots of $ARIMA(0,1,5)$.

This auto-Arima model, does not address the seasonal terms, and returns a RMSE greater than the previous two models. The high MA() argument can lead to overfitting, usually AR() and MA() arguments are best used with a maximum of three. This model does not use the differenced values, which can explain the high MA() argument.

3.1.2 Brooklyn Total Tonnage

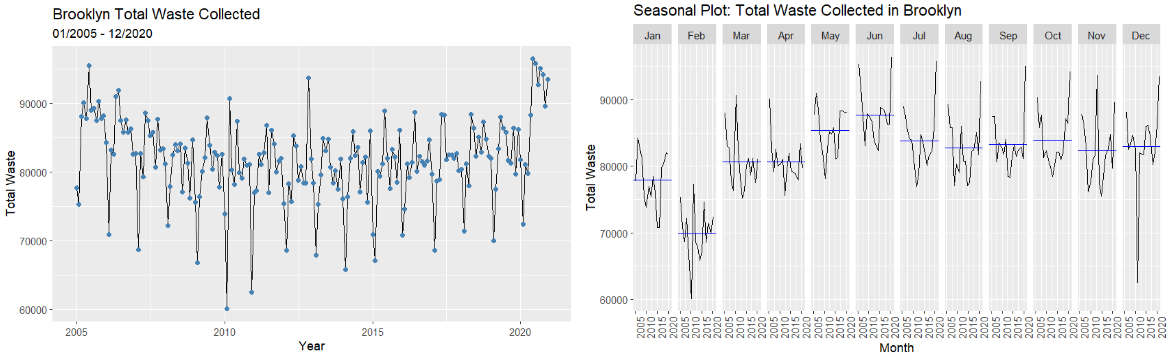

Figure 17 & 18: Total waste and subseries plots of Brooklyn refuse tonnage.

Most of the refuse tonnage values are bounded between (65000, 90000). From figure 18, we continue to see a decrease in tonnage values during the month of February for every year in the time series. When performing the KPSS test on the total waste values, the p-value of the test is greater than 0.05. We fail to reject the null-hypothesis and assume that total wase values are trend stationary. The same results hold for the differenced values.

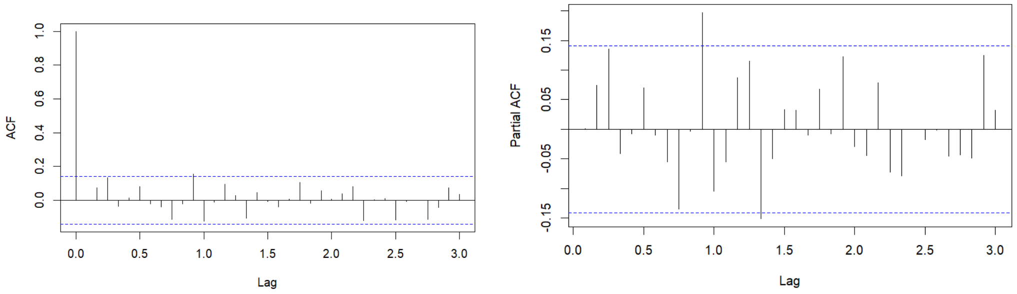

Figure 19 & 20: ACF and PACF of $ARIMA(0,1,2)(1,0,0)[12]$

After reviewing three Arima models, we believe that an $ARIMA(0,1,2)(1,0,0)[12]$ is the model that best captures the seasonality of the Brooklyn tonnage values. While returning good autocorrelation values both on the ACF and PACF plots. This model returns a RMSE = 4427.169.

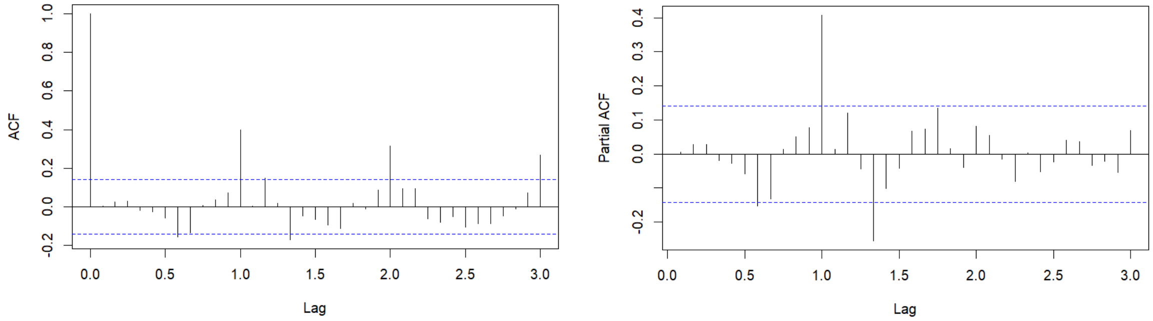

Figure 21 & 22: ACF and PACF of $ARIMA(3,0,3)$.

The auto-Arima model selected to be compared to our previous model is an $ARIMA(3,0,3)$. This model returns a RMSE = 5219. This model does not address any seasonality in the time series. It ACF plot, shows positive and significant autocorrelations at lags = 12, 24 and 36. A similar story for the PACF plot. The first 6 lags are not significant, but we do see a positively correlated and significant lag at lag = 12.

3.1.3 Queens Total Waste

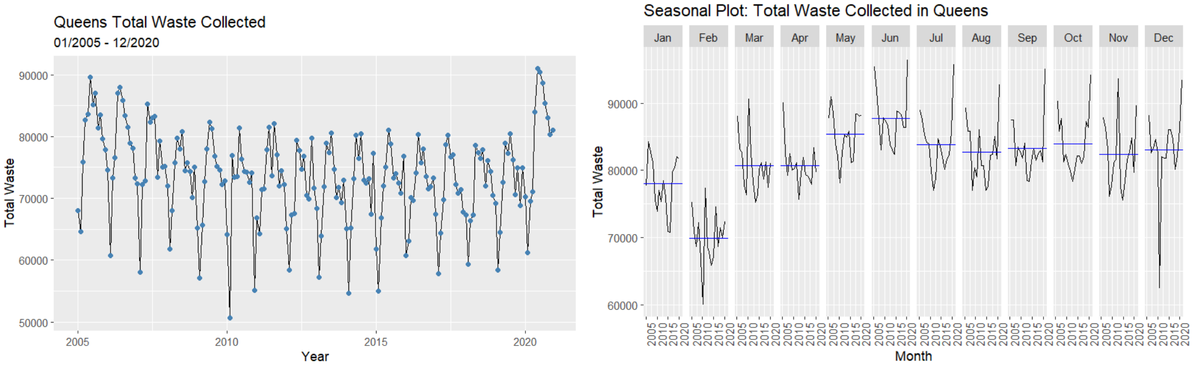

Figure 23 & 24: Queens total waste and subseries plot.

As with the previous boroughs, the KPSS test was performed both on the total waste time series and the differenced time series. In both tests, the p-value was greater than the $\alpha = 0.05$, which fails to reject the null hypothesis. We assume that both time series are trend stationary.

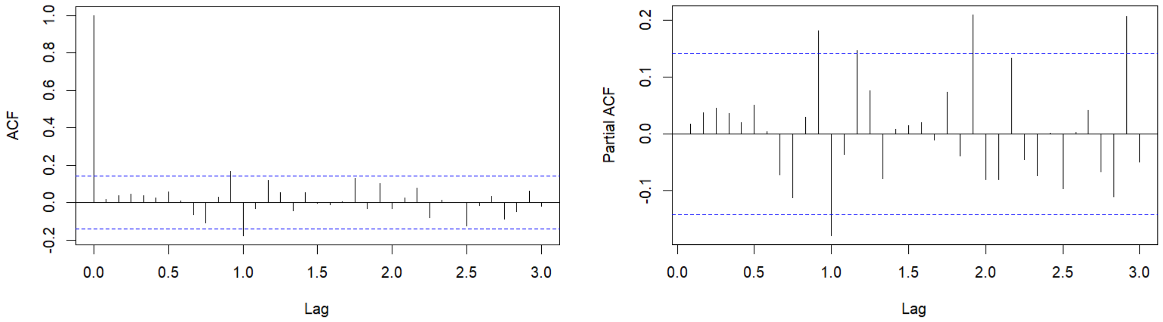

Figure 25 & 26: ACF and PACF of $ARIMA(3,1,1)(1,0,0)[12]$.

After a series of models evaluated on the differenced time series, we believe that an $ARIMA(3,1,1)(1,0,0)[12]$ can approximate the Queens total tonnage time series the best. This model returns a RMSE = 4111.87, and a AICc = 3766.23. The ACF plot displays a similar pattern of autocorrelated lags from previous Arima models of other boroughs. However, in the PACF plot, lags = (11, 23, 35) are all positively autocorrelated and significant. We also see that the first seasonal lag is negatively autocorrelated and significant. No further effort to remove the significant non-seasonal lags was made, as adding a higher order AR() argument will most likely lead to an over fitting model.

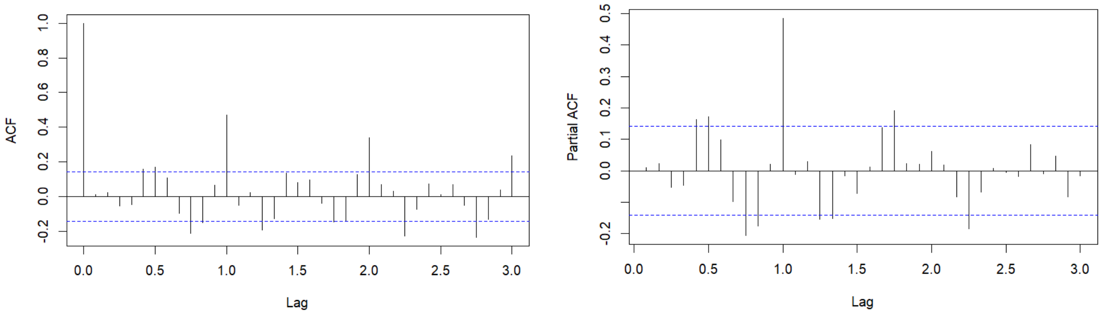

Figure 27 & 28: ACF and PACF plot of $ARIMA(3,0,2)$.

After reviewing two auto-Arima models, it was best to choose an $ARIMA(3,0,2)$ to report. This model returns RMSE = 5477.051, and an AICc = 3866.63. When looking at figure 27, there is a pattern of negative auto-correlated values that are significant. This model does not address the seasonality of the time series, which is why you see the positive lags that are significant in the same figure. This auto-Arima model returns the highest RMSE.

3.1.4 Staten Island tonnage values

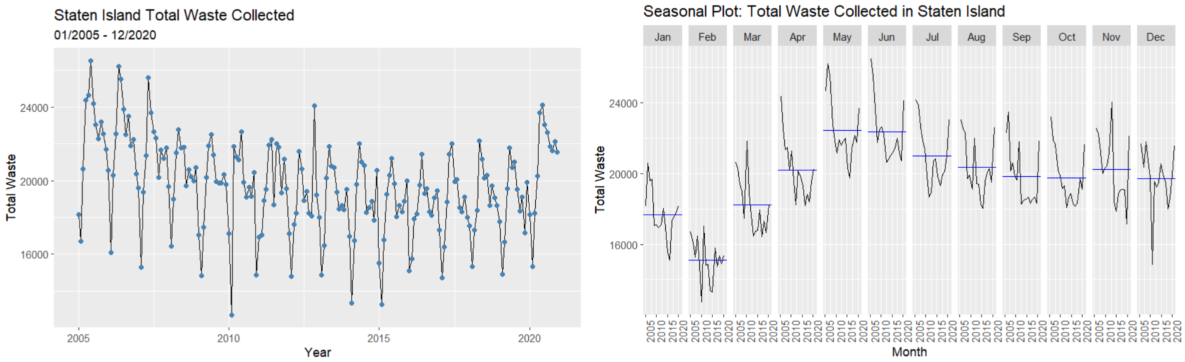

Figure 29 & 30: Staten Island total waste and subseries plot.

From figure 29, we see that most of the total tonnage is bounded between (15000, 26000). When performing the KPSS test on the total tonnage values, the results indicate that we can reject the null-hypothesis and assume this series is not trend stationary. When performing the test on the differenced values, the results show that we fail to reject the null hypothesis and we can assume the series is trend stationary.

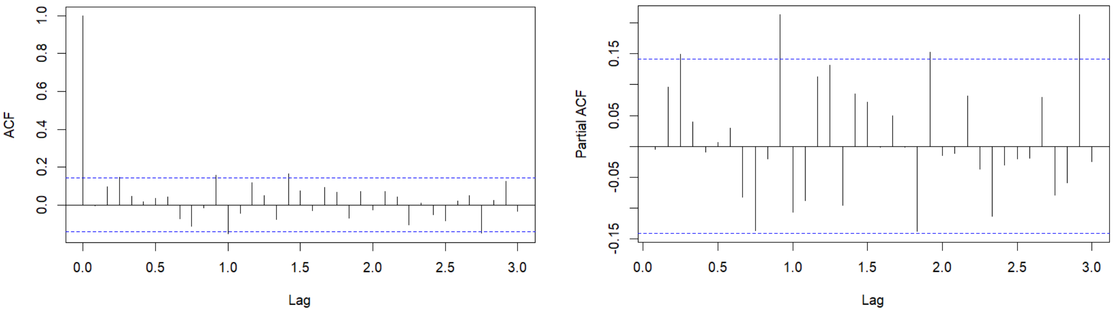

Figure 31 and 32: ACF and PACF plots of $ARIMA(2,1,3)(1,0,0)[12]$.

After researching numerous ARIMA models, we believe an $ARIMA(2,1,3)(1,0,0)[12]$ can best approximate and forecast the total tonnage values of Staten Island. This model has a RMSE = 1371, and AICc = 3331.18. All the lags in the ACF plot appear to be within bounds. Most of the lags in the PACF plot are within the bounds. But lags = (11,23,35) are significant, which is a pattern we have seen in previous models. Most of the lags are bounded between (-0.15, 0.20). Adding more parameters to make those lags insignificant would perhaps create an overfitted model.

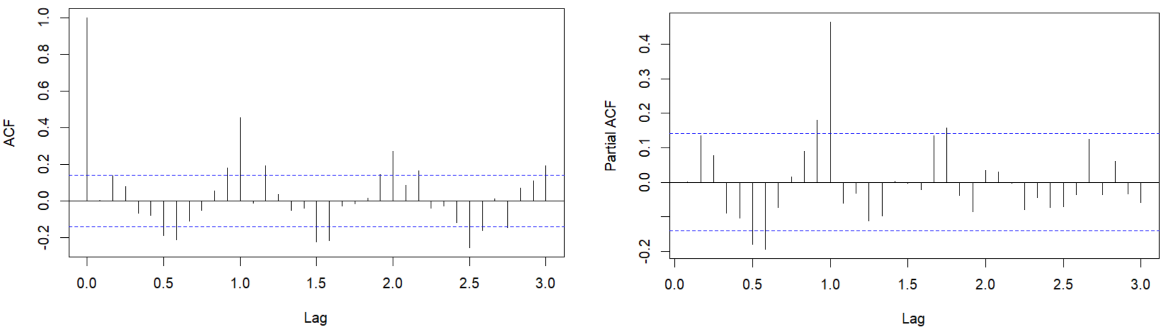

Figure 33 & 34: ACF and PACF plots of $ARIMA(2,1,4)$.

The auto-Arima model selected was an $ARIMA(2,1,4)$ with RMSE = 1746.309, and AICc = 3412.39. This model does not address the seasonality found in the time series. From the PACF() plot, we can see lags = (6, 7) also being significant and negatively auto-correlated. This would indicate that it would be best to add more MA() arguments. This would not be optimal and will lead to overfitting.

*3.2 **Linear Regression***

To use a different model to best approximate and forecast tonnage values, we will use multiple linear regression. For this method, we will not be producing linear models for each of the five boroughs. Instead, we will aggregate all the monthly tonnage values into one tonnage value per month. The tonnage value per month will represent the total tonnage collected in NYC, with a total of 192 time points.

With the help of the studies mentioned above, we were able to the consumer price index (cpi), unemployment rate, average precipitation, average temperature and average cooling degree days. The cpi is a measure of both New York and New Jersey, while the unemployment rate is a measure of New York City. The temperature, precipitation and cooling degree days are measured from Manhattan's Central Park.

The model that is used will regress the total tonnage values onto the five predictors. The TSLM function from the fable package allows us to fit a linear model with time series components.

$$\hat{y}_t = 257322.46 - 84.31x_1 + 48960.81x_2 + 38695.86x_3 + 1144.89x_4 - 1639.11x_5 + \epsilon_t$$

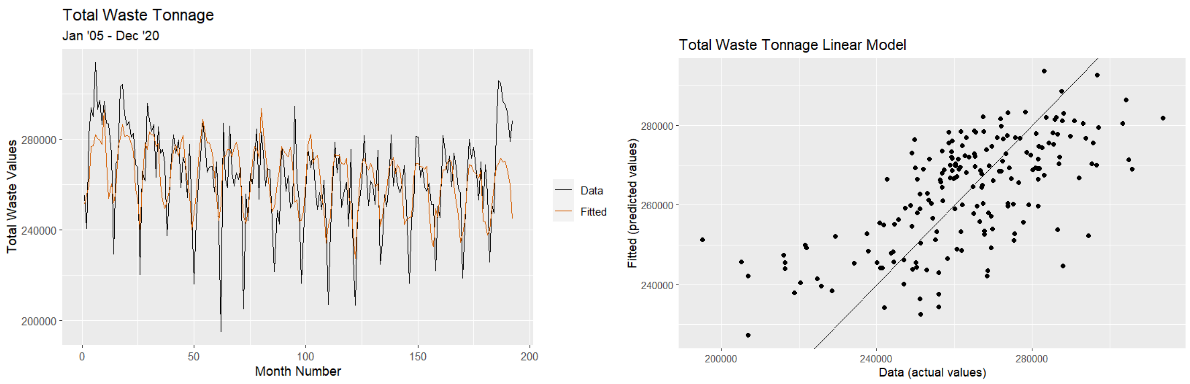

Figure 36 and 37: Fitted values on the total waste plot & fitted values plotted against the data.

From a glance, the fitted values do approximate the total waste tonnage well. Figure 37 does appear to show a linear trend when plotting the predicted values against the actual tonnage values in the series. We can look at the residuals of the predictors for further evaluation.

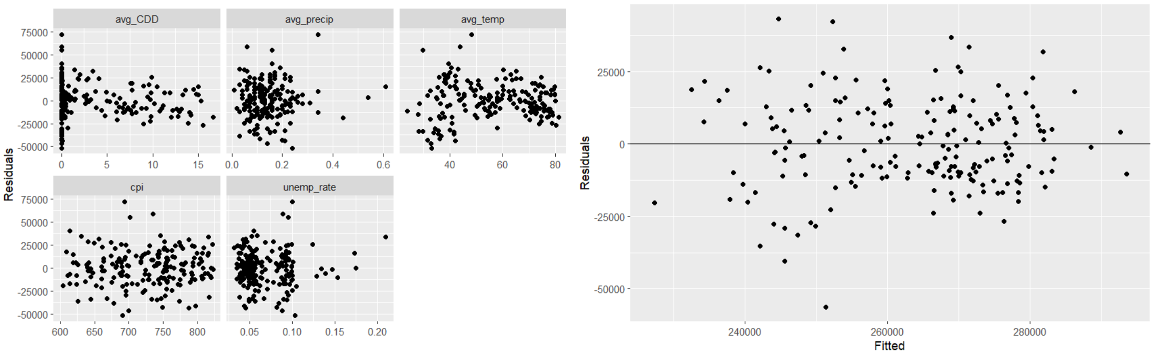

Figure 38 & 39: Residual plots against the predictors and fitted values.

From figure 38, the cpi and average temperature are the lone predictors where the residuals are randomly scattered. From figure 39, the model appears to show that the residuals are homoscedastic.

*3.3 **Dynamic Regression***

When using a dynamic regression model, we are allowing the errors from a regression model to contain autocorrelation. We will replace ε_t with η_t in the regression equation, so the error series η_t is assumed to follow an ARIMA model. These models will have two error terms - the error from the regression model, which we denote by η_t and the error from the ARIMA model, which we denote by ε_t. Only the ARIMA model errors are assumed to be white noise [insert reference].

We need to consider that all the variables in these models must first be stationary. It is common to difference all the variables if any of them need differencing. When performing the KPSS test on all the variables, the two variables that did not pass with a significant p-value were the total waste and CPI time series. Hence, we reject the null-hypothesis and assume that the time series' are not trend stationary.

After investigating the NYC total waste time series separately, and continuing to follow the modelling guide, the models that best approximate and forecast the time series is an $ARIMA(0,0,4)(1,0,0)[12]$ and an $ARIMA(0,1,4)(1,0,0)[12]$. Unfortunately, we can't investigate the $ARIMA(0,1,4)(1,0,0)[12]$ any further, as we are returned an error indicating that the design matrix is not invertible, and the dynamic regression model cannot be developed [insert stack exchange reference].

3.3.1 LM w/ ARIMA(0,0,4)(1,0,0)[12]

For a model including all five predictors and a trend parameter, we are returned a RMSE = 11339.34 and AICc = 4145.17.

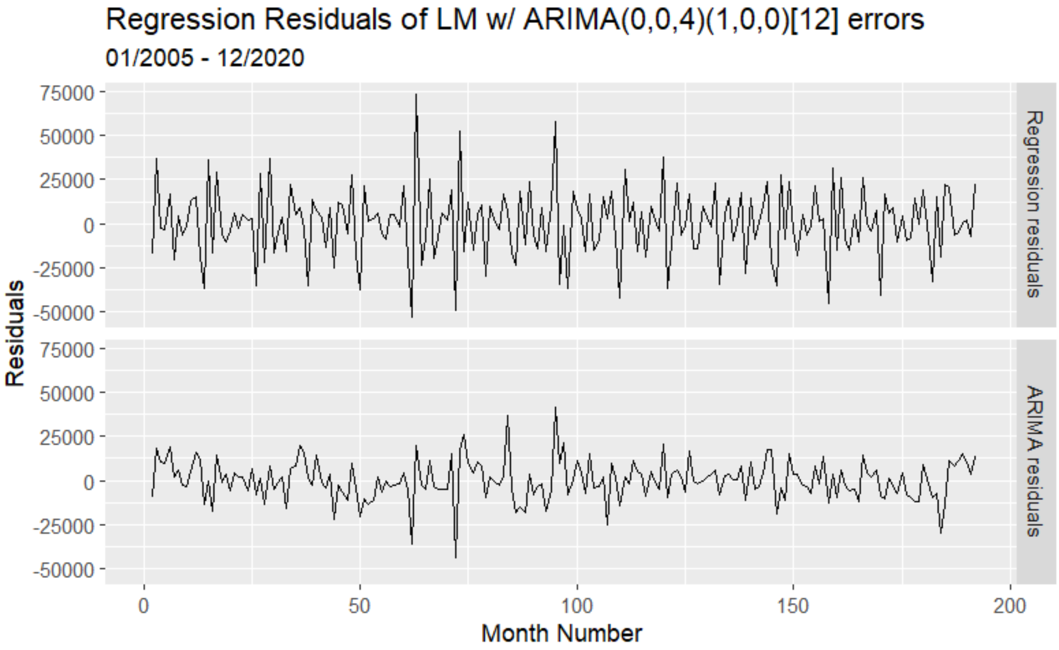

Figure 40: Regression residuals of this model.

3.3.2 LM w/ ARIMA(2,1,4) errors

The Arima errors in this model were found using auto-arima function and is used to compare the previous dynamic regression model. This model includes all five predictors, and no trend parameter. It has a RMSE = 13979.22 and AICc = 4208.78.

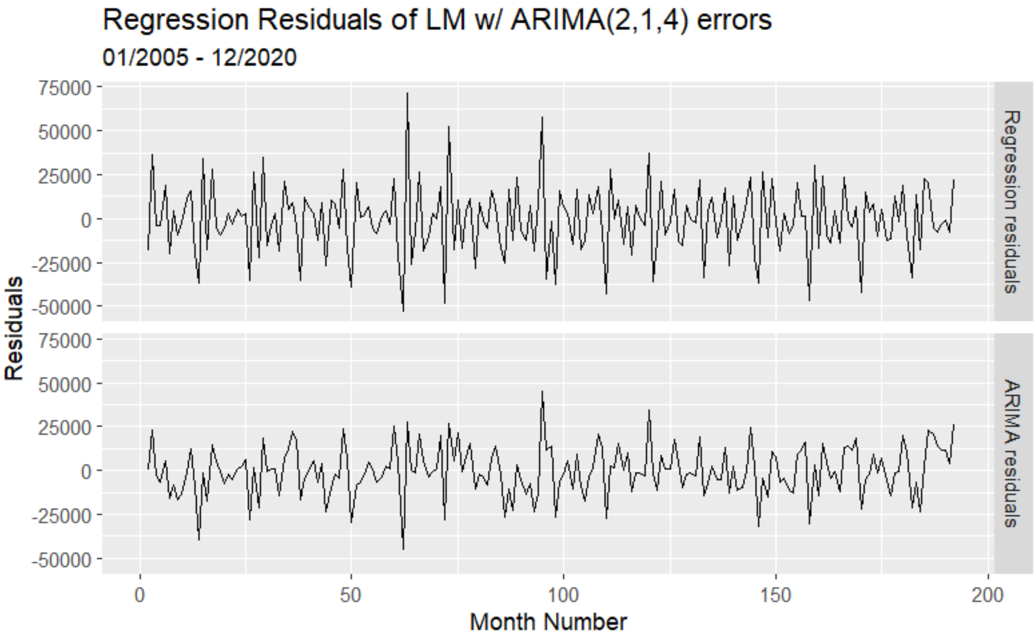

Figure 41: Regression residuals of this model.

4 Conclusion

To be written later.