

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 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 ([esri](#)).

This dataset used in this project can be found on the NYC Open Data website. 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. [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 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. 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 Waste

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

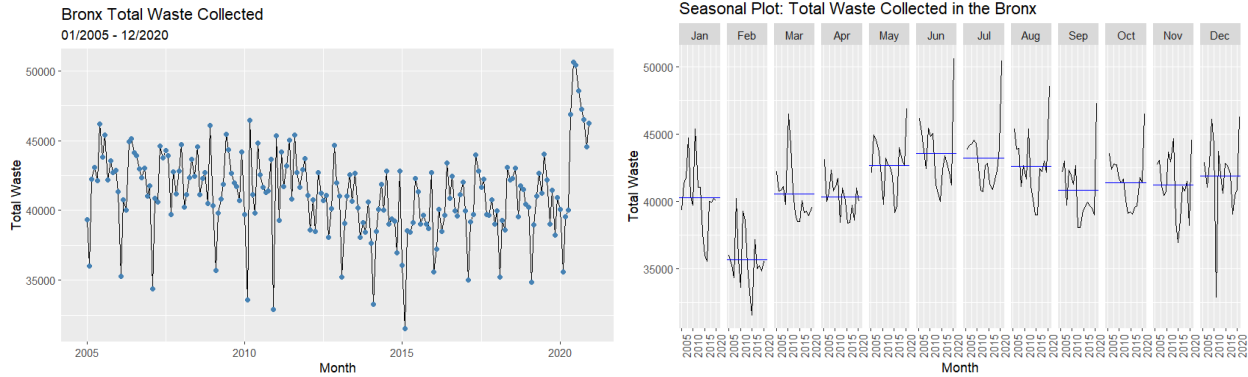


Figure 1 & 2: Bronx time series, and seasonal plot over time.

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, with a 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 assume 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 0.05, 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 $RMSE = 2304.198$, and $AICc = 3537.29$.

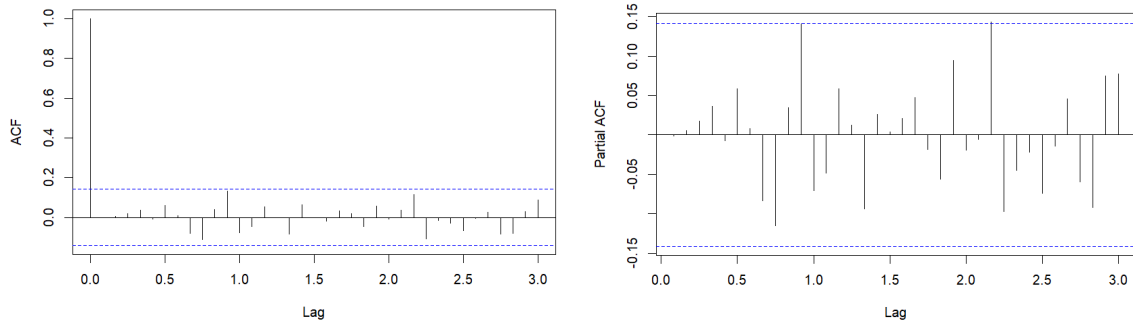


Figure 3 & 4: ACF and PACF plots of $ARIMA(0,0,4)(1,0,0)[12]$.

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 $RMSE = 2341.145$ and $AICc=3520.87$

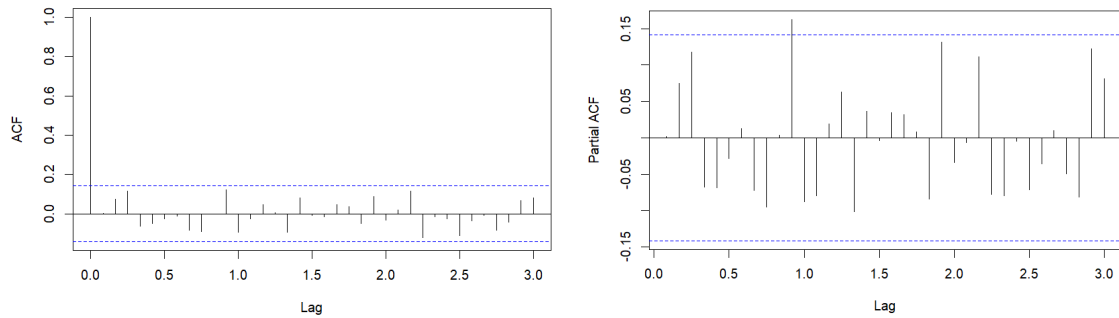


Figure: 5 & 6: ACF and PACF plots of $ARIMA(0,1,1)(1,0,0)[12]$.

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 $RMSE = 2688.7$ and $AICc=3587.93$.

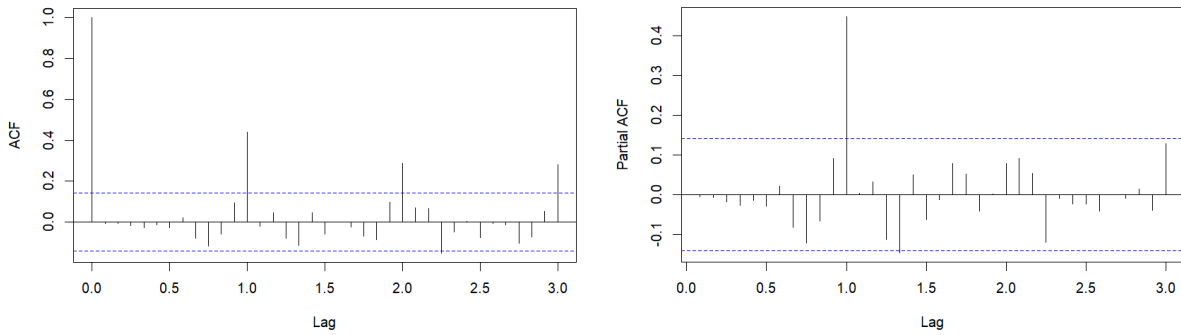


Figure 7 & 8: ACF and PACF plots of $ARIMA(0,0,3)$.

Looking at these plots further, we see that auto-arima does not use the differenced series. We do see that the first 11 lags have little autocorrelation and are not significant on both plots. 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.

3.1.2 Manhattan Total Waste

We begin with two preliminary plots.

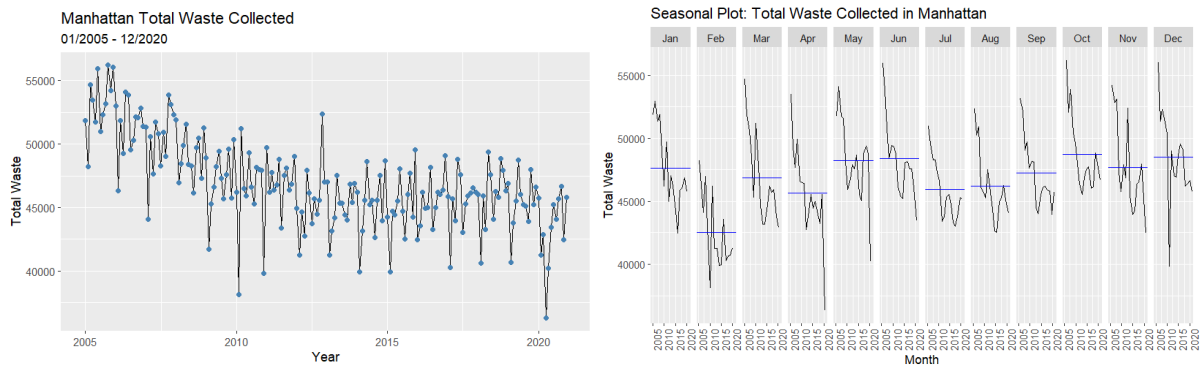


Figure 9 & 10: Total waste plot and seasonal plot of Manhattan.

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 $RMSE = 2215.12$ and $AICc = 3501.5$.

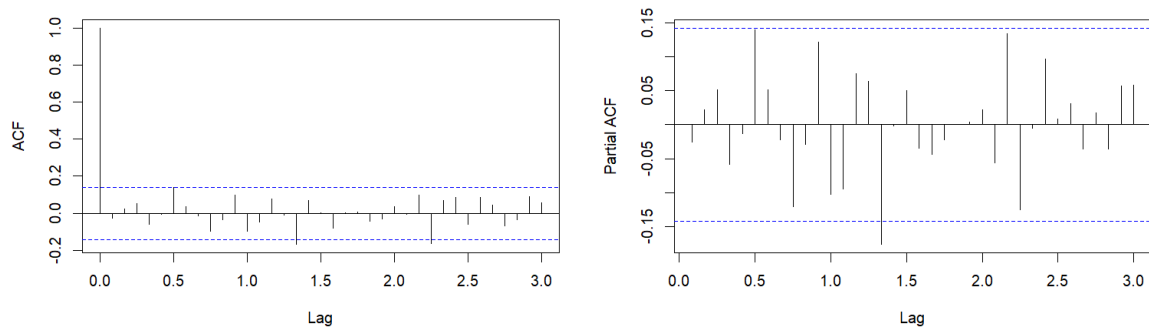


Figure 11 & 12: ACF and PACF plots of $ARIMA(0,1,2)(1,0,0)[12]$.

From the previous figure, the majority of the ACF and PACF lag values are contained within bounds. However, lag 16 in both plots is 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 $RMSE = 2208.185$ and $AICc = 3504.76$.

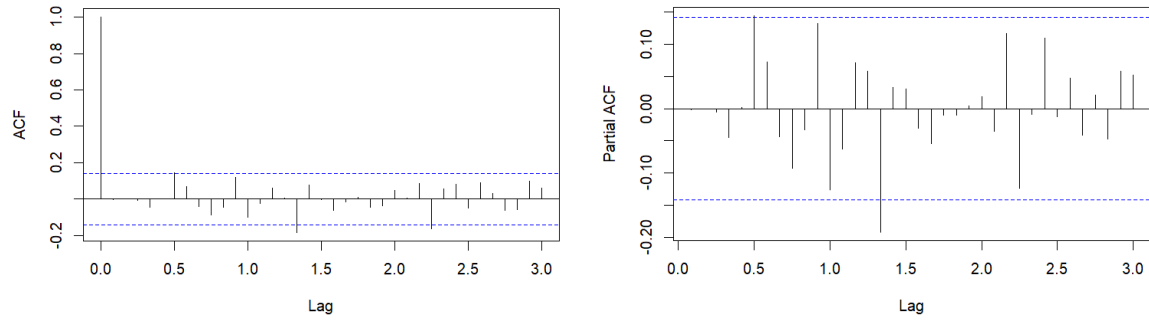


Figure 13 & 14: ACF and PACF plots of $ARIMA(0,1,4)(1,0,0)[12]$.

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 which returns a $ARIMA(0,1,5)$ model, with $RMSE = 2449.647$ and $AICc=3541.08$.

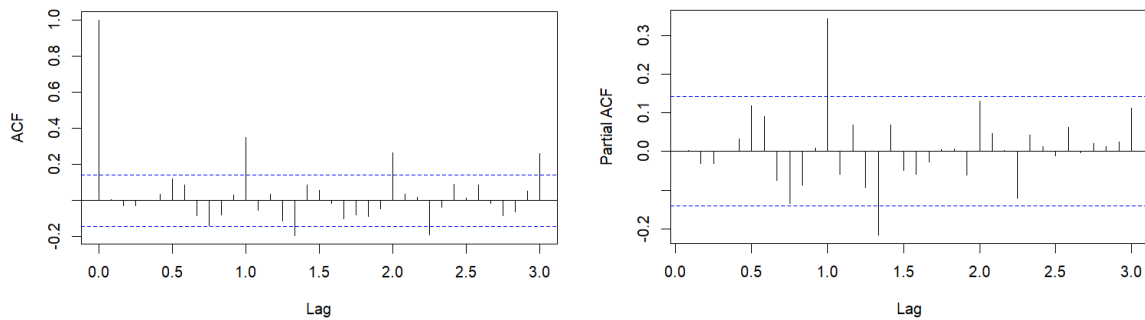


Figure 15 & 16: ACF and PACF plots of $ARIMA(0,1,5)$.

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.

3.1.3 Brooklyn Total Waste

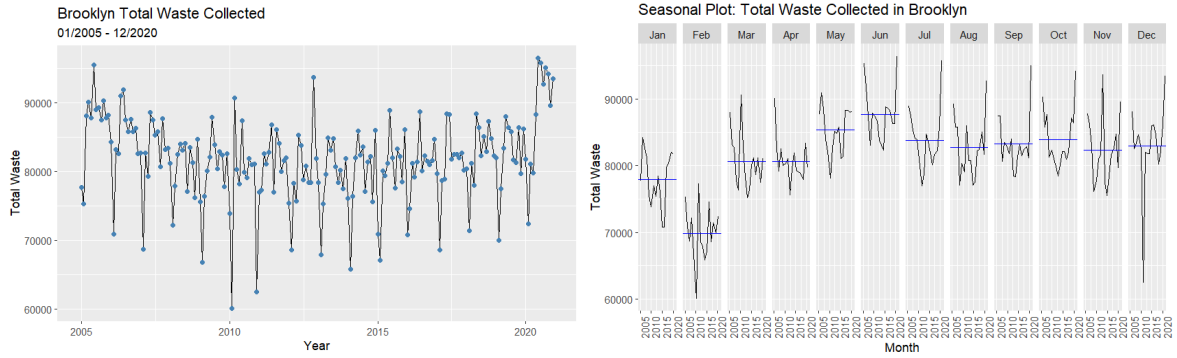


Figure 17 & 18: Total waste and subseries plots of Brooklyn refuse tonnage.

Most of the total 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 waste values are trend stationary. The same result holds for the differenced values.

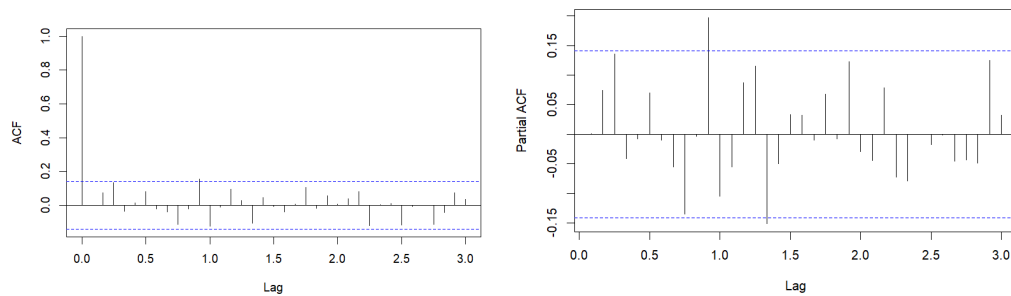


Figure 19 & 20: ACF and PACF of $ARIMA(0,1,2)(1,0,0)[12]$

After reviewing three Arima models, we believe that a $ARIMA(0,1,2)(1,0,0)[12]$ is the model that best captures the seasonality of the Brooklyn tonnage values. While also returning good autocorrelation values both on the ACF and PACF plots. This model returns a RMSE = 4427.169 and AICc = 3768.03.

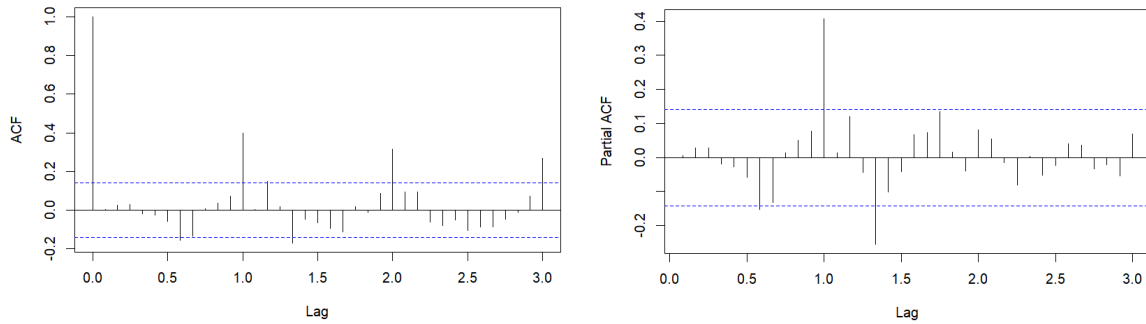


Figure 21 & 22: ACF and PACF of $ARIMA(3,0,3)$.

The auto-arma model selected to be compared to our previous model is a $ARIMA(3,0,3)$. This model returns a $RMSE = 5219.698$ and $AICc = 3849.87$. This model does not address any seasonality in the time series. The ACF plot shows positive and significant autocorrelations at lags 12, 24 and 36. For the PACF plot, the first 6 lags are not significant, but we do see a positively correlated and significant lag at lag = 12.

3.1.4 Queens Total Waste

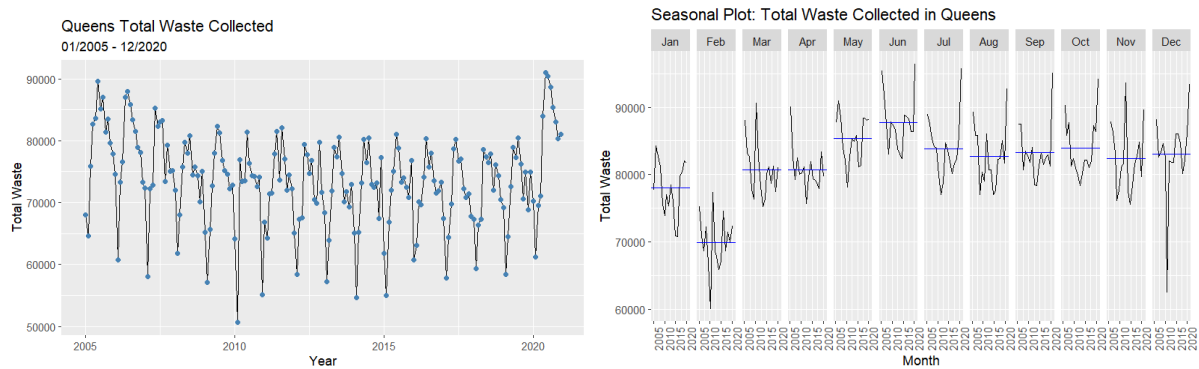


Figure 23 & 24: Queens total waste and subseries plot.

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.

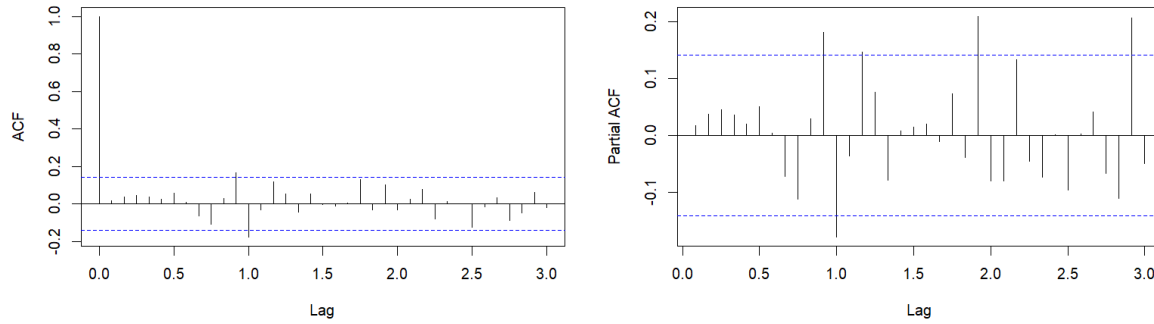


Figure 25 & 26: ACF and PACF of $ARIMA(3,1,1)(1,0,0)[12]$.

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.

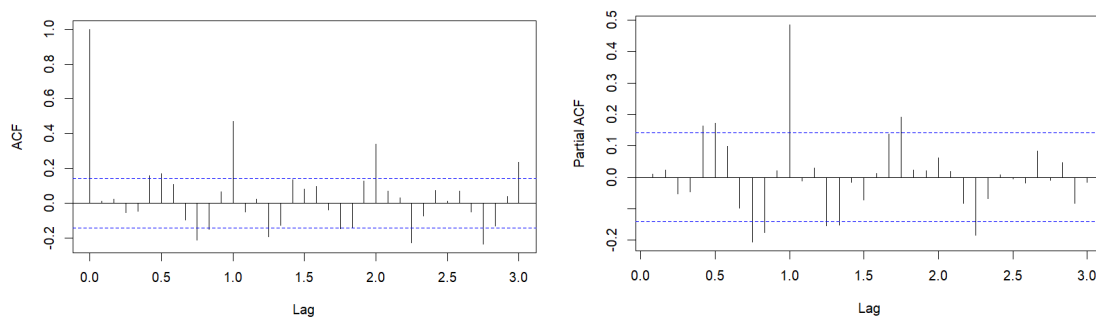


Figure 27 & 28: ACF and PACF plot of $ARIMA(3,0,2)$.

After reviewing two auto-arima models, it was best to report a $ARIMA(3,0,2)$. 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.

3.1.5 Staten Island tonnage values

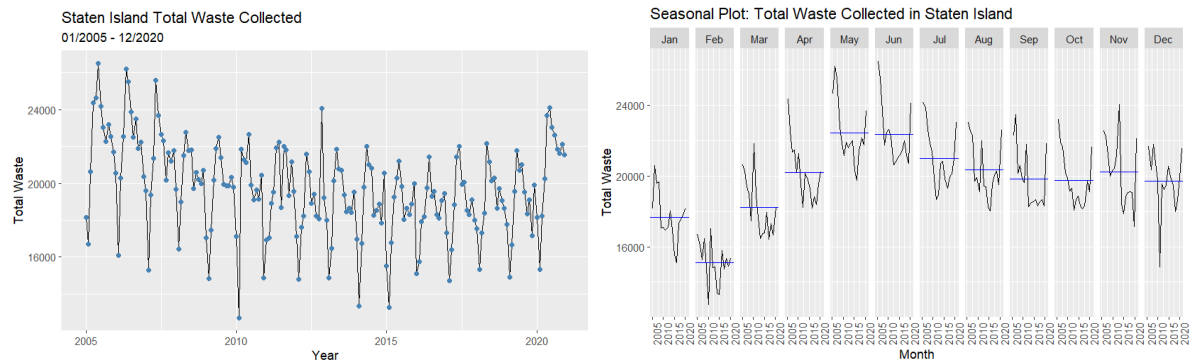


Figure 29 & 30: Staten Island total waste and subseries plot.

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.

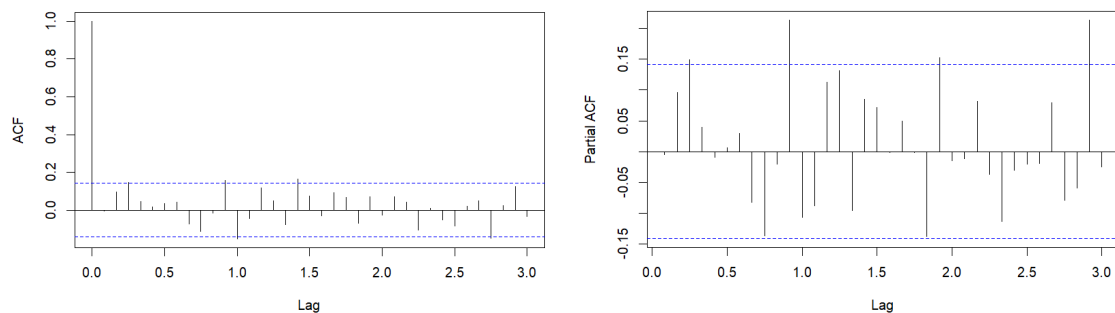


Figure 31 and 32: ACF and PACF plots of $ARIMA(2,1,3)(1,0,0)[12]$.

After researching numerous ARIMA models, we believe a $ARIMA(2,1,3)(1,0,0)[12]$ 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.

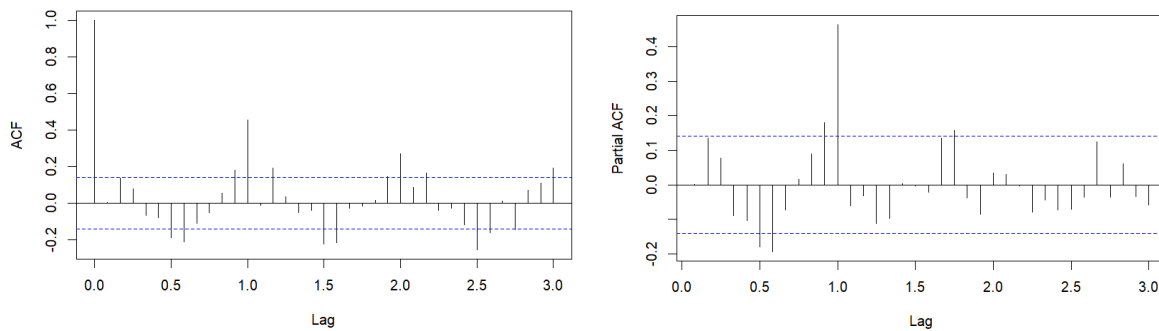


Figure 33 & 34: ACF and PACF plots of $ARIMA(2,1,4)$.

The auto-arma model selected was a $ARIMA(2,1,4)$ 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) are 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

We will use multiple linear regression model to best approximate and forecast tonnage values. 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 select the consumer price index (cpi), unemployment rate, average precipitation, average temperature and average cooling degree days as predictors for this model. 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.

This model 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.

$$\hat{y}_t = 257322.46 - 84.31 * cpi + 48960.81 * unemp_rate + 38695.86 * avg_precip + 1144.89 * avg_temp - 1639.11 * avg_CDD + \epsilon_t$$

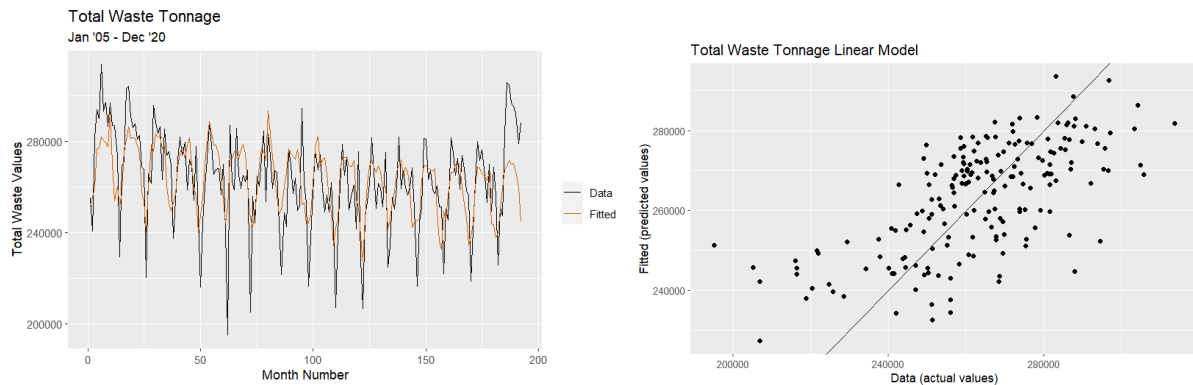


Figure 36 and 37: Fitted values on the total waste plot & fitted values plotted against the data.

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.

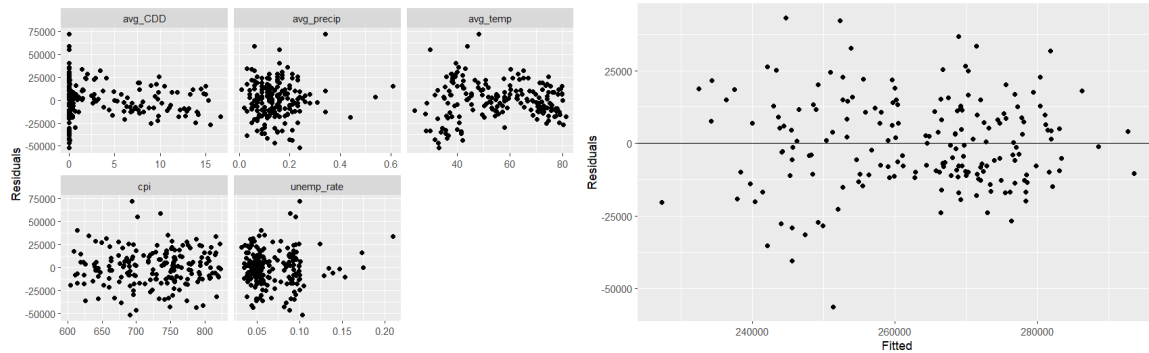


Figure 38 & 39: Residual plots against the predictors and fitted values.

From figure 38, the cpi and average temperature are the predictors where the residuals are randomly scattered. From figure 39, the model appears to show that the residuals are homoscedastic. This TSLM model returns an adjusted $R^2 = 0.4161$

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 ([fpp3](#)).

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. The two variables that are not stationary is the total waste tonnage and the CPI variable. We investigated the residuals of the model that contains the total waste tonnage being regressed onto the five predictors. And we investigated the residuals of the model of the differenced total waste tonnage being regressed onto the differenced values of each of the five predictors.

The residuals of both models, η_t , were stored as a time series separately. It is believed that the ARIMA model that best approximates η_t , is a $ARIMA(0,0,0)(1,0,0)[12]$ on both models.

3.3.1 Non-differenced variables - LM w/ $ARIMA(0,0,0)(1,0,0)[12]$ errors

For a model including all five predictors and a trend parameter, we are returned a RMSE = 12911.06 and AICc = 4204.69.

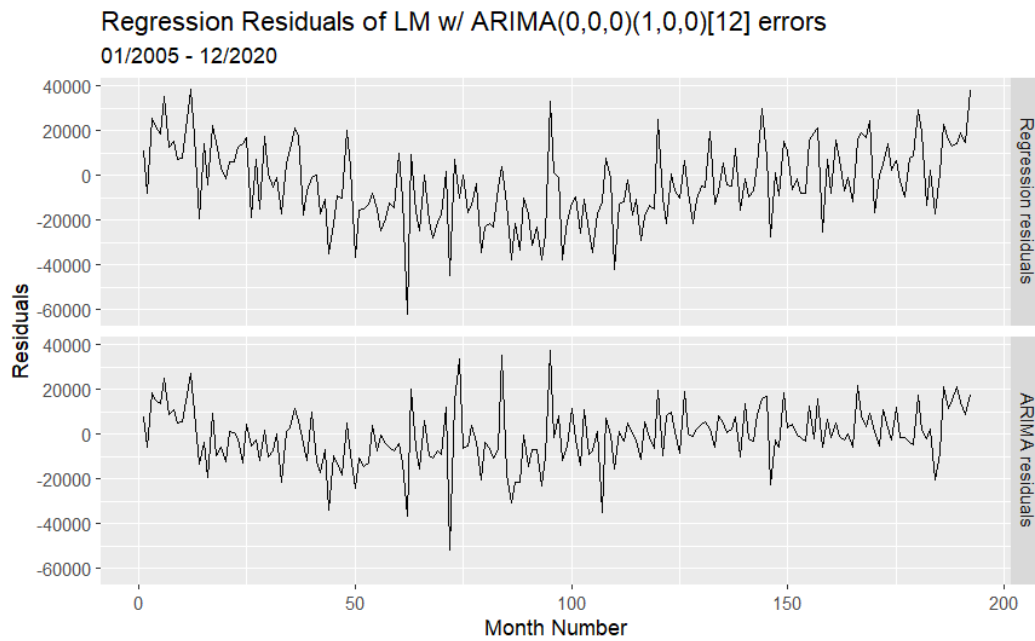


Figure 40: Regression residuals of this model.

3.3.2 Differenced variables - LM w/ $ARIMA(0,0,0)(1,0,0)[12]$ errors

This model also includes all five differenced predictors, and a trend parameter. It has a RMSE = 16898.46 and AICc = 4283.12.

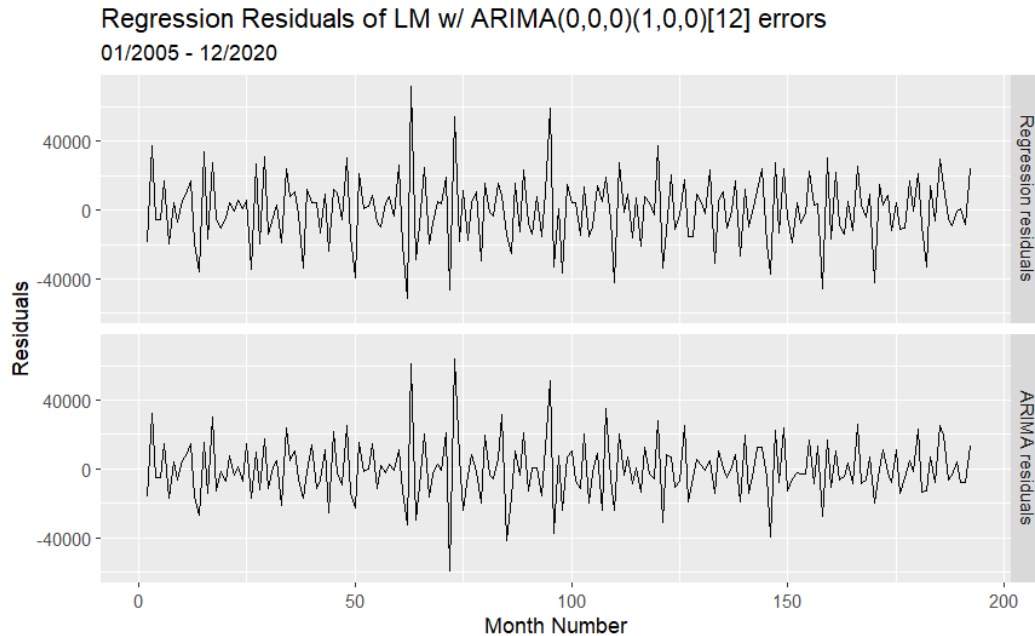


Figure 41: Regression residuals of this model.

4 Conclusion

We have written and discussed about three methods to forecast and analyze waste tonnage values collected in New York City. When investigating the total waste tonnage collected at the borough level, we have seen that seasonal auto-regressive integrated moving averaging models can best approximate the time series. There was not a single Arima model that can best summarize and forecast the tonnage values at each borough.

However, it was very common to see more than one non-seasonal MA() argument be used throughout the research. What we also saw in common, was the need to capture the seasonality of the time series. A pattern of significant and positively autocorrelated values at lags 12, 24, 36 meant that we should add a seasonal AR() argument to our models. In most cases, one seasonal AR() argument was enough to make the seasonal lags non-significant.

The decision to choose which predictors can be used for the multiple linear regression and the dynamic regression model was based on which predictors matched the time frequency of the total waste time series. Prior research studies have also drawn conclusions that weather variables such as the amount of rain, amount of snow, or average temperatures can help forecast waste generation. The unemployment rate and the consumer price index were also included in this research. Although these two variables may not be useful when forecasting waste tonnage, they matched the time scale and previous studies have also made use of socio-economic variables like these.

The assumptions of linearity between the predictors and the predicting variable were not thoroughly investigated beforehand. This was due to the time constraints when beginning to analyze and report the multiple linear regression and dynamic regression models. The values of the predicting variables were not all on the same scale. For example, average temperatures in Central Park varied between 40 and 80 degrees Fahrenheit. While the CPI ranged from 600 to 900. One option to consider would have been to scale and minimize all the tonnage and predicting values between 0 and 1.

This research was motivated by reading numerous papers that had a goal of predicting the amount of municipal waste collected within distinct cities. Predicting waste generation plays an important part in urban planning. With the population of cities growing, there is a need to manage and maintain systems to properly collect and dispose of waste and recyclables. Machine learning and other statistical methods play a big role in the studies mentioned above. The collection of waste is not linear, so there are times when a non-linear model can best summarize and forecast the amount of waste collected in the future. Continuing to add to this research topic was a pleasure, especially when working with the data of your hometown.

References

- Nicholas E. Johnson, Olga Ianiuk, Daniel Cazap, Linglan Liu, Daniel Starobin, Gregory Dobler, Masoud Ghandehari, Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City, *Waste Management*, Volume 62, 2017, Pages 3-11, ISSN 0956-053X, <https://doi.org/10.1016/j.wasman.2017.01.037>.
- Constantine E. Kontokosta, Boyeong Hong, Nicholas E. Johnson, Daniel Starobin, Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities, *Computers, Environment and Urban Systems*, Volume 70, 2018, Pages 151-162, ISSN 0198-9715, <https://doi.org/10.1016/j.compenvurbsys.2018.03.004>.
- J. Navarro-Esbrí, E. Diamadopoulos, D. Ginestar, Time series analysis and forecasting techniques for municipal solid waste management, *Resources, Conservation and Recycling*, Volume 35, Issue 3, 2002, Pages 201-214, ISSN 0921-3449, [https://doi.org/10.1016/S0921-3449\(02\)00002-2](https://doi.org/10.1016/S0921-3449(02)00002-2).
- X. Cuong Nguyen, T. Thanh Huyen Nguyen, D. Duong La, Gopalakrishnan Kumar, Eldon R. Rene, D. Duc Nguyen, S. Woong Chang, W. Jin Chung, X. Hoan Nguyen, V. Khanh Nguyen, Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam, *Resources, Conservation and Recycling*, Volume 167, 2021, 105381, ISSN 0921-3449, <https://doi.org/10.1016/j.resconrec.2020.105381>.
- Maryam Abbasi, Ali El Hanandeh, Forecasting municipal solid waste generation using artificial intelligence modelling approaches, *Waste Management*, Volume 56, 2016, Pages 13-22, ISSN 0956-053X, <https://doi.org/10.1016/j.wasman.2016.05.018>.
- Ezeudu OB, Ozoegwu CG, Madu CN. A Statistical Regression Method for Characterization of Household Solid Waste: A Case Study of Awka Municipality in Nigeria. *Recycling*. 2019; 4(1):1. <https://doi.org/10.3390/recycling4010001>
- Ajay Singh, Solid waste management through the applications of mathematical models, *Resources, Conservation and Recycling*, Volume 151, 2019, 104503, ISSN 0921-3449, <https://doi.org/10.1016/j.resconrec.2019.104503>.
- Křupka, Jiří & Kašparová, M. & Knížková, I.. (2013). Regression predictive models analysis of municipal waste. *International Journal of Mathematical Models and Methods in Applied Sciences*. 7. 111-120. https://www.researchgate.net/publication/289891560_Regression_predictive_models_analysis_of_municipal_waste
- <https://www1.nyc.gov/assets/dsny/site/about>
- <https://data.cityofnewyork.us/City-Government/DSNY-Monthly-Tonnage-Data/ebb7-mvp5>
- <https://people.duke.edu/~rnau/411arim3.htm>
- <https://fable.tidyverts.org/reference/ARIMA.html>
- <https://fable.tidyverts.org/reference/TSLM.html>
- <https://www.esri.com/about/newsroom/arcuser/fightsnow/#:~:text=During%20wintertime%2C%20DSNY%20is%20responsible,progress%20of%20snow%20removal%20operations.>

<https://otexts.com/fpp3/dynamic.html>