Time Series Forecast of Food Waste

Julia Troni

julia.troni@colorado.edu

Student ID: 109280095

Abel Iyasele

May 2023

Context

According to worldbank.org, the world generates 2.01 billion tons of solid waste annually, with at least 33 percent of that—extremely conservatively—not managed in an environmentally safe manner. And worse, global waste is expected to grow to 3.40 billion tons by 2050. These trajectories will have vast implications for the environment, health, and prosperity, thus requiring urgent action. [Trends in Solid Waste Management - World Bank]

Food scraps and yard waste make up more than 30% of our nation’s waste stream headed for the landfill. Landfills are the third-largest source of climate-damaging methane emissions in the United States. In a landfill, organic waste gets buried under mounds of toxic trash, so it can’t break down like it would in a compost pile. Instead, it rots and emits methane. What's worse is that these alarming climate implications are well known, however little action remains to be taken to change this.

Rather, all of that organic waste can be easily recycled and returned to the earth to benefit our environment through composting. Sadly, many cities do not have compost available and do not recycle organic waste.

For more cities to adopt waste collection and compost, they need to have accurate predictions of how much natural waste will be collected and the corresponding compost produced. This will give cities and corporations the necessary information to properly plan infrastructure and operations. It may also be beneficial to persuade more locations to implement similar composting programs.

Problem Definition

Given a time series dataset of tons of organic waste collected and estimated amount of compost that is generated as a result, the goal is to accurately predict the future amount of waste collected in order to provide more cities and governments with the necessary information to confidently begin an organic waste recycling and composting program.

Methodology

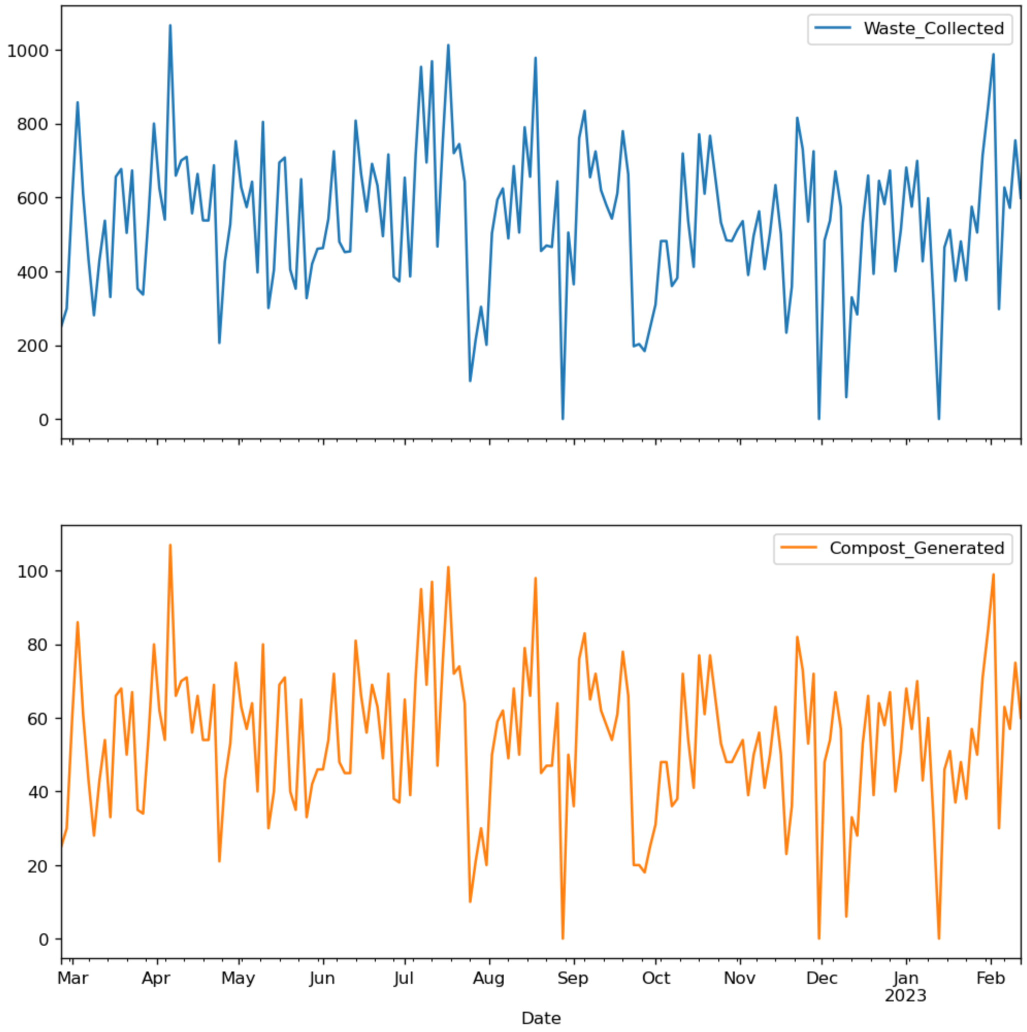
Data Set

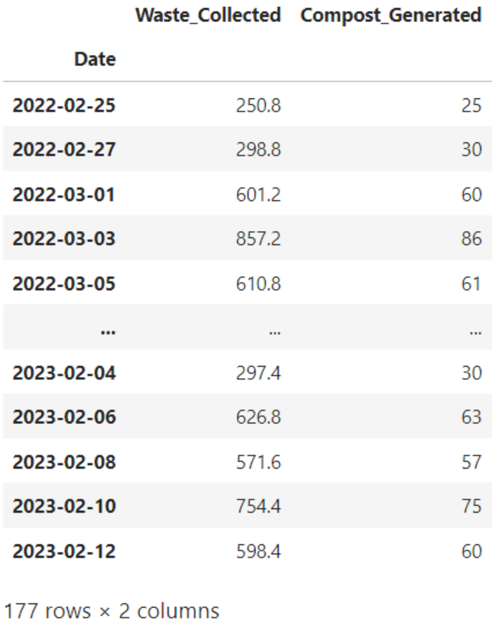
In some places, like Carey, North Carolina steps have been taken to reduce waste. Town of Cary's had a Food Waste Recycling Drop-Off program which is a local option to turn food and other organic waste into compost.

The proposed data source is a dataset from this program. The publicly available data includes the amount of pounds of waste collected and compost generated beginning February 7, 2022 to March 2023. [https://catalog.data.gov/dataset/food-waste-recycling-impact]

Data Analysis

First I examined the data on a high level. The original dataset is an unevenly spaced time series. Modeling unevenly spaced data presents challenges in many forecasting models. So, I applied a naive smoothing model. Given that the data was collected between February 2022 and March 2023, with 177 entries, I re-dated the data such that it is now sampled every 2 days. A view of the data can be seen in the following graphics.

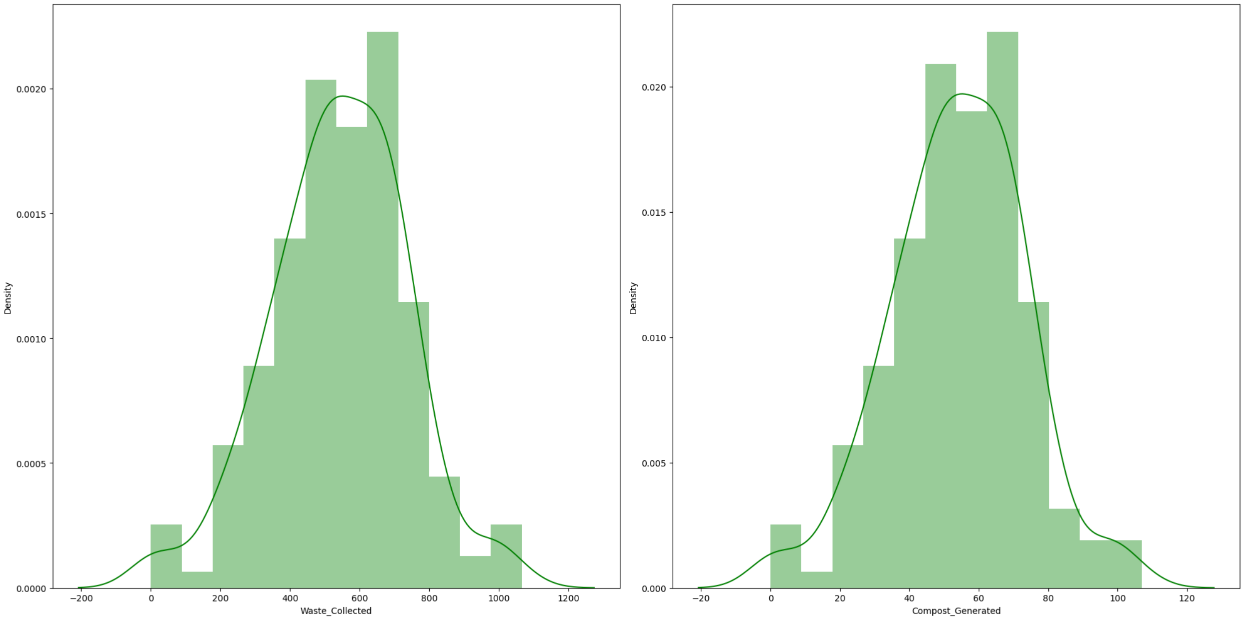




Date	Waste_Collected	Compost_Generated
2022-02-25	250.8	25
2022-02-27	298.8	30
2022-03-01	601.2	60
2022-03-03	857.2	86
2022-03-05	610.8	61
...	...	...
2023-02-04	297.4	30
2023-02-06	626.8	63
2023-02-08	571.6	57
2023-02-10	754.4	75
2023-02-12	598.4	60

177 rows × 2 columns

Next, I plotted the data shape. Upon first observation it appeared normally distributed. This gives us reason to believe that the data is stationary.



To confirm this, I ran an Augmented Dickey-Fuller (ADF) Test or Unit Root Test for checking stationarity. For lbs_collected the p-value (1.276730e-18) is smaller than 0.05, which further supports the rejection of the null hypothesis and the conclusion that the data is stationary. So no further transforming was necessary prior to modeling and the data for ‘lbs_collected’ is stationary. Since the data is stationary, no differencing is needed, that is, d=0 in my ARIMA model

Model

For modeling, the data was split into 80% training and 20% testing. Unfortunately, due to the very small dataset, this resulted in only 34 data points for testing.

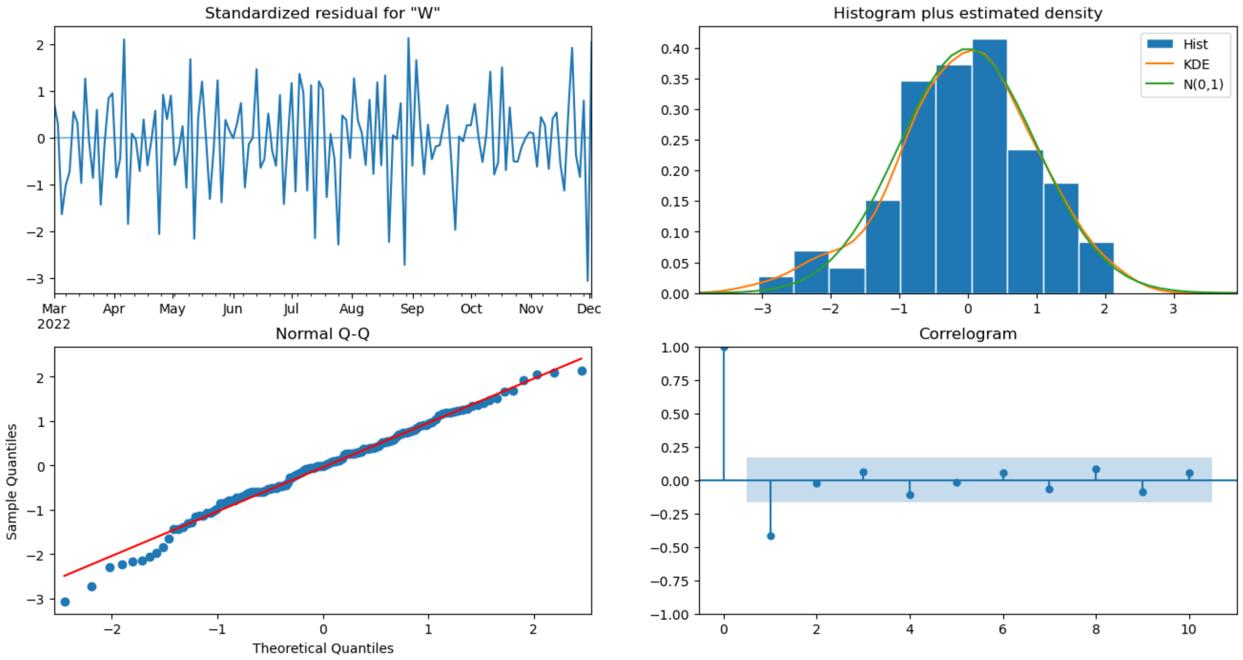
Since this is a time series dataset, I chose to implement an ARIMA (Autoregressive Integrated Moving Average) model which captures the linear relationships between data points and the lags or differences between them. They are best used when dealing with time series data to forecast future values based on historical patterns and trends and there is a need for short-term or medium-term forecasts. It is a good choice when there is evidence of autocorrelation in the data, trend or seasonality in the data, which can be adjusted using the integrated component of the model, and the data has already been preprocessed and cleaned without outliers, and it is stationary or can be made stationary through differencing.

Since my dataset is already stationary, as I determined above, I implemented ARIMA using the Python library “statsmodels.tsa.arima.model.ARIMA”. In an ARIMA model, the parameters for the model must be tuned and set. The parameters p,d, q are the lag order, differencing order or the number of times that the raw observations are differenced, and the size of the moving average window.

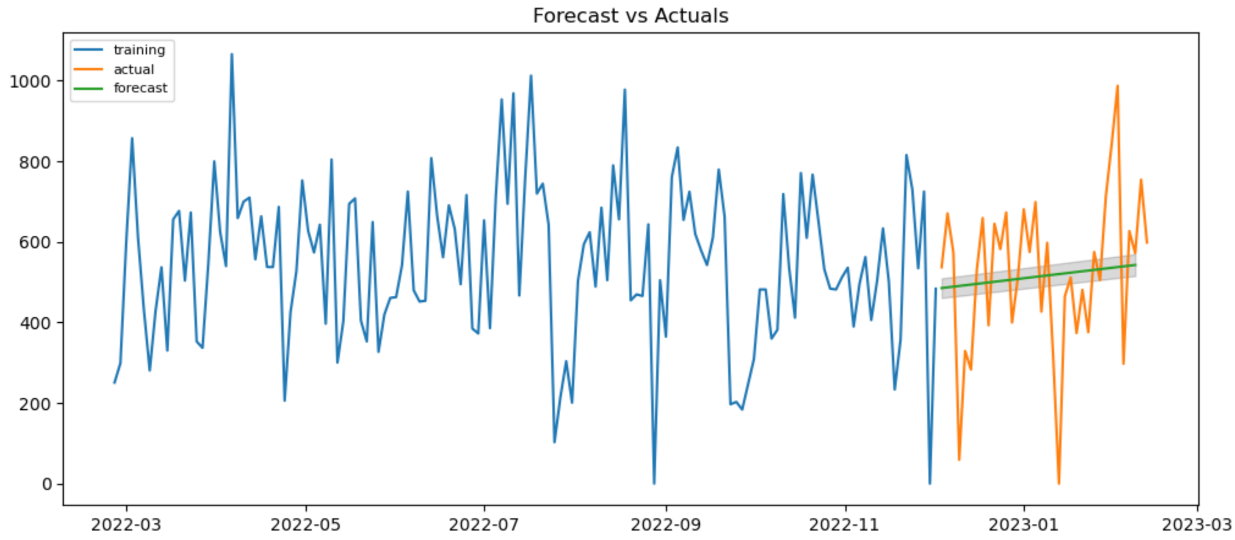
I implemented a grid search to find the optimal p, d, and q values, which was found to be 0,2,1 respectively. Next I fit the model and made predictions.

Description of Findings

To evaluate my model, I plotted various plots created using plot_diagnostics() shown below. The top left is the residual errors, which appear to have a uniform variance and a mean of zero which is promising. In the bottom left we see that the red line aligns with a majority of the dots. This is the desired behavior as any notable variations would indicate that the distribution is skewed. Then in the top right we see the density plot shows a normal distribution with a mean of zero. Lastly, the correlogram in the bottom right shows that residual errors are not very autocorrelated, however there are some. Autocorrelation would suggest that the residual errors exhibit a pattern that is not accounted for by the model and there is a need for more data, which as discussed below is a large limitation since the dataset is small.



In plotting the predicted values against the actuals, we can make many conclusions about our model's performance.



First, we see that the model has not accurately accounted for the fluctuations and has smoothed the data too much. This could be due to the small dataset and few train/test values, however more work needs to be done on this. One positive is that we see that the forecasted values accurately show the positive trend in the pounds of waste collected and the forecasts do lie within 95% of the actual values.

Unfortunately I did not have sufficient time to evaluate the summary statistics and that is something I continue to do and will update the project with as more work is done.

Limitations/Constraints

In this model the biggest constraint was the very limited data and dataset. Due to the lack of collection and compost data available, there were only 174 data points with which to model with. I believe this results in the poor quality of the predictions.

In my project my main challenge was time and finding data. Due to my niche interest in modeling food waste, I spent a great deal of time researching data sets. Unfortunately I even began cleaning and processing some datasets, before realizing that the dataset was either far too complex, or in one circumstance the dataset was removed from public availability. Hece with these efforts I wasted my time.

Once I finally began working with this dataset, I had less than a month prior to the submission deadline. Obviously this was less than ideal. For the sake of the grade, I am submitting as is, but I plan to continue work to tune the model, implement other models, and model the amount of compost generated, which I am particularly interested in.

Conclusion

Time series prediction models are important for forecasting trends in the waste and environment management sector because they can help decision-makers plan and allocate resources effectively. By analyzing historical data and identifying patterns, these models can predict future trends, providing insights into the amount of waste that will be generated, the quality of air and water, recycling rates, and energy consumption. These predictions can help waste management companies and government agencies to plan for resource allocation, infrastructure development, and policy formulation, reducing waste and environmental impacts. Additionally, time series prediction models can help environmental agencies issue warnings and take action to mitigate pollution levels, improving public health and the environment. Overall, the use of time series prediction models can help decision-makers make informed decisions, leading to more efficient resource use and a better quality of life for people and the planet.

Having an accurate forecast of expected waste drop off amounts is important so that the town can adequately allocate resources for safe collection, storage, and preparation for the composting. With an accurate understanding of the amount, the hope is that more towns will have the necessary information to implement similar composting programs as Carey, North Carolina did here.

References

https://data.townofcary.org/explore/dataset/food-waste-pilot/table/?sort=date_collection_datetime

https://catalog.data.gov/dataset/food-waste-recycling-impact

Trends in Solid Waste Management - World Bank

ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+ (machinelearningplus.com)

Introduction to Time series Modeling With -ARIMA - Analytics Vidhya
Stationarity | Statistical Tests to Check Stationarity in Time Series (analyticsvidhya.com)

Netflix Stock Prediction with ARIMA | Kaggle

Time Series Analysis- Part II. In the previous blog we learnt about… | by Madhu Ramiah | Medium

Time Series in Python — Part 2: Dealing with seasonal data | by Benjamin Etienne | Towards Data Science (medium.com)