Rat Activity Time Series Forecasting

1st Jiya Varma
GT 4801
Georgia Institute of Technology

***Abstract*—This paper provides insight into data cleaning, exploration, modeling, and forecasting of rat activity in Manhattan. Specifically, it explores the NYC Department of Health and Mental Hygiene's (DOHMH) database on rat inspections across the 5 major NYC boroughs that have been logged on a daily basis since 2008. The dataset was reduced into a time series of monthly totals of rat activity found from the inspections from 2018 to 2023 in Manhattan. The primary model used to forecast rat activity from the remainder of 2023 to 2026 was the seasonal autoregressive integrated moving average model (Seasonal ARIMA). These results may be useful to New York's rat mitigation efforts moving forward.**

***Index Terms*—Rat inspections, rat activity, New York rats, ARIMA, time series**

I. Introduction

The boroughs of New York have been plagued by rat infestations since the early 19th century and pose a significant public health risk for its residents as they can carry and spread diseases and contribute to property damage [1]. New York's rat problem is not just a matter of numbers but also of location, since lower-income neighborhoods tend to be disproportionately affected by rat infestations. Katheleen Corradi, New York City's Director of Rodent Mitigation puts it best as "Rats are a symptom of systemic issues, including sanitation, health, housing, and economic justice" [2].

NYC Open Data provides access to The New York Department of Health and Mental Hygiene (DOHMH)'s rat inspection data set [3], which includes important information such as whether the inspection discovered rat activity, location information, and inspection date (among several other columns that are extraneous for the purpose of this analysis). This information was used to extract trends and seasonality and make predictions for future rat activity on a monthly/yearly basis. Finding these correlations can help New York's rat mitigation task force to be better equipped to handle its rat infestation problem

This paper is organized into the following sections: Related Works, Data Cleaning, Data Exploration, Data Modeling, Results, Conclusion + Future Work, References.

II. Related Works

A. Rat sightings in New York City are associated with neighborhood sociodemographics, housing characteristics, and proximity to open public space [4]

This paper uses rat sighting dataset (similar to the rat inspection dataset, both provided by the NYC Dep of Health and Mental Hygiene) in an investigation to show how public spaces/subway lines, vacant housing units, and low education of population correlates to greater amounts of rat sightings. Provides public health officials with areas to target and deploy rat control efforts in, and the models used to find these results may be beneficial to epidemiologists studying the movement of pathogens in urban areas

B. Rat city: Visualizing New York City's rat problem [5]

This article uses rat sighting dataset from NYC DOHMH to create several heat map visualizations from data collected in 2017 to showcase rat sighting densities across areas like Lower Manhattan, Midtown, Upper East Side, Upper West Side, etc and even identifies specific streets with highest densities. It also showcases the progression of rat sightings throughout the years, as well as rat sightings in each borough each year. Another data plot that is presented is the location type (family appt, commercial building, 1-2 person appt, etc) and number of rat sightings.

C. Does New York City really have as many rats as people? [6]

This investigation also uses the rat sighting dataset from NYC DOHMH. It provides a map visualization of rat sightings using a subset of the data (Jan 2010 - July 2011) in a small part of Brooklyn. It provides data represented in different colors from two sample periods: the first half of 2010 and the first half of 2011 to draw attention to where there have been multiple rat sightings. Later, the author also draws attention to the fact that all of the reported data is based on the general population's 311 calls to report rat sightings, and thus there may be other factors contributing the the various densities of 311 calls throughout different neighborhoods. However, the investigation is confined to studying one neighborhood at a time, so it can ignore this possible confounding variable. Additionally, there is a 6 month buffer between sample periods to serve as a "cool-down" period under the assumption that the city made efforts to reduce the rat population based on rat sightings/311 calls from the initial period. Overall, this study aims to estimate the number of rats present in New York based on the number of rat-infested lots and average # rats present in rat colonies, and comes to the conclusion that there are approximately 2 million rats in New York.

III. Data Cleaning

The Rat Inspection dataset contains over 2.45 million rows of rat inspection information that has been logged on a daily basis since 2008. Because of this, this dataset was too large to be processed by RStudio's read_csv() function in a single run.

Thus, I decided to split up the .csv file into multiple .csv files that contained rat inspections by year using a simple python script. Then, I separately loaded in the .csv files from 2018 to 2023 (in an arbitrary decision) and cleaned them in the method outlined below.

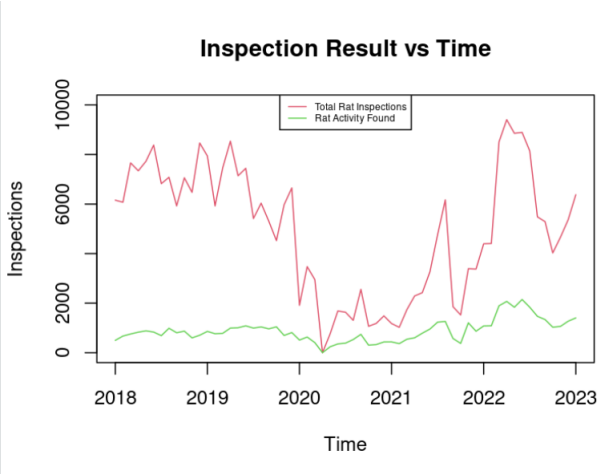
The dataset contains several columns of information that provide insight into the type of inspection conducted, exact coordinates, borough, inspection ID, date/time, result of the inspection, etc. Since this report's goal was to forecast future rat activity by month and year in Manhattan, the variables of interest were: BORO_CODE, INSPECTION_DATE, and RESULT. After extracting these columns for each year's .csv, I appended them all back into a single .csv and continued with the data reduction.

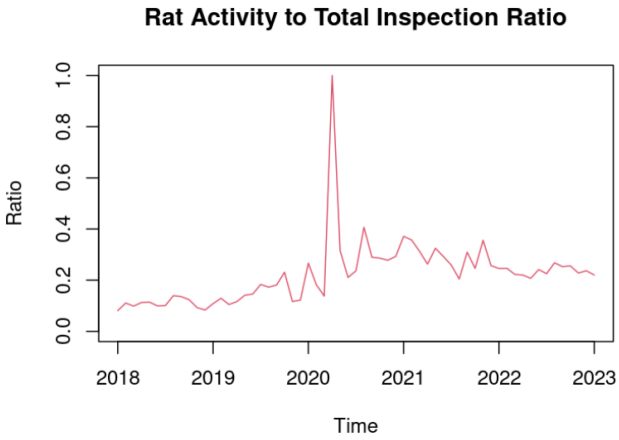
The BORO_CODE column provides a value ranging from 1-5 to specify which borough the inspection took place in (1 = Manhattan, 2 = Bronx, etc). Since we are only interested in Manhattan, we selected only the rows with BORO_CODE == 1. The INSPECTION_DATE provides a date-time value, but we need to separate this value into month and year in order to aggregate rat activity totals over month and year. This was done using the parse_date_time(), as.Date(), srftime() methods and created an additional 'month' and 'year' column. Rat inspections could RESULT in a variety of different outcomes such as success (no rats found), stoppage done, bait applied, cleanup done, monitoring visit, rat activity, etc. Out of these outcomes, the only one meaningful for this analysis was the inspections that resulted in 'Rat Activity.' Thus, we selected only the rows in which RESULT == 'Rat Activity.'

In order to make the aggregation sum easier, we add another column called 'Rat Activity' and assign each cell with '1' since each row now signifies a single inspection in which rat activity was found in Manhattan. We then aggregate 'Rat Activity' by 'month' and 'year' in order to extract the monthly totals using the aggregate() function. Since we know the 'Rat Activity' column contains rat activity totals ordered by month and year, so we extract this column into a new data frame and convert it into a Time Series object using the ts() function, with the start year = 2018, end year = 2023, and frequency = 12 for the months. This concludes the data cleaning step, and the next section will discuss the data models applied to this time series.

IV. Exploratory Analysis

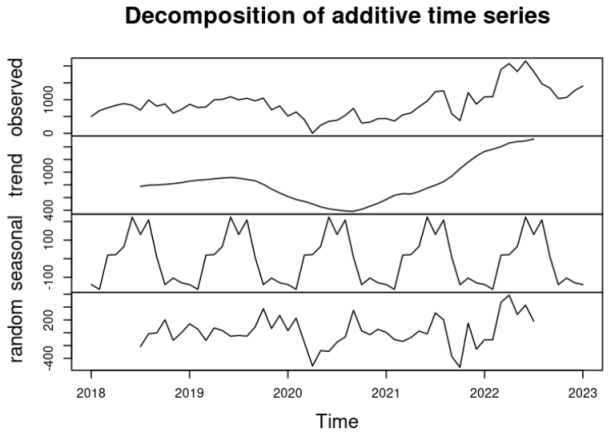
A ts.plot() function was used in order to visualize the possible trends and overall growth of the amount of rat activity from 2018 to 2023 on a monthly basis. It's also useful to see the amount of rat activity found compared to the total number of inspections.





It's interesting to note that while rat activity was lowest in early-mid 2020, the rat activity to total inspection ratio was the highest. Perhaps the pandemic may have had an impact on rat inspection efforts, as well as rat activity.

To gain a better insight into the trends and seasonality of rat activity, we apply an additive decomposition using the decompose() function. We assume additive decomposition since there is no drastic difference in seasonality



The 'trend' subgraph of Table 2 shows a slight growth in rat activity from 2018 to mid 2019, before exhibiting a

dip between mid 2019 and mid 2020, followed by a steeper incline from mid 2020 to 2023. Again, the pandemic and post-pandemic may have played a role in the dip and incline respectively. The 'seasonality' subgraph shows a cyclic pattern of increasing rat activity during the summer months, followed by a dip in the winter months. This aligns with the fact that rats are typically most active in summer months due to the increased availability of food and resources [7]

V. Data Model

There are several data models that are often used for time series forecasting such as exponential smoothing and ARIMA. In this analysis, I investigated the ARIMA model, which stands for Autoregressive Integrated Moving Average, and generally the framework outlined here [8] [9]. This model has 3 major components: AR (Autoregression), indicating a model with current values that are dependent on past values, I (Integrated) indicates the differencing needed to make the data stationary, and the MA (Moving Average) which uses past errors to calculate future values. ARIMA models have parameters (p, q, d) that represent the number of autoregressive terms, nonseasonal differences, and lagged forecast errors, corresponding to the AR, I, MA parts of the model.

The non-stationary/seasonal ARIMA model assumes univariate, stationary data. Thus, if there is a non-stationary component in the data, such as seasonal trends, we must difference the data in order to remove it. The augmented Dickey-Fuller test (ADF) can tell us whether data is stationary or not via hypothesis testing and p-value results. If not stationary, we can use the ndiff() method to find out how many times we must difference to reach stationary status. The next concept to consider is the autocorrelation function, or ACF, which provides insight on the extent to which a time series correlates to its lagged/shifted versions. The shape of the ACF-Lag function provided by the acf() command helps us in determining the correct p and q values. We can repeat these steps multiple times to figure out the best p,d,q values that will minimize the Akaike information criterion (AIC) as this indicates the best fit for the model without overfitting.

Like the non-stationary/seasonal ARIMA model, there is also a seasonal counterpart which accounts for both non-seasonal and seasonal factors. Here, there are additional P, D, Q values corresponding to the seasonal order of the ARIMA model that we must account for. The 'forecast' package offers an easy way to do this via the auto.arima() function, which automatically traverses through p,q,d, P, Q, D values to find the best model. Once we find these parameters for the model, we can feed it into the forecast() method and generate a plot containing the existing time series, and the forecasted values from the model.

VI. Result

Applying the auto.arima() function on the Rat Activity time series generated the ARIMA(0,1,0)x(0,1,1)[12] model, where 12 is the number of periods in the season (aka months). Note that you must set D = 1 so that auto.arima() can pick up on the seasonality. Each parameter of the model can be broken down as follows:

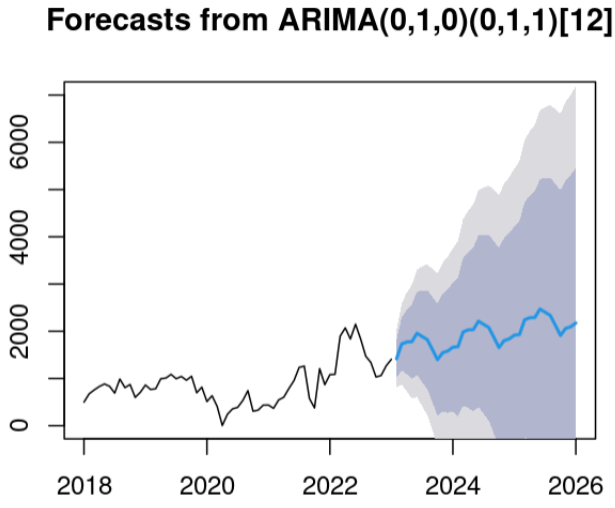
Non-seasonal components

- p = 0
  - indicating there is no autoregressive component in the model
- d = 1
  - indicating a first order difference was applied, effectively "de-trending" the data
- q = 0
  - indicating that no moving average component in the model

Seasonal components

- P = 0
  - indicating there is no seasonal autoregressive component in the model
- D = 1
  - indicating a first order seasonal differencing was applied
- Q = 1
  - indicating the presence of a seasonal moving average term

We then feed this ARIMA model to the forecast() function to produce the following plot that estimates rat activity until 2026 (choosing this year was an arbitrary decision), where the blue line contains the estimated values and the shaded purple regions indicate confidence intervals



We can see that the blue line exhibits both a seasonal component and an upwards trend, which is fairly consistent with the original rat activity time series.

A closer look at the forecasted values can be seen below

	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
Oct 2023	1397.585	215.0033	2580.167	-411.01772	3206.188
Nov 2023	1544.174	297.6237	2790.725	-362.26040	3450.609
Dec 2023	1580.991	273.5980	2888.385	-418.49429	3580.477
Jan 2024	1660.498	294.9699	3026.026	-427.89693	3748.892
Feb 2024	1672.657	206.0559	3139.258	-570.31570	3915.629
Mar 2024	1984.471	423.3267	3545.614	-403.09285	4372.034

VII. Conclusion and Future Considerations

The forecasted rat activity in Manhattan exhibits an upwards linear growth with a seasonal component for the remainder for 2023, up until 2026 and beyond. This analysis mainly addressed the ”how many” aspect of New York’s rat infestation problem, but failed to address the ”where” aspect. Thus, a future iteration of this analysis would aim to conduct a multivariate time series model to forecast rat activity in all five boroughs. Additionally, it would be interesting to somehow investigate the rat activity to inspection ratio versus the surface area of each borough to extract information on rat activity density across New York.

References

[1] Pasley, J. (2023, March 18). How the rat population in New York City grew by 800

[2] Mayor Adams Anoints Kathleen Corradi as NYC’s first-ever “rat czar.” The official website of the City of New York. (2023, April 12). https://www.nyc.gov/office-of-the-mayor/news/249-23/mayor-adams-anoints-kathleen-corradi-nyc-s-first-ever-rat-czar-#/0

[3] Department of Health and Mental Hygiene, (DOHMH). (n.d.). Rodent Inspection. New York. https://data.cityofnewyork.us/Health/Rodent-Inspection/p937-wjvj

[4] Walsh M. G. (2014). Rat sightings in New York City are associated with neighborhood sociodemographics, housing characteristics, and proximity to open public space. PeerJ, 2, e533. https://doi.org/10.7717/peerj.533

[5] Frei, L. (2019, January 20). Rat city: Visualizing New York City’s rat problem. Medium. https://towardsdatascience.com/rat-city-visualizing-new-york-citys-rat-problem-f7aabd69 00b2

[6] Auerbach, J. (2014), Does New York City really have as many rats as people?. Significance, 11: 22-27. https://doi.org/10.1111/j.1740-9713.2014.00764.x

[7] Dowd, B. (2023, August 11). Where do rats go during summer. Skedaddle Humane Wildlife Control. https://www.skedaddlewildlife.com/blog/where-do-rats-go-during-summer/#: :text=Summer

[8] Chatterjee, S. (2018, January 18). Time series analysis using Arima model in R. DataScience+. https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

[9] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on Sept 24, 2023