Modeling the Number of Rat Activities in NYC Rodent Inspection

Ziyou Hu

May 2023

1 Introduction

The data set "Rodent Inspection" records rat inspections in New York City from 1918 to today. I specifically focused on a particular region in Manhattan from 2010-2014. I investigated the probability of encountering rat activities among initial inspections. I used a Poisson random variable to model the number of rat activities among initial inspections in a given month and a given year. I found that high number of inspections have lower probability for rat activities and vice versa for low number of inspections. The Poisson model does not account for the fluctuation of percentage with the number of inspections. The model performs well in predicting the number of rat activities, but the prediction has some errors when the number of inspections is particularly high or low.

1.1 Description of the Data Set

The data set is available on NYC Open Data, which provides data created by organizations associated with New York City to the public. "Rodent Inspection" is collected by the Department of Health and Mental Hygiene (DOHMH) and updates daily. As of May. 1,2023, there are 2,363,896 observations. The earliest inspection

dates back to 1918 (with only 1 inspection). However, very few observations were made between 1918-2009 compared to the rest of the data set. The numbers of inspections are more consistent after 2009. Among all data collected, 32% was collected in in Manhattan, 29% was collected in Bronx, 28% was collected in Brooklyn, 9% was collected in Queens, 3% was collected in Staten Island. Thus, Manhattan is the most inspected borough.

1.2 Variables in the Data Set

There are 20 columns in the data. They record the inspection's type, location, number of inspections before, date, and result. I focused on the type, date, and result, which correspond to the columns: INSPECTION_TYPE, INSPECTION_DATE, and RESULT.

Types of inspections:
a) **Initial**: response to complaints or a proactive inspection as part of the neighborhood indexing program. The 3 possible results of initial inspection are: "Failed for Other R", "Passed", and "Rat Activity".
b) **Compliance**: returning after a failed inspection. The possible results are the same as initial inspections.
c) **Bait**: Using rodenticide or conducting a monitoring visit. The possible results are "Bait applied" or "Monitoring Visit".
d) **Clean Up**: Emptying trash and clutter, where rodents tend to live. The only result is "Cleanup done".
e) **Stoppage**: Filling up holes and cracks to reduce rodents' movements. The only result is "Stoppage done".

Among all observations in the data set as of May 17, 2023, 70% of the inspections are Initial inspections, 14% is Bait, 16% are Compliance inspections. The Stoppage and Clean up inspections make up almost 0% of the inspections. Thus, most of the inspections are initial inspections.

In summary, here are the

Types of results

a) **Bait Applied**
b) **Cleanup done**
c) **Failed for Other R**
d) **Monitoring visit**
e) **Passed**
f) **Rat Activity**
g) **Stoppage done**

(DOHMH, 2015)

The determination of Rat Activity and Failed for Other R are not well-documented in the user guideline for the data set. Based on the guideline, I infer that Rat Activity is recorded if there are Active Rat Signs. Active Rat Signs include "1) fresh tracks, 2) fresh droppings, 3) active burrows, 4) active runways and rub marks, 5) fresh gnawing marks, and 6) live rats" (DOHMH, 2015). "Failed for Other R" are conditions that foster rat population, such as trash, clutter, vegetation, and mice (DOHMH, 2015).

As emphasized by the DOHMH, the number of inspections at a particular location does not necessarily indicate the abundance of rats in that location and the absence of rats elsewhere (DOHMH, 2015). The number of inspections has to do with number of rats, how well the rats hide, how many people live in that area, the importance of hygiene in that area, the availability of inspectors, and so on.

1.3 Question of the project

What is the probability of finding rat activities for an initial inspection given a month and a year between 2010 and 2014?
Rats tend to appear in April, May, June, October, and November (DC Health),

so I expect these months to have higher percentage of rat activities. Also, rat populations might thrive more in some years than other years due to New York City's policy and human population change. Therefore, I expected to observe a monthly and a yearly change for the percentage of rat activities found in initial inspections.

1.4 Narrowing Down the Original Data Set

From the original data set, I decided to study the probability of encountering rat activities each month among initial inspections in a particular area at Manhattan between 2010-2014.

Narrowing down the type of inspections:
I only selected initial inspections because it is the only inspection type that has no previous or current impacts to the probability of encountering rat activities. Other types either has previous or current influence on the probability of encountering rat activities. For example, Compliance follows from a previous visit, and Clean up remove the clutter, which makes rat populations less likely to survive.

Narrowing down the date of the inspections:
I narrowed down the data to 2010-2014 because it is the five years with consistent observations. Consistency in data means lower chance of observing a trend by chance.

Narrowing down the area of the inspections:
I narrowed down to a region in Manhattan with X_COORD from 984005 and 991147, and Y_COORD between 197160 and 209995.

There are 2 reasons for narrowing down the area of study. First, to observe yearly or monthly trend in my data, I want to make the observations I study as uniform as possible. Therefore, I want to reduce geographic variations. Second, I want to reduce the chance of observing a trend by chance by using inconsistent

observations. Manhattan is one of the 3 boroughs that have the most inspections. However, not all areas in Manhattan receive similar amount of inspections. I used the map on Rat Information Portal provided by NYC health to determine an area that is densely inspected. More information is included in the appendix.

2 Exploratory Data Analysis

2.1 Basic Information about the Selected Data

The number of inspections from 2010-2012 is consistently around 9000. The number of inspections in 2013 and 2014 dropped to around 5000. In 2010, there are 9313 initial inspections. In 2011, there are 9093. In 2012, there are 8838. In 2013, there are 5090. In 2014, there are 5774. Initial inspections makes up about 80% of the inspection in each year I studied, so the type of inspections is consistent.

2.2 Examine Number of initial inspections and the Percentage of Rat Activities

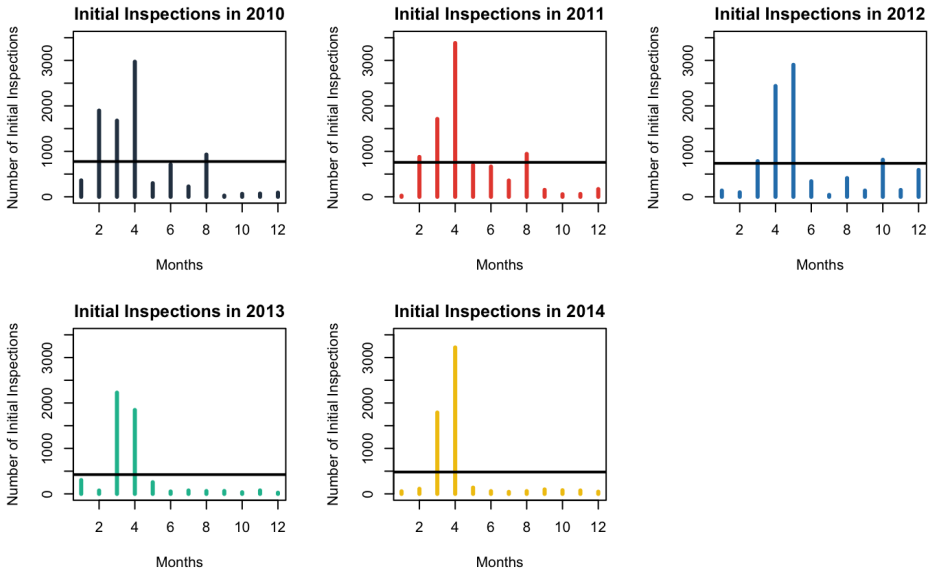

Figure 1: Number of initial inspections across months from 2010-2014. The horizontal line is the average number of initial inspections per month in that year.

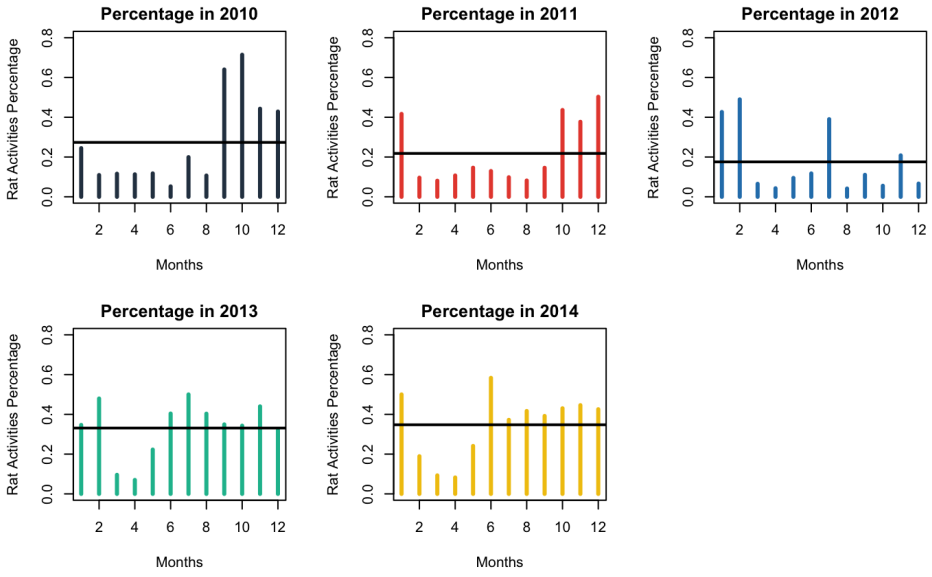

Figure 2: Percentage of rat activities across months from 2010-2014. The horizontal line is the average percentage per month in that year.

I expected to observe yearly and monthly trend for the percentage of rat activities. According to Figure 2, winter months such as January, October, November and February usually have the higher likelihood for encountering rat activities. Spring months and early summer months, such as March, April, and May have lower percentage of finding rats. 2013 and and 2014 in general have higher percentage of encountering rat activities.

However, when I compare to Figure 1, the months with higher percentage are the months that have lower number of inspections, and vice versa for the months with lower percentage. This holds true for the years as well. 2013 and 2014 are the years with the lowest number of inspections.

To further explore the relationship between the number of inspections and the percentage of rat activities, I plotted the number of rat activities against the number of initial inspections. I expected to find the rate of increase of the number of rat activities to reduce as the number of initial inspections increase.

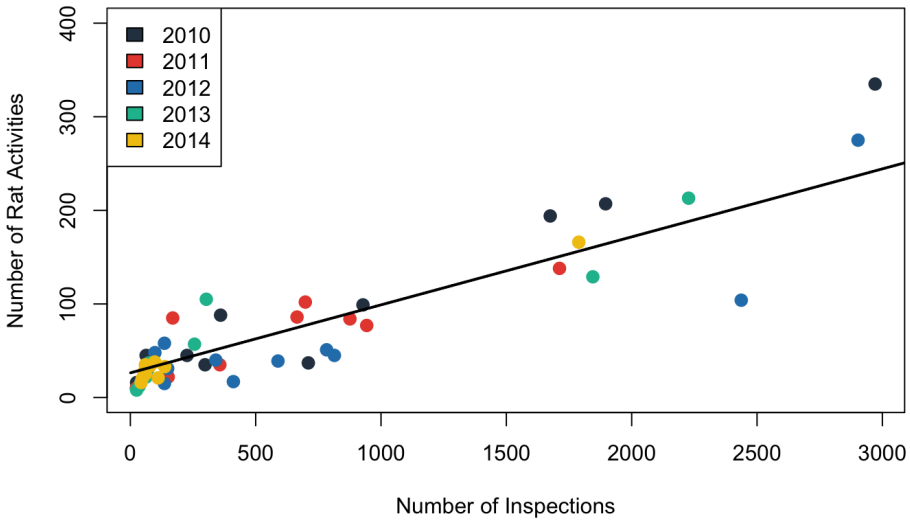

Figure 3: Number of Rat Activities against Initial Inspections Across Months from 2010-2014

Contrary to my expectation, the plot illustrates a linear relationship between the number of inspections and the number of rat activities. Through a linear regression, I found the slope is -0.07265 and the y-intercept is -26.36643. I plotted the line onto the graph. Number of inspections can be correlated with the percentage of rat activities, but this correlation cannot be captured by eyes.

3 Model Fitting

The linear relationship found between the number of inspections and the number of rat activities motivated me to embed the linear relationship into the model.

I modeled the number of rat activities spotted in a month in a year as a random variable X distributed as a Poisson random variable. $X \sim Poisson(\lambda(t))$, where $\lambda(t) = p \times t$.

Note: t is the number of initial inspections in that month in that year. p is the

probability of encountering rat activity for an initial inspection. k is the observed number of rat activities in that month.
I first found the MLE (Maximum Likelihood Estimator) for p. The likelihood of the data is:

$$L(D|p) = \prod f_x(X = k|p) \tag{1}$$

Take log on both side of the equation, we get:

$$\begin{aligned} logL(D|p) &= \sum log(f_x(X = k|p)) \\ &= \sum log(\frac{e^{-\lambda}\lambda^k}{k!}) \\ &= \sum log(e^{-\lambda}) + log(\lambda^k) - log(k!) \\ &= \sum -\lambda log(e) + klog(\lambda) - log(k!) \\ &= \sum -\lambda + klog(\lambda) - log(k!) \\ &= \sum -pt + klog(pt) - log(k!) \\ &= \sum -pt + klog(p) + klog(t) - log(k!) \end{aligned}$$

Take the derivative in terms of p, we obtain

$$\frac{d}{dp} logL(D|p) = \sum -t + \frac{k}{p} \tag{2}$$

Suppose $\frac{d}{dp} logL(D|p) = 0$. Then,

$$\begin{aligned} 0 &= \sum -t + \frac{k}{p} \\ \sum t &= \frac{1}{p}\sum k \\ p &= \frac{\sum k}{\sum t} \end{aligned}$$

Indeed, p is the average percentage of rat activities encountered per month from 2010 to 2014. Finally, I found the MLE for p is 0.113467. Therefore, the number of rat activities found per month is $X \sim Poisson(0.113467t)$.

3.1 Model Evaluation

The Fraction of Explained Variation is 0.7590249. Thus, the model explains the number of rat activities relatively well.

I compared the number of rat activities my model predicts and the actual number of rat activities.

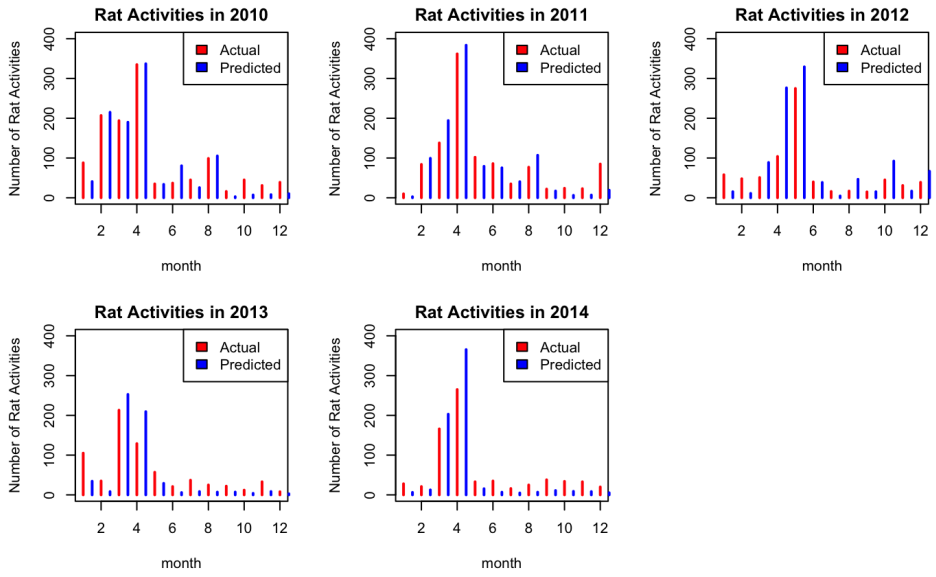

Figure 4: Comparison Between the Number of Rat Activities Observed and Predicted

The red lines and blue lines in general are closely aligned. However, there are some overpredictions and underpredictions. For example, from October to December in 2010, my model consistently predicts less than the actual number of rat activities. In April 2013, my model predicts significantly higher than the actual of rat activities.

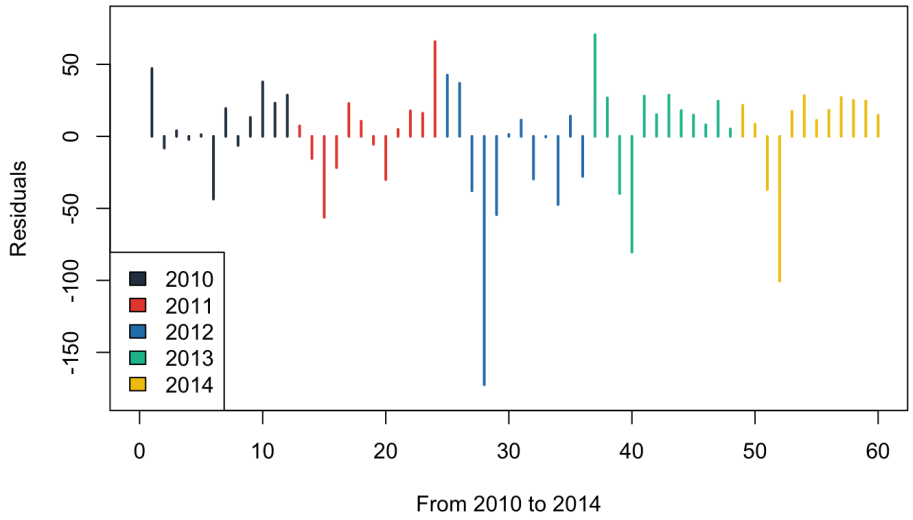

Figure 5: The Residuals of the Model

Based on figure 5, the residual is usually low but can be as high as around 150. Referring back to Figure 1, the overestimating happens at months with high levels of inspection. Underestimating happens at months with low levels of inspections. For example, October to December in 2010 have very low number of initial inspections. April in 2012 have high number of inspections. This further supports the finding in Figure 2, months with higher number of inspections have lower percentage of finding rat activities. Months with lower number of inspections have higher percentage of finding rat activities.

3.2 Model Conclusion

Months with high number of inspections have lower likelihood of encountering rat activity for each initial inspection, and vice versa for months with low number of inspections.

The overprediction and the underprediction might have been because the number of rats in New York City is relatively fixed at a given month. With more

inspections, it is harder to discover new rat activities. Also, in months with little inspections, the number of undiscovered rat activities is low, so it is more likely to find new rat activity.
Given Figure 4 and Fraction of Explained Variation, the model without accounting for the effects of number of inspection still predicts well the number of rat activities.
I cannot determine the monthly and yearly effects without accounting for the fluctuation of likelihood by the number of inspections.

4 Statement of Likelihoods

Null hypothesis: There is no correlation between the observed rat activities and the predicted number of rat activities based on my model. In other words, $R^2 = 0$. I conducted a bootstrap for 1000 iterations to find the 99% confidence interval for R^2 is 0.12 and 0.92. The graph is significantly skewed to the right. Therefore, there is correlation between the observed number of rat activities and the predictions by my model.

5 Limits to my investigation and Potential for Future Investigation

My model does not take into account how the number of investigations will affect the probability of finding rat activities. I would suggest future models to adjust the probability based on the number of investigations. Then, we can assess whether there is a monthly or yearly effect to the percentage

I also suggest exploring the JOB_PROGRESS column in the data. JOB_PROGRESS keeps track of the order of the inspection that took place. Future investigation can find the probability of finding rat activities for each JOB_PROGRESS and inves-

tigate where the probability decreases with higher JOB_PROGRESS.

Finally, I have observed large yearly and monthly fluctuations for the number of inspections. Future research can investigate the reasons behind the increase and decrease for the number of inspections throughout the years and months.

6 Appendix

6.1 Rough estimate of the area included in my study

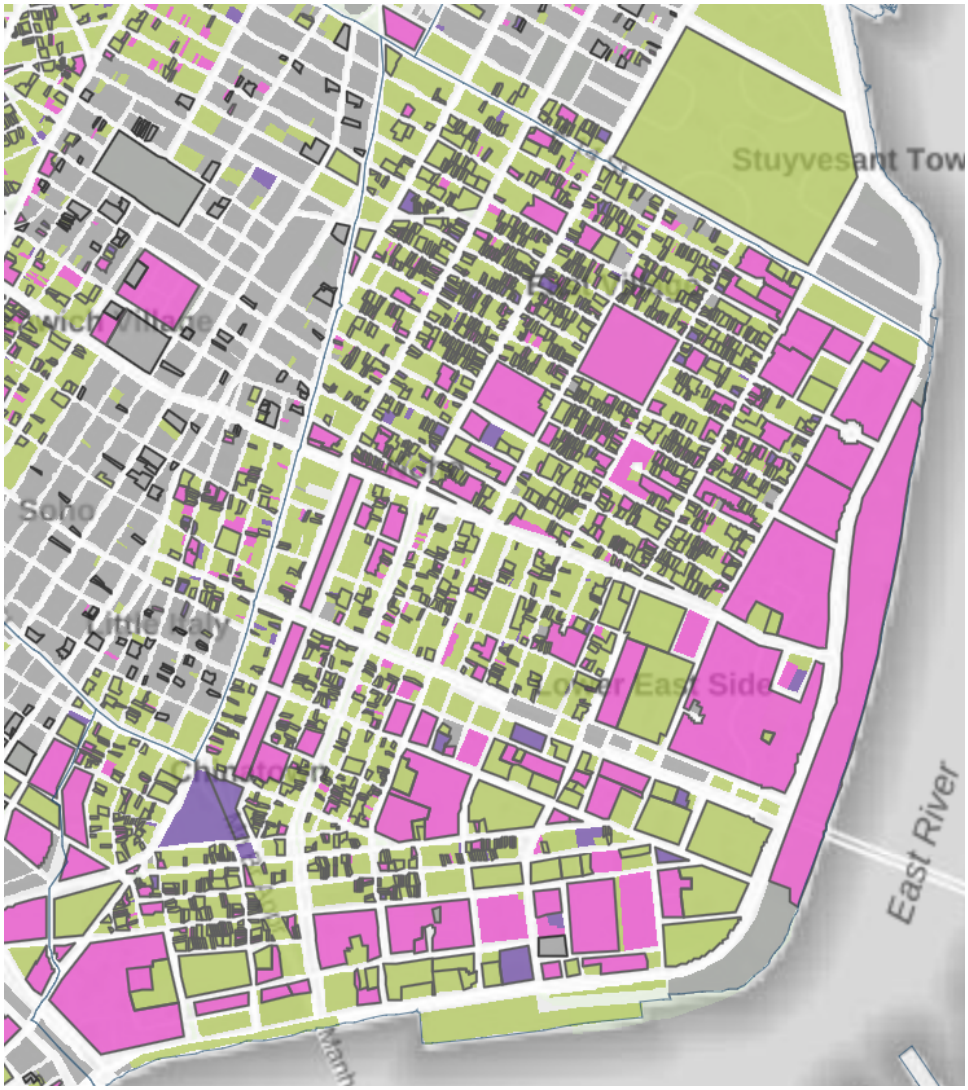

Figure 6: The area I attempt to investigate in this project.