Time Series Analysis on NYC 311 Request

Yang Kang Chua[a]

[a] *Univeristy of Connecticut, Mechanical Engineering Department, 191 Auditorium Rd. U-3139, Storrs, (06269)*

Abstract

This study aims to investigate the time series trend of noise complaints in New York City (NYC) using NYC 311 data. The research hypothesis is that noise complaints are similar across all boroughs, and the findings will be of interest to property developers and managers for making informed decisions about where to build new developments and how to mitigate noise impacts. The study employs various statistical methods, including time series plots, bar charts, and a Poisson regression model, to analyze the data and predict the number of noise complaints. The results of this investigation have the potential to benefit the real estate industry by identifying areas with high noise levels and taking measures to reduce noise pollution, making properties more appealing to potential buyers or tenants. The study concludes that analyzing noise complaints in time series data can provide valuable insights into the patterns of noise complaints in NYC and offer practical guidance for the real estate industry to develop properties with lower noise complaints.

Keywords: Time Series Analysis, NYC Open Data, Noise Complaints

1. Introduction

The impact of noise complaints on the property and real estate industry has garnered increasing attention in recent years. The volume of noise complaints within an area can have a significant effect on the value and selling potential of a property, making it a critical issue for real estate agents. In some cases, properties situated in areas with a high number of noise complaints may be difficult to sell, and real estate agents may struggle to identify locations with low noise complaint rates. To address this issue, there is a growing need for statistical analyses that can help identify patterns and trends in noise complaints throughout New York City, ultimately providing valuable insights for real estate professionals.

Extensive research has been conducted in the field of analyzing noise complaints, with most studies starting by analyzing the NYC 311 service request data. For instance, Fisher (2021) conducted a study on the noise complaints related to COVID-19 and developed a model that predicts the type of noise based on specific variables [1]. Such research allows real estate agents to identify specific types of complaints based on certain conditions. Additionally, Niki (2021) used NYC 311 data to perform an analysis on response time from the request. This information is beneficial for identifying areas that contain the fastest response time [2]. According to the research, Bronx had the quickest response time compared to all other boroughs. However, these studies are not the most updated, and it is essential to develop new statistical analyses to identify patterns related to noise complaints in New York City, particularly their impact on property values.

This study aims to investigate the temporal patterns of noise complaints and to develop a time series analysis to predict the future number of noise complaints in a given area. The results of this analysis will be a valuable tool for real estate agents, allowing them to make informed decisions about which neighborhoods

*Corresponding author
Email address: yang_kang.chua@uconn.edu (Yang Kang Chua)

to invest in and which to avoid. Additionally, the findings may enable agents to negotiate better deals for their clients. For instance, they may use the data to persuade sellers to lower their asking price or offer concessions to address noise issues. Ultimately, this research will provide a valuable contribution to the understanding of noise complaints and their impact on the real estate market.

In this paper, we will begin by providing an overview of the importance of noise complaints in the real estate industry, as well as a review of relevant literature on the topic. Next, we will describe the data used in this study and the methods we employed for our time series analysis. We will then present our findings, including an analysis of noise complaint trends and a prediction for future noise complaints in the selected areas. Finally, we will conclude with a discussion of the implications of our results for real estate agents and suggestions for future research in this area.

2. Data

In this study, we aim to investigate noise complaints in New York City from 2010 to 2022 by utilizing publicly available data from the NYC Open Data Portal [3]. The dataset provides comprehensive information on the location, date, and type of each complaint. We have focused our analysis on the time frame of 2021 to 2022, which comprises 1,507,956 observations and 41 variables of interest. These variables include complaint type, borough, latitude, and longitude, among others. The data is a combination of quantitative and categorical variables, including time series variables, such as created date, and nominal variables, such as borough.

To ensure the reliability of our analysis, the data underwent thorough cleaning and preprocessing. We removed duplicate entries, handled missing values, and aggregated the data by year and borough. Additionally, we merged the dataset with temperature data, which contains four variables of interest, namely average temperature, maximum temperature, minimum temperature, and snow precipitation. Furthermore, we incorporated zipcode-level data to obtain population density, home value, and household income. Other relevant data, such as whether it was a holiday or a weekend, was also included to aid our analysis.

This comprehensive dataset provides us with an excellent opportunity to conduct a detailed analysis of noise complaints and their patterns over a year. The additional data that we have incorporated will enable us to draw more informed conclusions and provide greater insights into the relationships between noise complaints and various socio-economic factors. Figure 1 show the plot of the full data related to noise complaints.

3. Research Design and Methods

In this study, we aim to analyze the pattern of noise complaints in New York City from 2021 to 2022 using a time series analysis approach. To begin, we will examine the descriptive statistics of the data and visualize the trends and patterns using graphical tools such as line plots and bar plots.

Our analysis will include hypothesis testing on the time series to study whether noise complaints are similar across all boroughs. We will utilize Poisson regression and negative binomial regression models to analyze the data and predict the number of noise complaints for one month [4]. These two models will be compared to discuss the performance of the model, and to provide insights into the factors that contribute to noise complaints and how they can be mitigated in the future.

Our study will provide valuable insights for real estate developers to identify locations where noise complaints are high and plan new developments accordingly. The introductory model will be explained below to provide a clear understanding of the methodologies used in our analysis.

All analyses are performed using the Python programming language and relevant packages such as pandas, statsmodels, and scikit-learn.

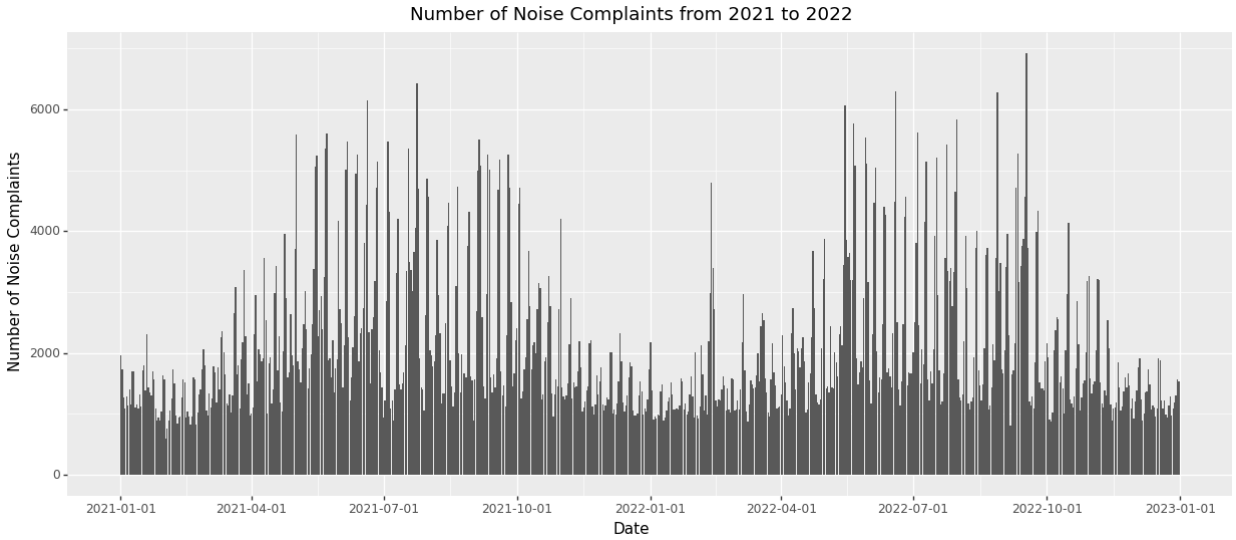

Figure 1: Noise Complaint by year

3.1. Poisson Regression Model [4]

The Poisson distribion has the following Probabilty Mass Function:

$$P_X(k) = \frac{e^{-\lambda t} * (\lambda t)^k}{k!} \tag{1}$$

Where $P_X(k)$ is the probability of seeing k events in time t, and λ is the number of events per unit time. The expected value (mean) for a poisson distribution is λ.

When the event rate λ is constant, the predicted values will also be constant and equal to λ. However, in real-world scenarios, the event rate is influenced by a vector of explanatory variables, also known as **predictors**, **regression variables**, or **regressors**. This matrix of regression variables is denoted by X. The goal of the regression model is to fit the observed count y to the matrix X. This requires estimating the values of a vector of regression coefficients β, which will be used to model the relationship between the dependent variable and the explanatory variables.

The link function below works great for possion regression model because it keep λ non-negative even when the regressors X or the regression coefficient β have negative values.

$$\lambda = e^{\mathbf{X}\beta} \tag{2}$$

After the model is trained on the data set, the regression coefficients β are determined, and the model is then used to make predictions. To predict the event count y_p associated with an input row of regressors x_p, we use the following formula:

$$y_p = \lambda_p = e^{x_p\beta} \tag{3}$$

The regression coefficients β can be estimated using maximum likelihood estimation (MLE), which involves finding the values of β that maximize the likelihood function given the observed data.

3.2. Negative Binomial Regression Model

Negative binomial regression is a type of regression analysis that is useful for modeling count data when the variance of the dependent variable is greater than the mean, violating the assumptions of the Poisson regression model. It is a generalization of Poisson regression, which assumes that the variance of the dependent variable is equal to its mean.

In this research, we will be using NB2 model where the variance formula are shown below:

$$Variance = mean + \alpha * mean^2 \tag{4}$$

$$\frac{(y_i - \lambda_i)^2 - \lambda_i}{\lambda_i} = \alpha\lambda_i \tag{5}$$

We estimated α through an auxiliary OLS regression using Equation 5, using λ values obtained from the training results of the Poisson regression model. The NB2 model is widely used in fields such as public health, ecology, and economics to model count data, such as the number of hospital admissions, species in a particular area, or sales within a given time frame. It is also applied in social science research to model voting behavior, crime rates, and other phenomena that can be counted.

4. Results

Our research paper delves into the distribution of noise complaints across different boroughs in New York City. To begin with, we analyzed the number of noise complaints by borough as shown in Figure 2, and found that Staten Island had the lowest number of complaints while the Bronx had the highest. This initial analysis provided valuable insight into the overall distribution of noise complaints across the boroughs, although it is important to note that factors like population density, neighborhood types, and traffic levels could also influence these results. Nonetheless, this analysis laid the foundation for further investigation.

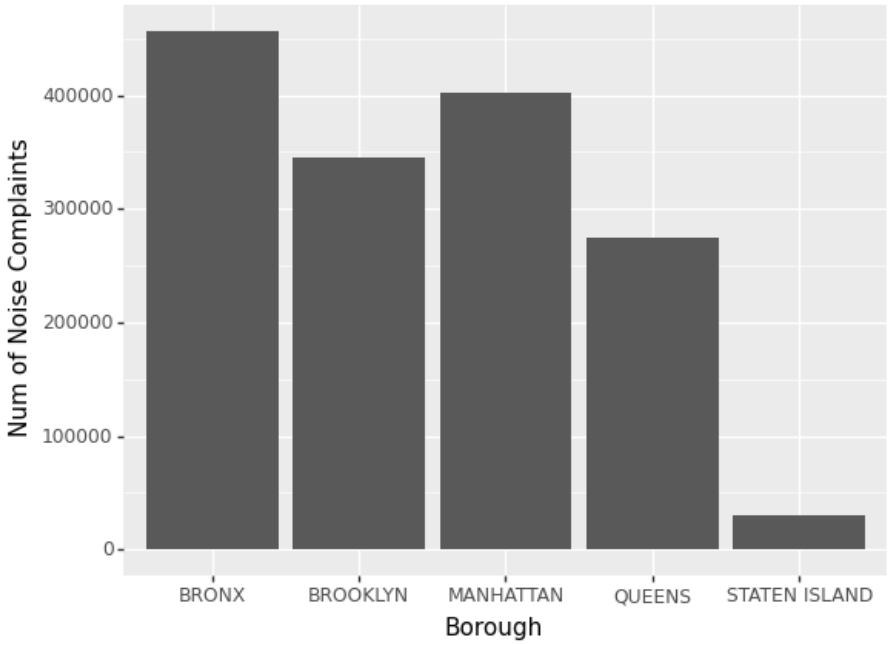

Figure 2: Noise Complaint by brough

We then explored the hypothesis based on the distribution of the Number of Complaints across weekdays and weekends, as shown in Figure 3. Our null hypothesis was that there is no difference in the distribution of the Number of Complaints between weekdays and weekends. To test this, we employed the Mann-Whitney U test and found a p-value of less than 0.05. This result led us to reject the null hypothesis and conclude that the distribution of the Number of Complaints is indeed different between weekdays and weekends.

Furthermore, we also tested the null hypothesis that there is no significant median difference in the distribution of the Number of Complaints across the borough. We used the Kruskal-Wallis test for this purpose and found that the p-value was less than 0.05. As a result, we rejected the null hypothesis and concluded that there is a significant difference in the median Number of Complaints across boroughs.

Overall, our research provides valuable insights into the distribution of noise complaints across different boroughs in New York City, and highlights the importance of considering factors like weekdays vs. weekends and boroughs when analyzing noise complaints.

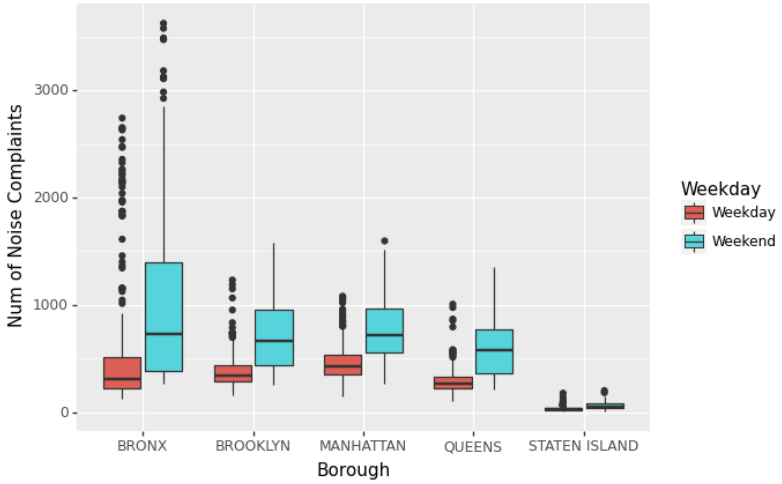

Figure 3: Noise Complaint by brough by week type

In this research, we analyze the performance of Poisson regression and negative binomial regression models on time series data for different boroughs. First, we evaluate the performance of the Poisson regression model on the Bronx data set, as shown in Figure 4a. The Deviance and Pearson chi-squared values reported for the model are extremely high, indicating a poor fit to the training data. We use the 2 table to assess the goodness-of-fit at a 95% confidence level and find that the Poisson regression model has fit the training data poorly.

Next, we compare the performance of the NB2 model to the Poisson regression model for the Bronx data set. The NB2 model, as illustrated in Figure 4b, reports Deviance and Pearson chi-squared values of 515.14 and 584, respectively, indicating a better fit than the Poisson regression model. However, the chi-squared table shows that the model is still sub-optimal.

We visualize the comparison between the predicted and actual number of complaints in the Bronx for the month of January in Figure 5. This allows us to gain insight into the performance of the model and identify any potential areas for improvement.

To quantitatively compare the performance of the NB2 and Poisson regression models, we use the Likelihood-ratio test. The LR test statistic, which is calculated as negative two times the difference in the fitted log-likelihoods of the two models, shows that the trained NB2 regression model has demonstrated a much better goodness-of-fit than the Poisson regression model on the bicyclists data set.

Generalized Linear Model Regression Results

Dep. Variable:	Noise_COUNT	No. Observations:	730
Model:	GLM	Df Residuals:	718
Model Family:	Poisson	Df Model:	11
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-37311.
Date:	Thu, 20 Apr 2023	Deviance:	68823.
Time:	13:50:45	Pearson chi2:	7.47e+04
No. Iterations:	5	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.7669	0.154	-17.952	0.000	-3.069	-2.465
is_holiday	0.3353	0.007	48.566	0.000	0.322	0.349
MONTH	-0.0338	0.001	-54.756	0.000	-0.035	-0.033
DAY_OF_WEEK	0.0207	0.001	15.239	0.000	0.018	0.023
DAY	0.0035	0.000	20.362	0.000	0.003	0.004
Day_Type	0.6025	0.006	107.640	0.000	0.592	0.614
TAVG	0.0079	3.93e-05	201.076	0.000	0.008	0.008
TMAX	0.0103	0.000	32.696	0.000	0.010	0.011
TMIN	0.0055	0.000	16.248	0.000	0.005	0.006
PRCP	-0.0957	0.003	-28.230	0.000	-0.102	-0.089
Population	-1.12e-06	3.77e-07	-2.970	0.003	-1.86e-06	-3.81e-07
Home_Value	1.544e-05	3.95e-07	39.120	0.000	1.47e-05	1.62e-05
House_Income	5.241e-05	6.16e-07	85.137	0.000	5.12e-05	5.36e-05

(a) Poisson Regression model for Bronx

Generalized Linear Model Regression Results

Dep. Variable:	Noise_COUNT	No. Observations:	730
Model:	GLM	Df Residuals:	718
Model Family:	NegativeBinomial	Df Model:	11
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-4739.9
Date:	Thu, 20 Apr 2023	Deviance:	515.14
Time:	13:50:45	Pearson chi2:	584.
No. Iterations:	8	Pseudo R-squ. (CS):	0.9637
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6936	1.044	-0.665	0.506	-2.739	1.352
is_holiday	0.3786	0.079	4.794	0.000	0.224	0.533
MONTH	-0.0306	0.005	-6.253	0.000	-0.040	-0.021
DAY_OF_WEEK	0.0299	0.012	2.408	0.016	0.006	0.054
DAY	0.0019	0.002	1.110	0.267	-0.001	0.005
Day_Type	0.5969	0.055	10.894	0.000	0.490	0.704
TAVG	0.0075	0.000	21.160	0.000	0.007	0.008
TMAX	0.0148	0.003	4.982	0.000	0.009	0.021
TMIN	0.0002	0.003	0.078	0.938	-0.006	0.006
PRCP	-0.1333	0.034	-3.905	0.000	-0.200	-0.066
Population	2.459e-06	3.41e-06	0.721	0.471	-4.22e-06	9.14e-06
Home_Value	8.725e-06	2.66e-06	3.277	0.001	3.51e-06	1.39e-05
House_Income	6.259e-05	5.76e-06	10.869	0.000	5.13e-05	7.39e-05

(b) Negative Binomial Regression Model for Bronx

Figure 4: Model Result Report

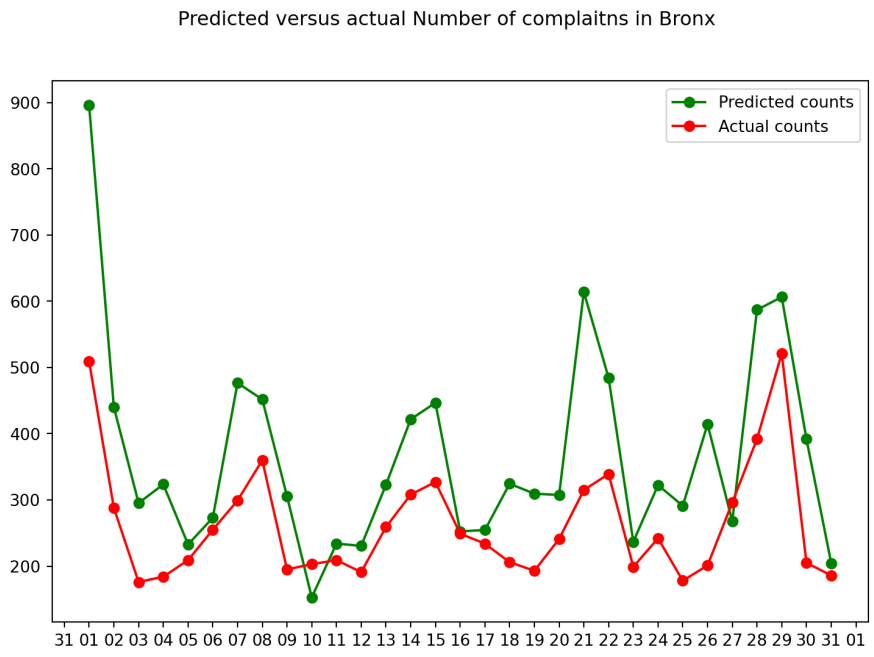

Figure 5: Predicted vs actual result for Bronx

Overall, our research demonstrates that the negative binomial regression model is a better choice than the Poisson regression model for modeling the time series data for noise complaints in the Bronx. Further details and model reports are available in the github repository.

5. Conclusions

In conclusion, this research has shed light on the trends of noise complaints in New York City, providing insights on the number of complaints per borough, the impact of weekdays versus weekends, and the difference in median complaints across boroughs. Our findings suggest that the Negative Binomial Regression model is more effective than the Poisson Regression model for predicting future complaints. The real estate industry can use this information to make properties more appealing to potential buyers or tenants by implementing strategies to reduce noise pollution. Overall, this study has contributed to a better understanding of noise complaints in New York City and can guide the development of properties with lower noise complaints.

References

[1] A. Fisher, Analyzing and modelling nyc 311 service requests (Jan 2021).
URL https://towardsdatascience.com/analyzing-and-modelling-nyc-311-service-requests-eb6a9c9adc7c

[2] A. Niki, Detailed data analysis: The rise of nyc 311 noise complaints (Dec 2021).
URL https://nycdatascience.com/blog/student-works/detailed-data-analysis-the-rise-of-nyc-311-noise-complaints/

[3] Nyc 311 open data, https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9, accessed: [Mar 23 2023] (2023).

[4] S. Date, The poisson regression model (Nov 2022).
URL https://timeseriesreasoning.com/contents/poisson-regression-model/