Time-Series Analysis / Statistical Inference
Group Project | Research Project
December 22nd, 2023

An Empirical Application of Spectral Analysis: Fourier Series and Transforms to Predict Noise Complaints in New York City

Professor Yisha Yao
Federico Panariello, fep2120
Daniele Prevedello, dp3252

Contents

Research Question

We employ spectral analysis for the examination and decomposition of a time series dataset encompassing noise complaints recorded in New York City spanning from 2010 to 2020. Subsequently, we develop a predictive model capable of forecasting the volume of noise complaints in an out-of-sample dataset, covering the period from 2021 to 2023. Our underlying research hypothesis posits that the application of spectral analysis, along with associated techniques, will enable us to disentangle the inherent seasonality within the time-series data.

Our overarching goal is to make future predictions regarding noise complaints based on a set of observable independent variables readily available within the selected dataset. To evaluate the performance of our model, we employed the Mean Squared Error (MSE) as a metric during out-of-sample testing. The MSE serves as a measure of the accuracy of our forecasts. When needed, we adopted alternative measures, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

We expect our model to be of significant utility to professionals in fields such as social security, public infrastructure, and urban planning. These sectors often require insights into noise complaint trends and their correlation with specific neighborhoods. Furthermore, potential applications extend to individuals and corporations contemplating relocation to New York City. It provides them with a valuable tool to consider noise complaint levels when making decisions about their ideal residential or business locations.

Resources Allocation

To retrieve the data and run a preliminary analysis we employed 10 hours in total. After that, we performed model specification for 5 hours and 5 hours for model training, cross-validation and out-of-sample testing. Lastly, we spent 7 hours drafting the final report and 10 hours preparing the presentation deck to be shared with the class.

Dataset

NYC 311 Open Data on Noise Complaints is a subset of the larger NYC 311 Open Source Data platform that specifically focuses on complaints related to noise disturbances within New York City. It provides detailed information about noise complaints made by residents, including the type of noise, location, date, and time of the complaint. Users can freely access this open-source data to gain insights into noise-related problems, allocate resources efficiently, and develop targeted solutions to mitigate noise pollution.

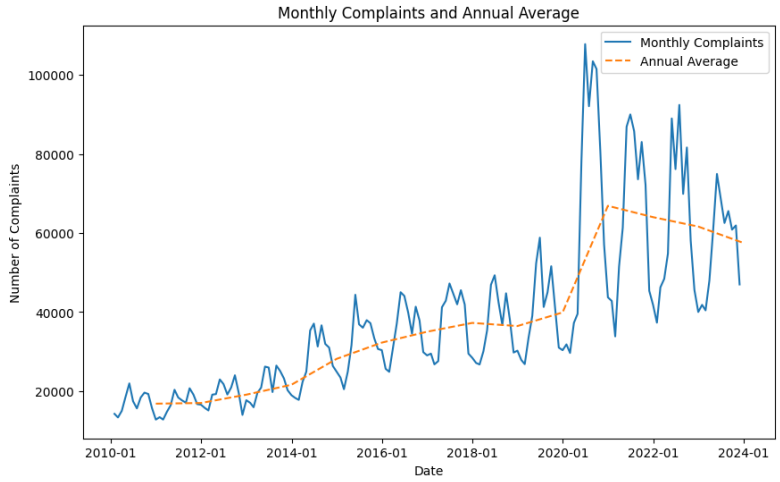
Exploratory Analysis and Visualization

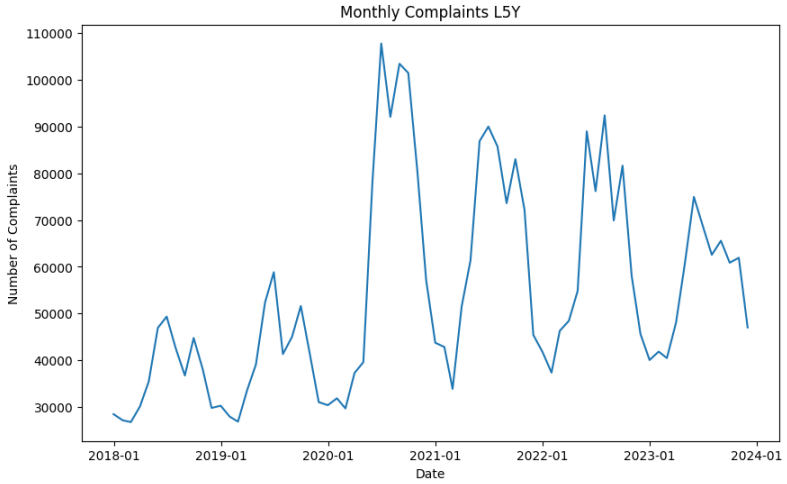
From an initial analysis, the daily, weekly and monthly frequencies of noise complaints across the entire dataset –notwithstanding of geographical location– exhibit significant seasonality, which limits statistical inference through conventional Auto-Regressive Moving-Average (ARMA) and Vector Auto-Regressive (VAR) methods alone.

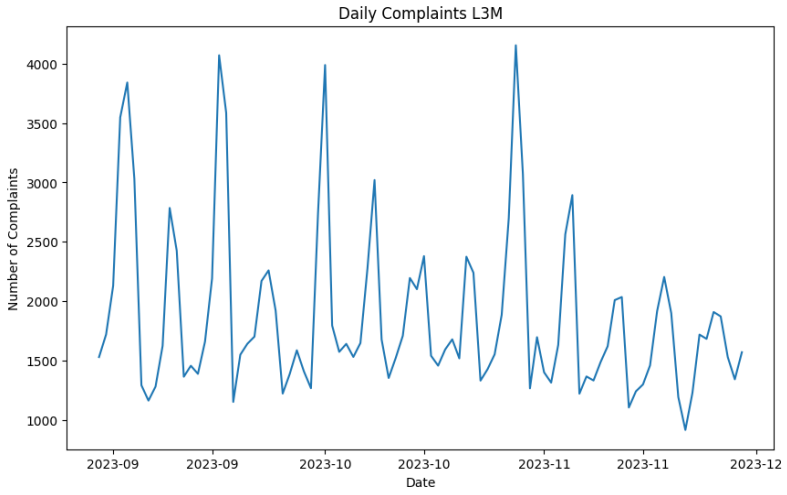
The following plots highlight intuitive trends, namely (i) a daily seasonality, which entails higher frequency of complaints on Fridays, Saturdays and Sundays, and (ii) a monthly seasonality, which is reflected in more complaints during summer months.

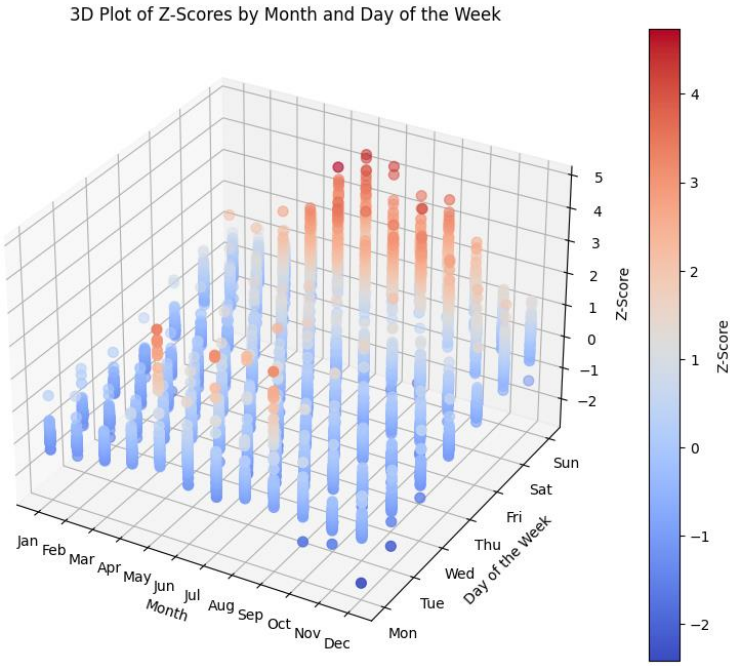
Moreover, the average monthly complaints drift positively over time, with a substantial spike after 2020, likely driven by the impact of Covid-19. The standardized 3D plot summarizes the above-mentioned findings, where each daily noise complaints frequency has been standardized vis-à-vis its annual average.

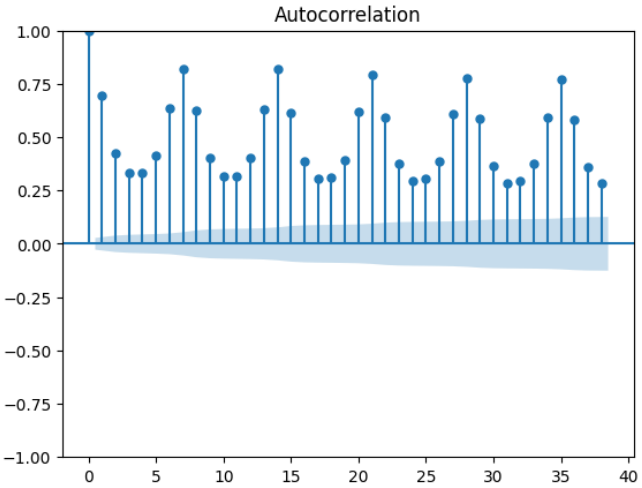
From a geographical standpoint –though not the focus of this research project– total frequency over the last twelve months appears to be the highest in some areas of the Upper West Side, Financial District and Uptown as far as Manhattan is concerned. As for the broader NYC area, some residential areas of Brooklyn, Rockaway Beach and Staten Island add on top of the west side of Manhattan as the higher frequency neighborhoods.

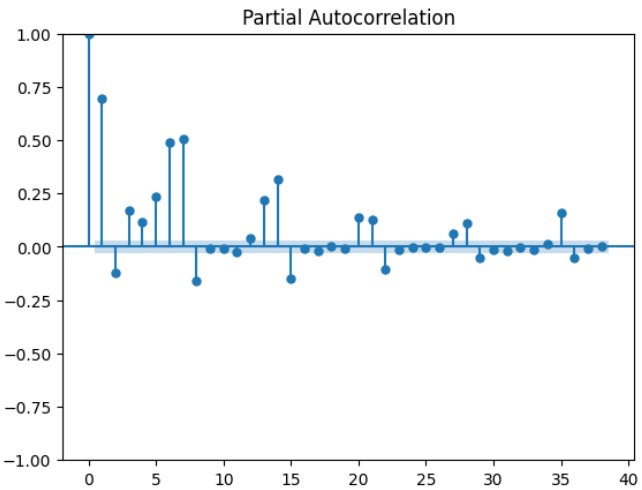

Monthly Complaints and Annual Average
Monthly Complaints
Annual Average
Number of Complaints
Date


Monthly Complaints L5Y
Number of Complaints
Date

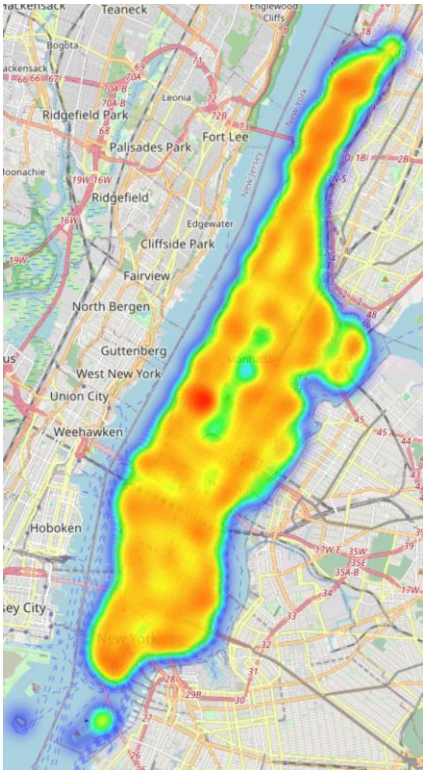

Daily Complaints L3M
Number of Complaints
Date

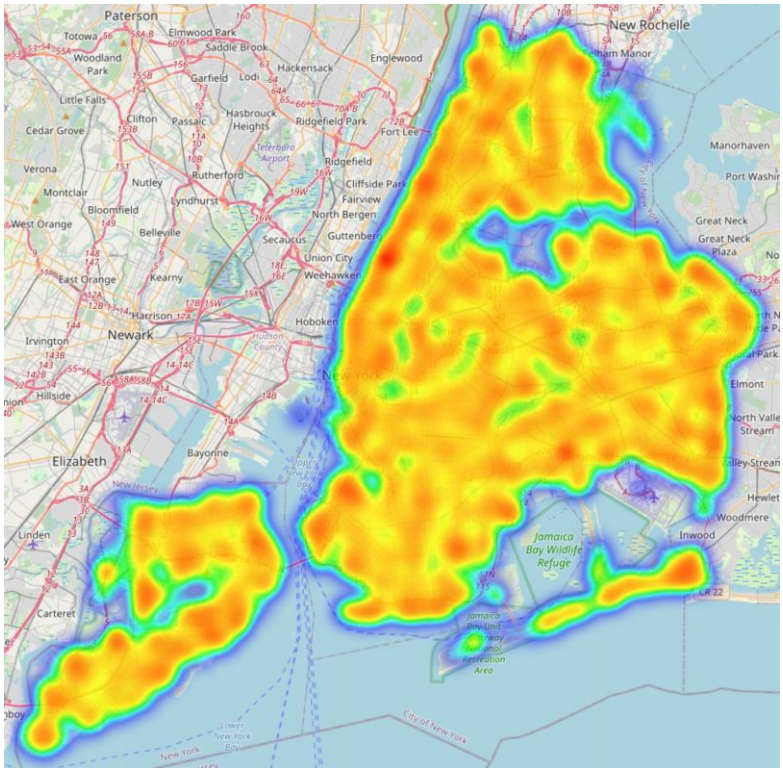

3D Plot of Z-Scores by Month and Day of the Week
Z-Score
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of the Week


Autocorrelation


Partial Autocorrelation

Geographical Locations of Last Twelve Months Noise Complaints in Manhattan and NYC





Literature Review

"From Fourier to Koopman: Spectral Methods for Long-term Time Series Prediction" by Henning Lange, Steven L. Brunton, and J. Nathan Kutz, published in 2021, discusses spectral methods for forecasting long-term temporal signals from both linear and nonlinear quasi-periodic dynamical systems.

"An Approach to Enhance Time Series Forecasting by Fast Fourier Transform," examines the Fast Fourier Transform (FFT) and its capacity to extract frequency-domain features from time series data, exploring its potential for feature engineering to enhance the accuracy and efficiency of time-series forecasting models.

“A Survey on Deep Learning based Time Series Analysis with Fourier Transforms" conducted by Kun Yi et al., published in 2023, explores the advantages of Fourier Transform (FT) in time series analysis, highlighting its efficiency and global perspective in the context of deep learning.

"An Approach to Enhance Time Series Forecasting by Fast Fourier Transform", authored by José María Luna-Romera, was first online on 31 August 2023. It emphasizes feature engineering in time series forecasting, focusing on the use of Fast Fourier Transform (FFT).

Time-Series Modelling Alternatives

Introduction

The above-mentioned characteristics of this time series demand for more sophisticated modeling techniques in order to capture both the drift and the short-term and long-term seasonality. We plan on running Spectral Analysis on the dataset and using a selected subsect of its techniques for studying the frequency content and properties of signals, which include Dynamic Harmonic Regression and Fast Fourier Transform (FFT).

All alternative models are trained on the common training set from November 2010 to November 2020 and tested from December 2020 to November 2023. As previously highlighted, the primary measure used to evaluate the models is the Mean Squared Error (MSE).

Fast Fourier Transform (FFT) with ARIMA-fitted Residuals

Fast Fourier Transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. FFT is a mathematical tool used for analyzing and processing signals and time-series data in various fields, including signal processing, image analysis, audio processing, and more.

The DFT is a mathematical operation that transforms a time-domain signal into its frequency-domain representation. By converting the signal into its constituent frequencies, it reveals the underlying frequency components. The DFT of a signal results in a complex number, where the magnitude represents the amplitude of the frequency component, and the phase represents the phase angle of that component.

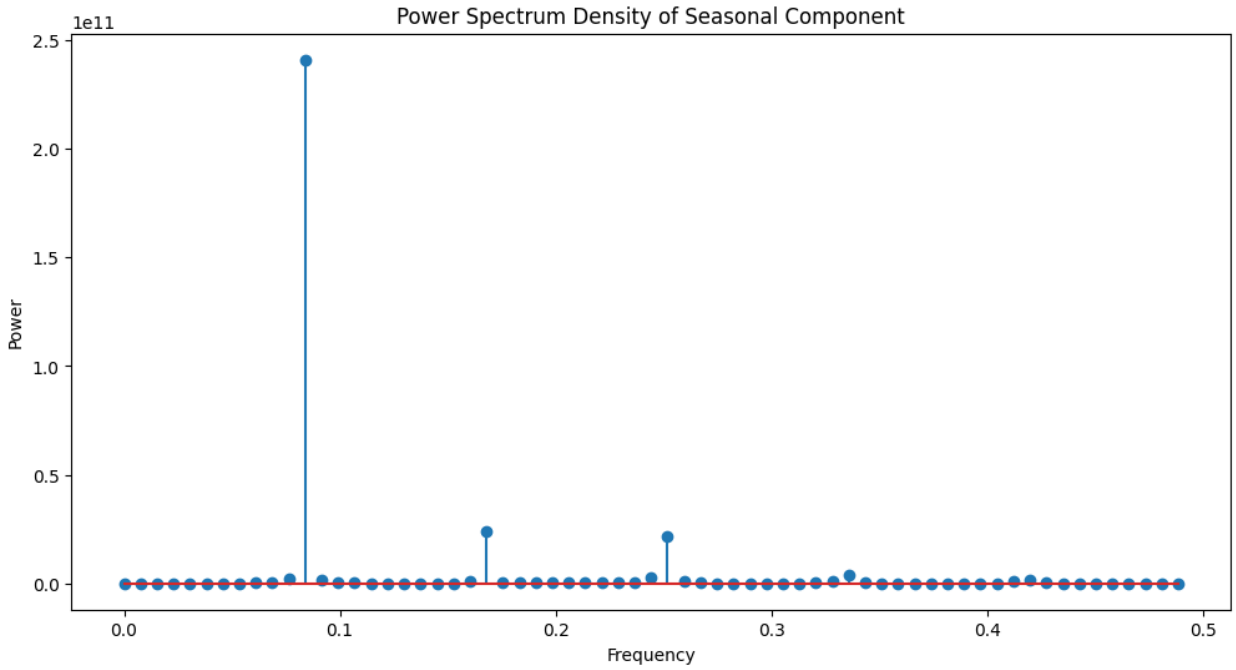
$$DFT := S(k) \;=\; \sum_{t=0}^{T-1} s(n) e^{-\frac{j2\pi kn}{N}} = \sum_{t=0}^{T-1} s(n) \left[\cos\left(\frac{2\pi Kt}{T}\right) + isin(\frac{2\pi Kt}{T})\right]$$

Where $S(k)$ is the *k*-th Fourier Coefficient and i is an imaginary number ($i = \sqrt{-1}$), meaning that each coefficient is a complex exponential with a real and imaginary component containing the amplitudes of the underlying harmonics to the observed wave.

The applicable use of the DFT in our model is to fit the month-of-the-year seasonal component. To pursue this, we first performed a seasonal decomposition of the training data. Seasonal decomposition separates a time series into three components: trend, seasonal, and residual of the following form:

$$y(t) = g(t) + s(t) + \varepsilon(t)$$

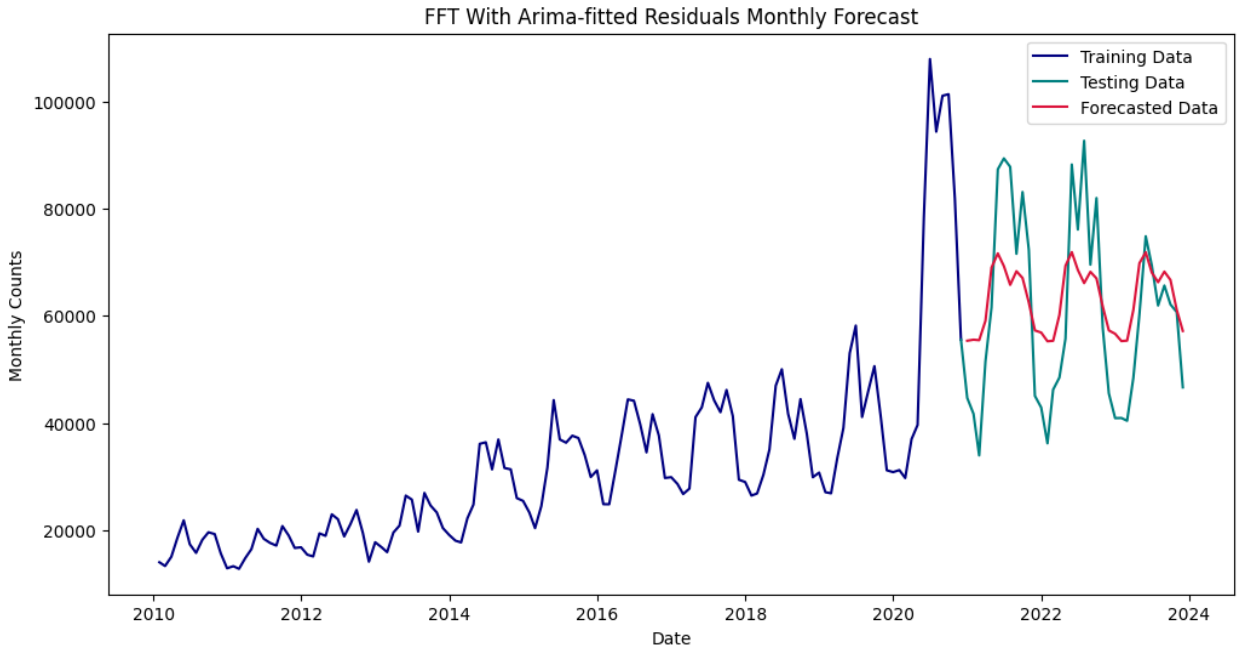
Subsequently, we used an additive model with a seasonal period of 12 to resemble the observed seasonality. At this stage, we applied the FFT to the seasonal component and computed the Power Spectrum Density (PSD) of the transformed results to identify the dominant frequencies of noise complaints seasonal patterns. To filter statistically significant frequencies, we applied a threshold of 5% of the maximum PSD.



On the x-axis it is shown the frequency, which represents the number of cycles per time unit. This is limited to the Nyquist frequency, defined as half the sampling rate of 1 per month of the monthly time series. The plot highlights a significant peak at the frequency of *0.08 cycles per month*. The periodicity T in months is given by $T = \frac{1}{f}$ where f is the frequency. The peak frequency corresponds to a periodicity of approximately *12.5 months*, confirming our hypothesis of a yearly cycle due to temperature and season affecting the level of noise complaints.

After having defined the model for the seasonal component, we aggregated trend plus residuals and used an ARIMA (2,1,1) to account for non-stationary trend with significant partial autocorrelation at lag 2 and autocorrelation at lag 1.

Lastly, we combined forecasted trend, seasonality and residuals and assessed the model fit with out-of-sample data by calculating the Mean Squared Error of 166,562,616.



Dynamic Harmonic Regression (DHR) with ARIMA-fitted Residuals

Dynamic Harmonic Regression is a technique that combines linear regression with harmonic components to capture both trend and seasonality in the data. The intuition behind using sine and cosine functions as harmonic components lies in their ability to model periodic patterns, such as seasonality in time series data. Once again, we decomposed the time series in trend, seasonality and residuals. The purpose of Dynamic Harmonic Regression is to aid in estimating the seasonal component $s(t)$.

$$y(t) = g(t) + s(t) + \varepsilon(t)$$

To model $s(t)$ we adopt a linear model of the form $s(t) = X(t)\beta$ where $X(t)$ is a $K\ x\ N, k \in K$ matrix of sines and cosines that, contrary to the DFT, requires to be determined ex-ante or through cross-validation. The intuition is that an order *k* of harmonics of increasing frequency approximates a Fourier series, resembling the periodicity of the seasonal component. As a result:

$$s(t)\ \approx a_0 + \sum_{k=1}^{K} y_t \left[a_k \cos\left(\frac{2\pi KT}{N}\right) + b_k sin(\frac{2\pi KT}{N})\right]$$

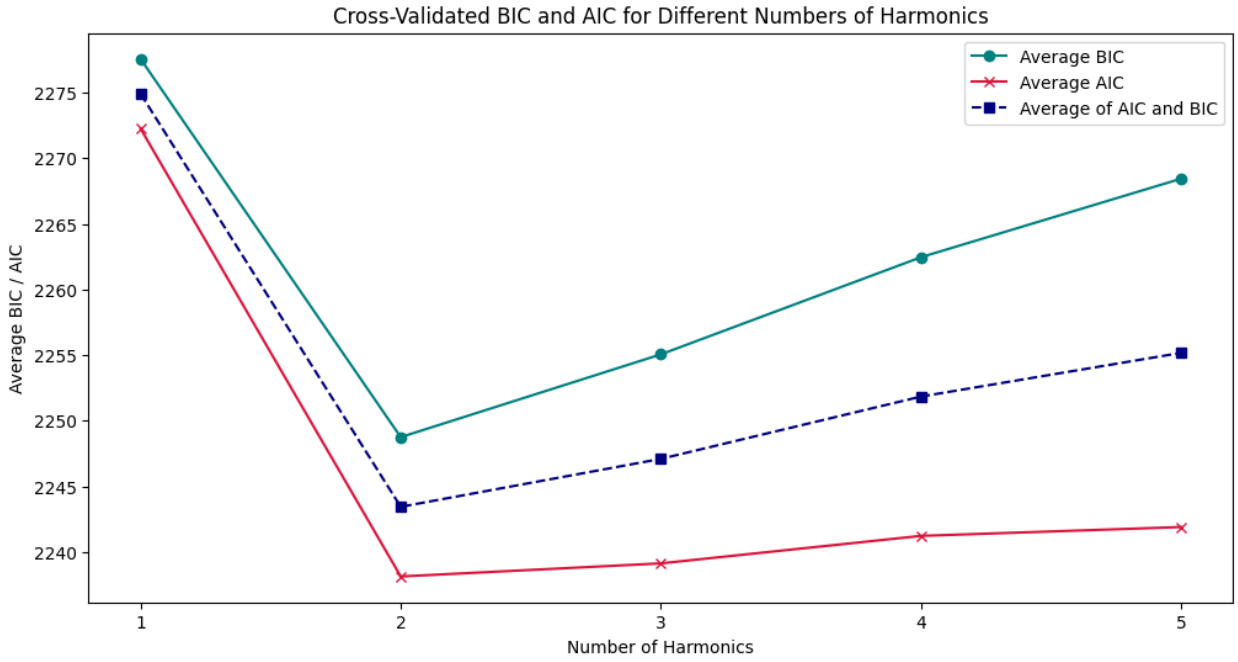
Guiding the choice of the number of harmonics to include to estimate the seasonality components are the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC), which impose penalties on the number of parameters to estimate while minimizing the MSE.

$$AIC = 2k - 2\ln(\hat{L}) \qquad BIC = k\ln(n) - 2\ln(\hat{L})$$

Where $(\hat{L})$ is the likelihood.

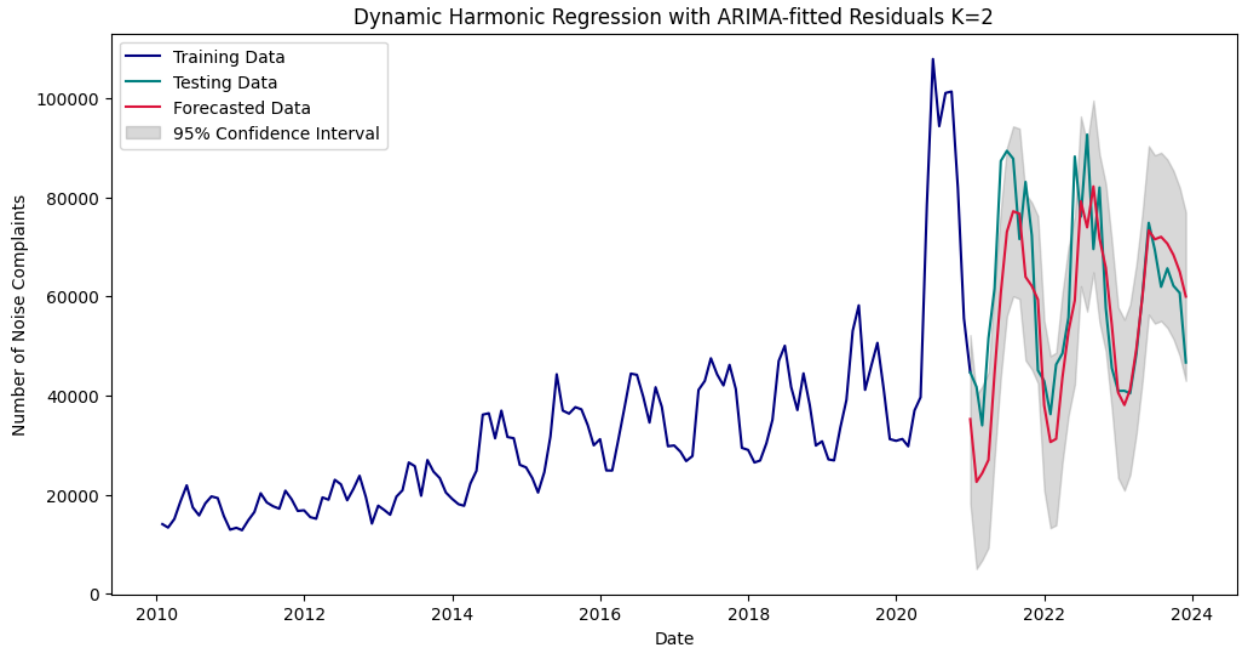
The number of harmonics *K* is a hyperparameter and needs to be chosen using either (i) in-sample fit, (ii) using a stored validation set or (iii) via cross-validation. We adopted the latter by bootstrapping 40

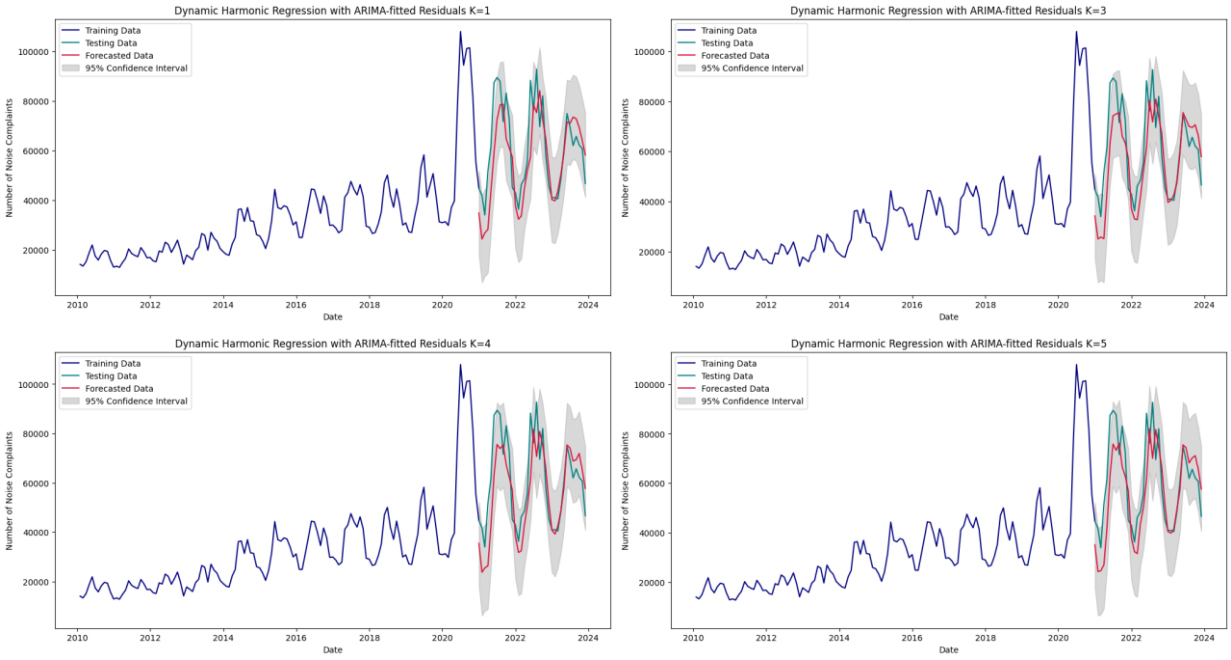
random samples and testing the model fit retrieving BIC and AIC values. The plot below shows that *K=3* harmonics minimizes both average AIC and BIC.



Subsequently, we specified the periodicity of the seasonality cycle equal to 12 months, which resembles the results from the DFT. Following the seasonality component definition, similarly to the DFT case, we aggregated trend plus residuals and estimated an ARIMA (2,1,1) model.

Lastly, we combined forecasted trend, seasonality and residuals and assessed the model fit with out-of-sample data by calculating the Mean Squared Error of 153,854,172.





Conclusions

To summarize, dealing with complex periodic and seasonal datasets requires sophisticated modelling techniques to be specified ad-hoc on time-series-specific characteristics. Discrete Fourier Transform and Fast Fourier Transform as well as Dynamic Harmonic Regression are valuable tools to properly model the seasonality component after the time-series decomposition. This results in composite specifications including ARIMA models to capture trend and residuals. The two approaches discussed exhibit a unique set of advantages and disadvantages, which we summarize in detail below.

DFT (and FFT) is a powerful tool that facilitates the decomposition of the time series data into its constituent frequency components, unveiling underlying periodic patterns. Analyzing the frequency domain is crucial to distinguish real cyclicality and seasonality from noise. However, when employed for forecasting, this technique is very sensitive to the thresholding level for significant frequencies and lacks performance vis-à-vis more conventional seasonal autoregressive models.

On the other hand, Dynamic Harmonic Regression allows explicitly model the seasonal components, providing enhanced control and interpretability. It is flexible in the choice of hyperparameters based on metrics such as AIC and BIC, allowing for precise forecasts through cross-validation of the number of harmonics. It performs well out-of-sample when the number of harmonics is well specified. At the same time, it requires in-depth analysis of the time-series to understand potentially multiple seasonal periodicities. Exploratory analysis plays a significant role in the choice of the hyperparameter K which is fundamental for model performance – when the number of harmonics is too low, the model is not complex enough and lags conventional seasonal autoregressive models, while when K is too high, this setup is prone to overfitting and loses out-of-sample performance. Moreover, setting aside a validation subset is not feasible with limited data and cross-validating K is computationally expensive vis-à-vis techniques that do not require hyperparameters tuning.

We propose that both instruments should be employed in conjunction: DFT (and FFT) to conduct thorough analysis in the frequency domain and discern significant seasonal patterns and their periodicity, and Dynamic Harmonic Regression to maintain control over the number of parameters and explicitly model the seasonal component.

Further Research

There exist promising avenues for further exploration and development of the 311 NYC dataset. One potential direction involves conducting classification analysis to delve into time-series variations across different boroughs within New York City. This would enable a more granular understanding of noise complaint trends and their geographical disparities, offering valuable insights for targeted interventions and policy decisions.

Furthermore, there is an opportunity to devise an effective method for forecasting the daily frequency of noise complaints. Currently, the research primarily focuses on monthly predictions. Extending the forecasting capabilities to a daily level would empower authorities and stakeholders to respond promptly to fluctuations in noise complaints, leading to more efficient resource allocation and improved noise pollution management.

Additionally, there is potential to leverage the frequency of noise complaints as a proxy for estimating non-observable variables using state space models. This approach could uncover hidden patterns and relationships within the dataset, shedding light on factors influencing noise complaints that might not be directly measurable. Such an application of state space models has the potential to enhance our understanding of the broader urban environment and its impact on noise-related issues, paving the way for more comprehensive and data-driven urban planning and policy-making.

These potential research extensions demonstrate the versatility and applicability of spectral analysis techniques and time-series modeling in addressing complex urban challenges related to noise complaints and public well-being.

References

[1] “A Time Series Technique Fourier All Seasons.” | (Lawrence R. de Geest. LDG).
https://lrdegeest.github.io/blog/fourier-series

[2] “Istanbul Weather Forecast Using SARIMA.” | (n.d.). https://github.com/cnzdgr/Weather-Forecast/commits?author=cnzdgr

[3] “Time-Series Forecasting Using TBATS model” | (n.d.). https://medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model-ce8c429442a9

[4] “If History Repeats Itself, Fourier Transform is a Key” | (Piero Paialunga). https://towardsdatascience.com/if-history-repeats-itself-fourier-transform-is-a-key-a593ddfa246e

[5] “Time Series Analysis” | (Michael Foley). https://bookdown.org/mpfoley1973/time-series/dynamic-harmonic-regression.html

[6] “Forecasting: Principles and Practice.” 3rd ed. Otexts | (Rob J Hyndman, George Athanasopoulos.). https://otexts.com/fpp3/.