Analyzing NYC Property Sale Using Random Forest and VARIMA

SI 671 Final Project Report

Yifei Sun

Abstract:

This project focuses on predicting housing prices through spatiotemporal methods, considering the dynamic interaction of time, location, and property-specific attributes. Leveraging a dataset spanning 2003 to 2023 from the NYC government website, including property category, sale date, price, and various features, the analysis employs techniques such as Fast Fourier Transform, VARIMA models, and Random Forest. Results highlight the significance of spatiotemporal considerations in housing predictions, influencing decision-making for stakeholders and aiding governmental policy-making. The study explores correlations between sale numbers and prices, revealing collaborative trends. Despite challenges like data discrepancies, the project provides valuable insights for the real estate industry, contributing to transparency and efficiency.

Code link: https://github.com/YifeiSun01/SI-671-Project

1. Motivation

Predicting housing prices using spatiotemporal methods is vital due to the intricate interplay of three key factors: time, position, and property-specific attributes. The temporal aspect, including sale dates, reveals market trends and historical context. Spatial features, like latitude and longitude, offer insights into surroundings, encompassing school districts, crime rates, and overall livability. Property-specific attributes—size, condition, type, age, and amenities—further shape pricing dynamics, covering gross area, furnishing, property type, and amenities like elevators. The independence of these dimensions creates a dynamic landscape.

The housing sales data represents samples from a dynamic series, akin to a grid where each cell, defined by latitude and longitude, harbors a baseline property price. This value evolves continuously over time and is influenced by factors like age, property type, and more. The temporal evolution occurs uniformly within each cell, while spatially, neighboring cells exhibit smooth price variations. This data, seen as observations from this dynamic pool, is leveraged statistically to infer population characteristics. This holistic understanding facilitates more accurate predictions, bridging temporal and spatial dimensions, and empowering stakeholders in navigating the real estate market's fluctuations.

Accurate predictions impact decision-making for buyers, sellers, and agents, streamlining investments. Real estate agents, armed with precise predictions, optimize sales strategies and cater to market demands effectively. At a broader level, spatiotemporal housing predictions enhance market efficiency, aiding economic growth. Governments use this data for policy-making, addressing housing shortages, and resource allocation. The fusion of temporal, spatial, and property-specific data transforms the real estate industry, making it transparent, efficient, and responsive to societal needs.

2. Related Works

The literature on housing price prediction has witnessed significant advancements, particularly in addressing the spatiotemporal dimensions of the real estate market. Researchers have increasingly recognized the importance of incorporating geographic and temporal aspects to enhance the accuracy of predictive models. Notable contributions include the work of Ali Soltani et al. (year), who conducted a comprehensive study in Australia utilizing machine learning models. Their research focused on analyzing the impact of various features, such as property attributes and neighborhood quality, on housing price variations. The study emphasized the significance of non-linear tree-based models and ensemble techniques, showcasing the utility of spatiotemporal lag variables to enhance prediction accuracy. This work represents a pioneering exploration into understanding the dynamics of the Australian property market, highlighting the potential of cutting-edge technologies for business and property valuation.

Another noteworthy contribution comes from Lei Xu et al. (year), who conducted a comprehensive review on spatiotemporal forecasting (STF) methods. The study explored the extension of traditional time series forecasting to space and time dimensions. The authors provided insights into statistical, physical, and artificial intelligence (AI) methods, elucidating the strengths and

limitations of each approach. They emphasized the importance of integrating data-driven and physical model-driven methods to improve interpretability and forecasting accuracy. The review not only offered insights into the methods and uncertainties in STF but also underscored the need for user-friendly, intelligent STF systems for real-time forecasting services. Recent research in housing price prediction emphasizes spatiotemporal dimensions. Wang et al. (2022) introduce the Flexible Spatiotemporal Model (FSTM), tailored for middle-small cities, considering factors like governmental policies, road density, and shared area. Fotheringham et al. (2015) explore spatiotemporal variations in London's house prices, using geographically weighted regression (GWR) and proposing GWR–TS, highlighting the importance of spatiotemporal dynamics in housing price modeling. These studies underscore the evolving landscape of housing price prediction, emphasizing nuanced spatiotemporal models for diverse urban contexts. These studies collectively underscore the growing emphasis on spatiotemporal considerations in housing price prediction, showcasing the interdisciplinary nature of research in this domain.

3. Methodology

Due to time constraints, we'll use a hybrid approach combining VARIMA time series analysis, Random Forest machine learning for prediction, signal processing methods like Fast Fourier Transform, and traditional statistical methods for testing and analysis. This integrated strategy aims to capture temporal patterns, handle complex relationships using machine learning, and provide rigorous testing and analysis through traditional statistical methods.

Random Forest is a versatile ensemble learning method, adept at handling both classification and regression tasks. Utilizing bagging (Bootstrap Aggregating) techniques, the algorithm constructs multiple decision trees during training, using different subsets of the training data created through bootstrapping. Notably, it introduces randomness by considering only a subset of features at each split, fostering diversity among the trees and mitigating overfitting. The Random Forest method is particularly suitable for this task as it adeptly manages both numeric and categorical variables. In our dataset, we encounter various features such as day, month, year, latitude, longitude, area, age, and more, which exhibit a combination of numeric and categorical characteristics, especially in the property category. Notably, scikit-learn's Random Forest lacks native support for handling both types simultaneously. While options like label encoding may introduce assumptions about ordinality and one-hot encoding might lead to a high-dimensional feature space. H2O's Random Forest efficiently accommodates both numeric and categorical variables without the need for extensive preprocessing, making it an optimal choice for our diverse set of predictors.

The VARIMA (Vector Autoregressive Integrated Moving Average) model provides a versatile framework for analyzing and predicting multivariate time series data. Comprising autoregressive (AR), integrated (I), and moving average (MA) components, it excels in capturing temporal dynamics within and among time series variables. Well-suited for tasks involving complex systems with interdependencies, VARIMA handles both numeric and categorical variables effectively. Its adaptability extends to economic forecasting, where multiple indicators influence each other, and diverse domains like finance, epidemiology, and environmental science. VARIMA's ability to address non-stationary time series data with trends and seasonality sets it apart from traditional ARIMA models, making it valuable for accurate predictions and decision-making. Here we will use VARIMA model to predict the sale number and average sale price in a day of different categories of properties, because there may be interdependence among these variables.

Fast Fourier Transform (FFT) and Discrete Fourier Transform (DFT) are powerful tools in time series analysis, especially in financial markets. FFT efficiently computes the DFT of a sequence, aiding in the identification of cyclic patterns and dominant frequencies in market price data. This transformation reveals hidden cycles and periodicities, such as daily, weekly, or yearly patterns, providing insights into market behavior. DFT, applied to equally spaced samples of a function, decomposes time series into constituent frequencies, helping analysts detect seasonal trends, analyze frequency components, and filter out noise. In financial markets, these techniques are crucial for understanding periodic market phenomena, uncovering hidden patterns, and making informed investment decisions based on the temporal characteristics of asset prices.

I utilized Geopandas in Python to perform calculations and exploratory data analysis (EDA) on the dataset to do spatial analysis,

employing the sjoin function to associate each property sale point with a corresponding polygon representing a community district.

4. Dataset

The dataset utilized in this project is sourced from the NYC government website, where I downloaded and consolidated multiple files spanning property sales transactions from 2003 to 2023, covering approximately 20 years. The data can be retrieved at https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page. It includes various property details such as category, sale date, sale price, as well as features like built year, gross area, and geographical coordinates. Notably, some features contain missing data. Two main preprocessing steps were applied: for the time series analysis of daily sale numbers, all available data was used, and missing values in the pivoted dataframe were filled with 0. In the analysis of average sale price per gross area, records lacking gross square feet values were excluded, and missing prices were forward-filled based on the closest previous day's value. For the Random Forest method, rows with essential features and minimal missing data were retained, limiting the data to the period from 2016 to 2023. It's important to note that the data may lack accuracy as it was manually combined from different files. The presence of unexplained jumps at two points in the sale number time series could be attributed to the manual merging process, introducing potential discrepancies.

5. Results

Initially, I explored the distribution of prices to understand its characteristics. Given that prices are inherently positive values, they typically exhibit a skewed distribution, akin to income distributions, often resembling a pyramid shape with near-0 density near zero. Analysis of the plots revealed these expected patterns. Notably, the straight lines observed on the right side of the plots suggested a power-law tail in the distribution, indicating a propensity for extreme values. This observation held true for prices, prices per gross area, and prices per land area. (Fig 1) The QQ plot against the log-normal distribution illustrates a nearly straight line, suggesting a close resemblance to a log-normal distribution. (Fig 2)

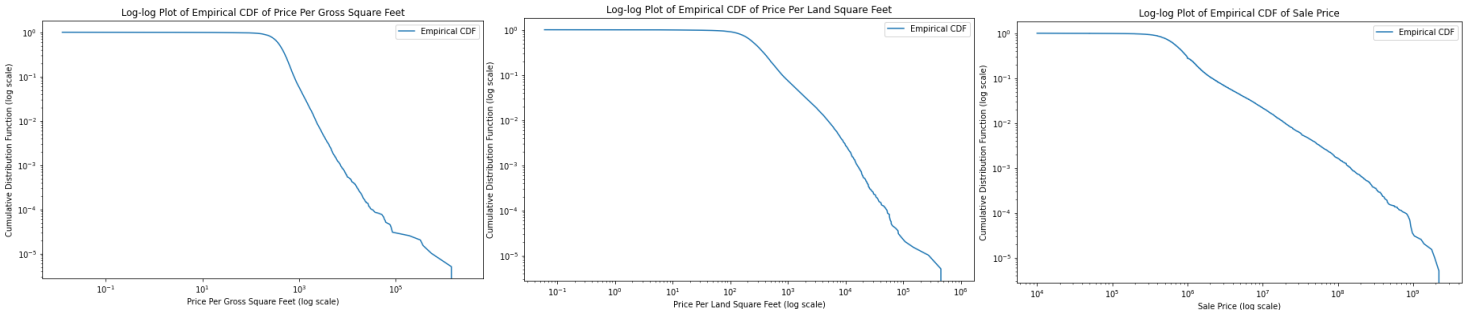


Figure 1. Loglog plot of the surviving function of Sale Price, Sale Price Per Land Square Feet, and Sale Price Per Gross Square Feet

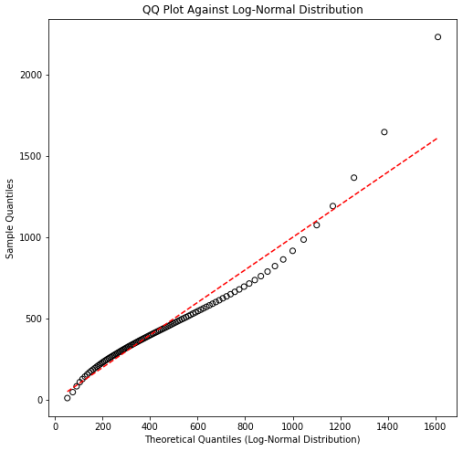


Figure 2. QQ plot of Sale Price Per Gross Square Feet against Log Normal Distribution

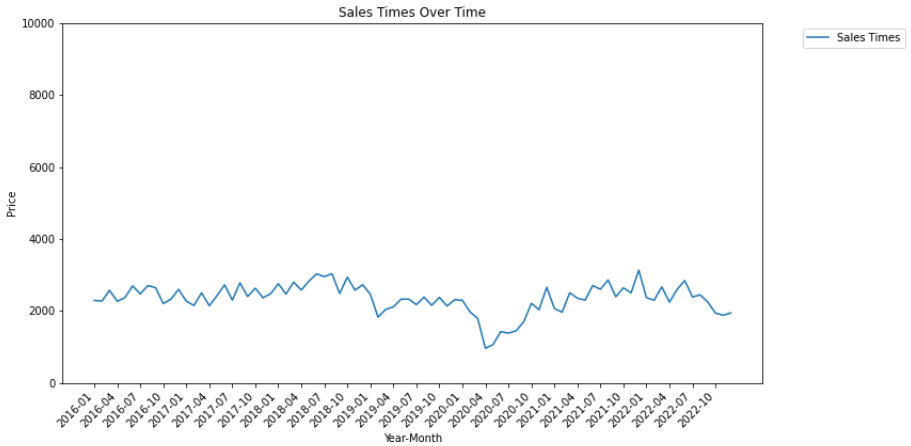


Figure 3. QQ plot of Sale Price Per Gross Square Feet against Log Normal Distribution

In this sale times per month plot (Fig 3), we indeed sees a drop in early 2020, which is largely due to Covid. The sales time kept

very steady from 2016 to 2018 with slight increase.

In Fig 4, we can see that though there is a slow growth in the price per gross square feet for residential buildings. Condominiums and elevator apartments have larger fluctuations. Though there is a large difference in price per land square feet between low-density residential buildings, such as one or two family dwellings and high-density residential buildings, such as elevator apartments and condos, their difference in price per gross square feet is relatively small. Elevator apartments even have lower price per gross square feet. In these plots we can also see that medians tend to give steadier trend than means, because it can avoid the affects of outliers.

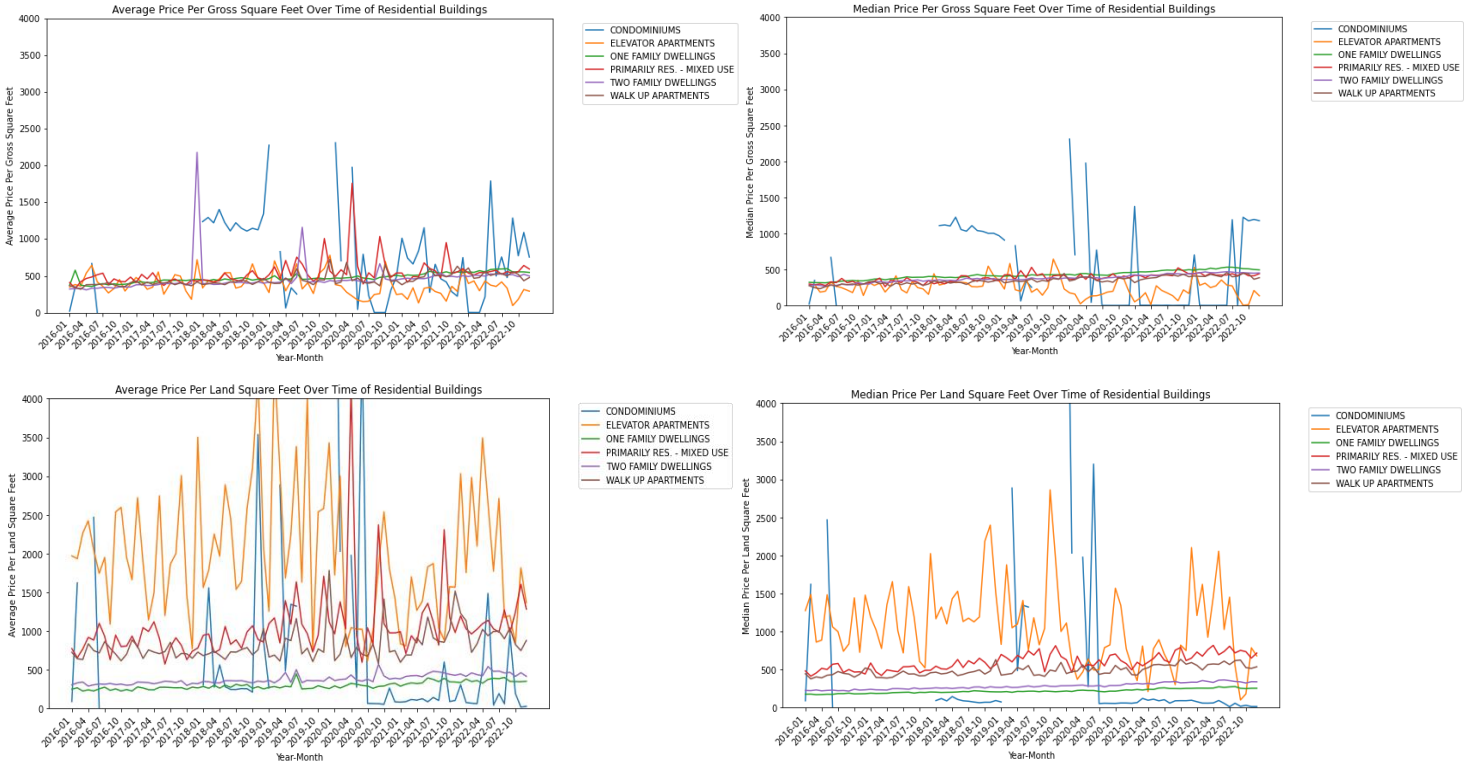


Figure 4. Median and Mean Price Per Land Square Feet and Price Per Gross Square Feet in Every Month

From the three maps in Fig 5, we can see that Manhattan has a very high gross area over land area ratio, which could be an indicator for the average height of buildings. Also, buildings in Manhattan and some parts of Brooklyn have significantly older age, which is more than 100 years old. We could also see that there is an obvious citywide drop in residential property sales in 2020, probably due to covid. The price per gross area percent change rates in 2020 do not have a clear trend across different community districts. Some districts witnessed increased values while others saw decrease.

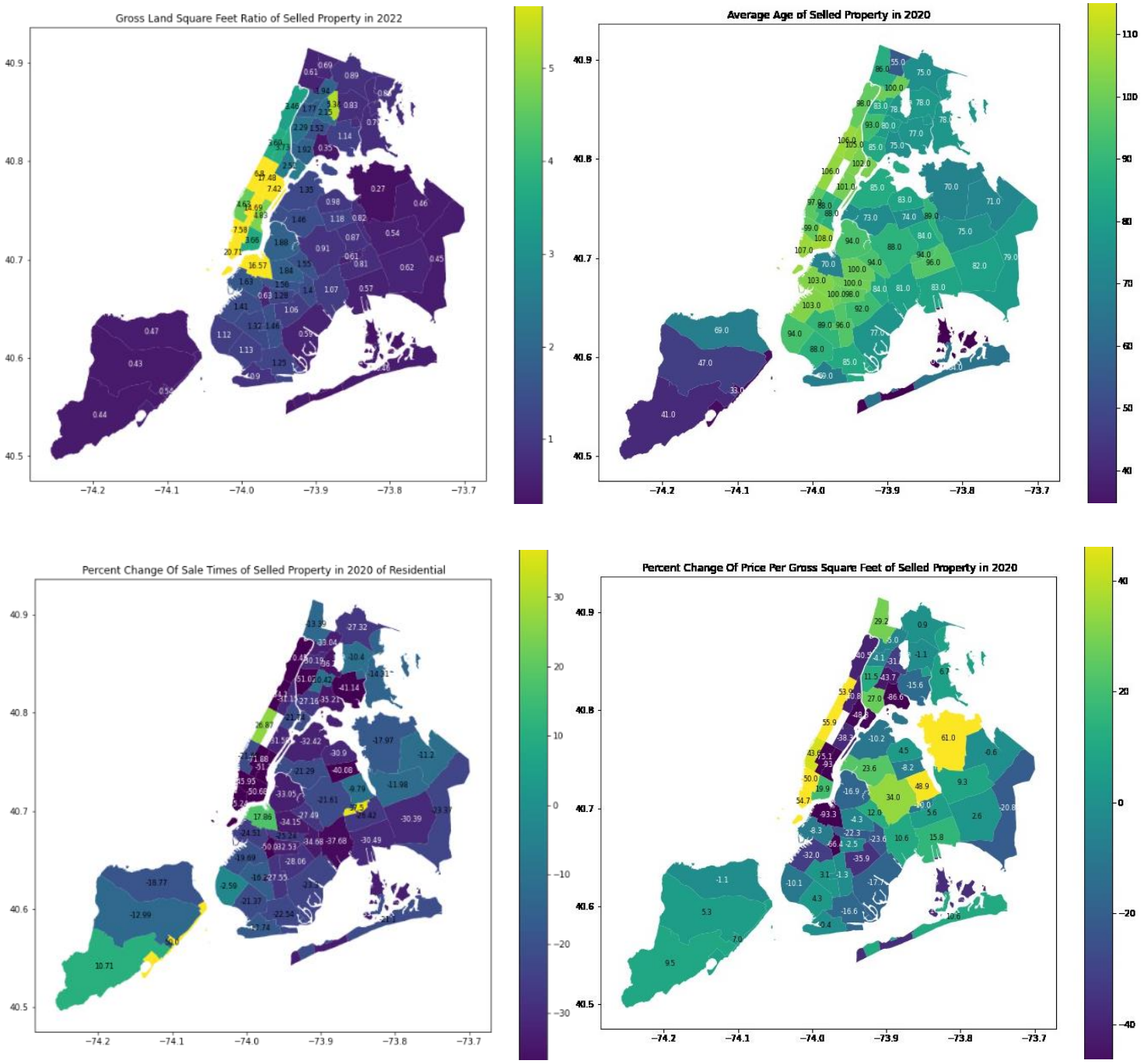


Figure 5. Map of Gross Area over Land Area Ratio in 2022, Map of Average Sold Property Age in 2020, Map of Percent Change of Sale Times of Residential Buildings in 2020, Map of Percent Change of Price Per Gross Square Feet of Residential Buildings in 2020

From the daily property sale number, we do see a weekly periodic pattern in the zoom-in plot, where the sale number is high during weekdays and low during weekends. Condos tend to exhibit some extrenely large values. We also see the drop in property sale numbers during the sart of the pandemic, and an inexplicable jump during 2022, which I reckon to be data collection discrepencies.

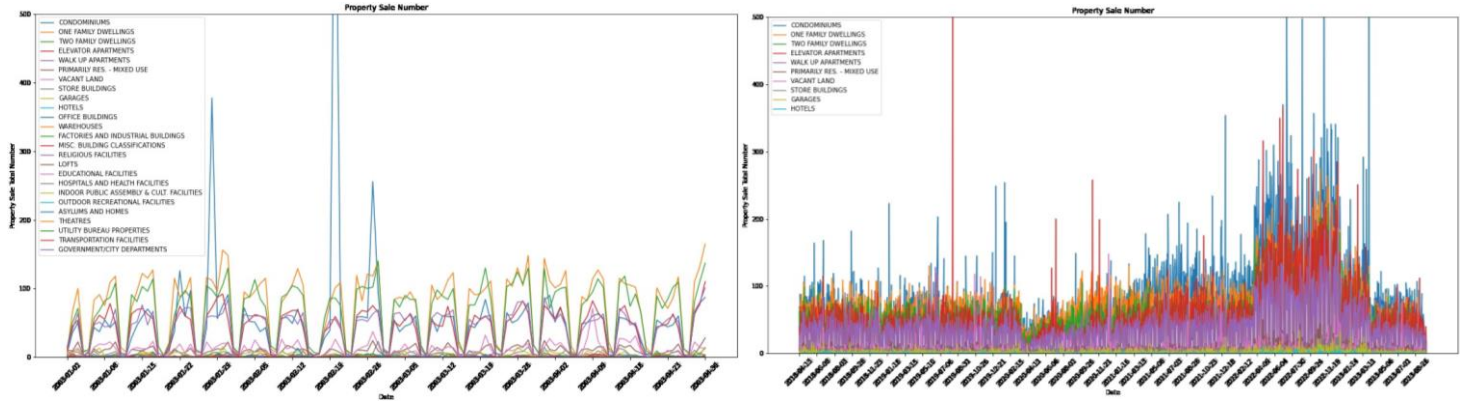


Figure 6. Property Sale Number of Every Day of Different Property Categories

When doing FFT to the daily sale number, we see clear peaks at $0, \pm 1/7, \pm 2/7, \pm 3/7$. (Fig.6) The peak at 0 indicates the series does not have a mean of 0. 0 frequency equals to a infinitely large period, and muliplied by an amplitude, is equivalent to a constant added to the signal. The frequency of 1/7 corresponds to our observation of the series being periodical weekly. When the series is detrended by subtracting the moving averge, we see the peak at 0 disappears and the magnitudes of frequencies near 0 become very small. Because of the periodic nature of the signal, we could filter out all frequencies besides $0, \pm 1/7, \pm 2/7, \pm 3/7$, and largely still get the same signal as the original one. (Fig.7)

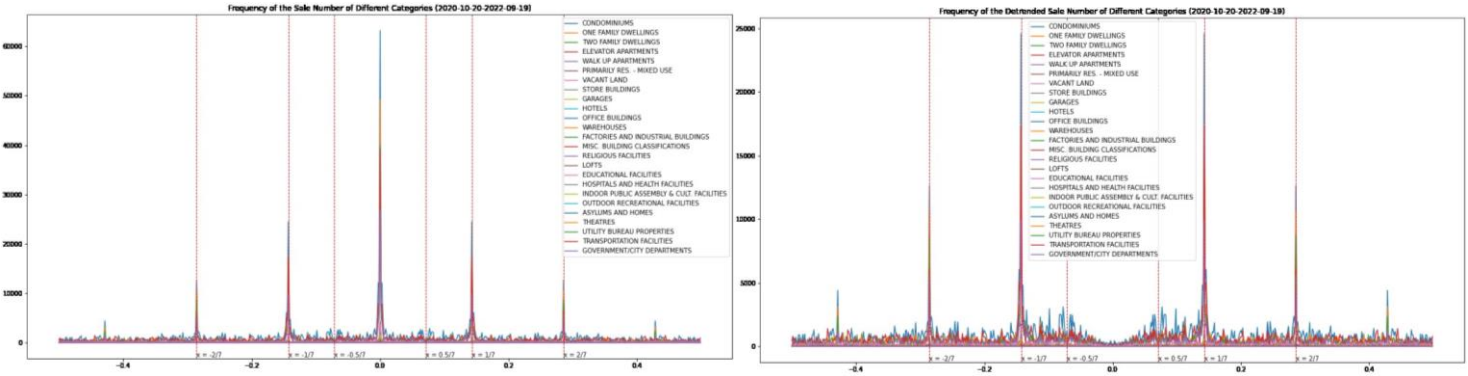


Figure 6. Magnitudes of Frequencies of the Original Daily Sale Numbers and Detrended Daily Sale Numbers

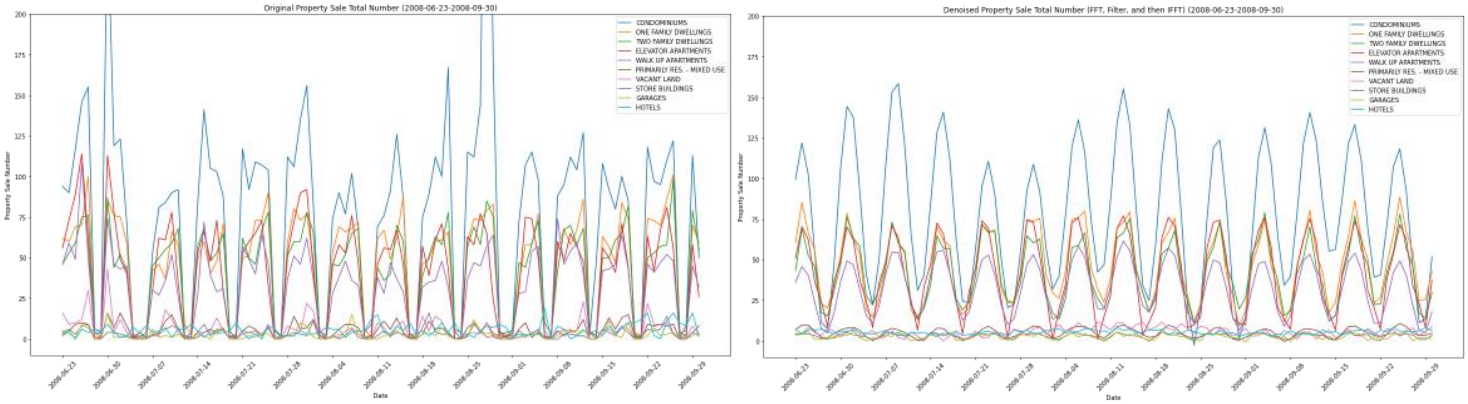


Figure 7. Original and Denoised Daily Sale Numbers Using FFT, Filter and IFFT

Using the Trend, Seasonal, Residual Decomposition method from statsmodels.tsa.seasonal, we can see that condos and one family dwellings have very simlar patterns, they all have a period of 7 days. And their trends are almost identical.

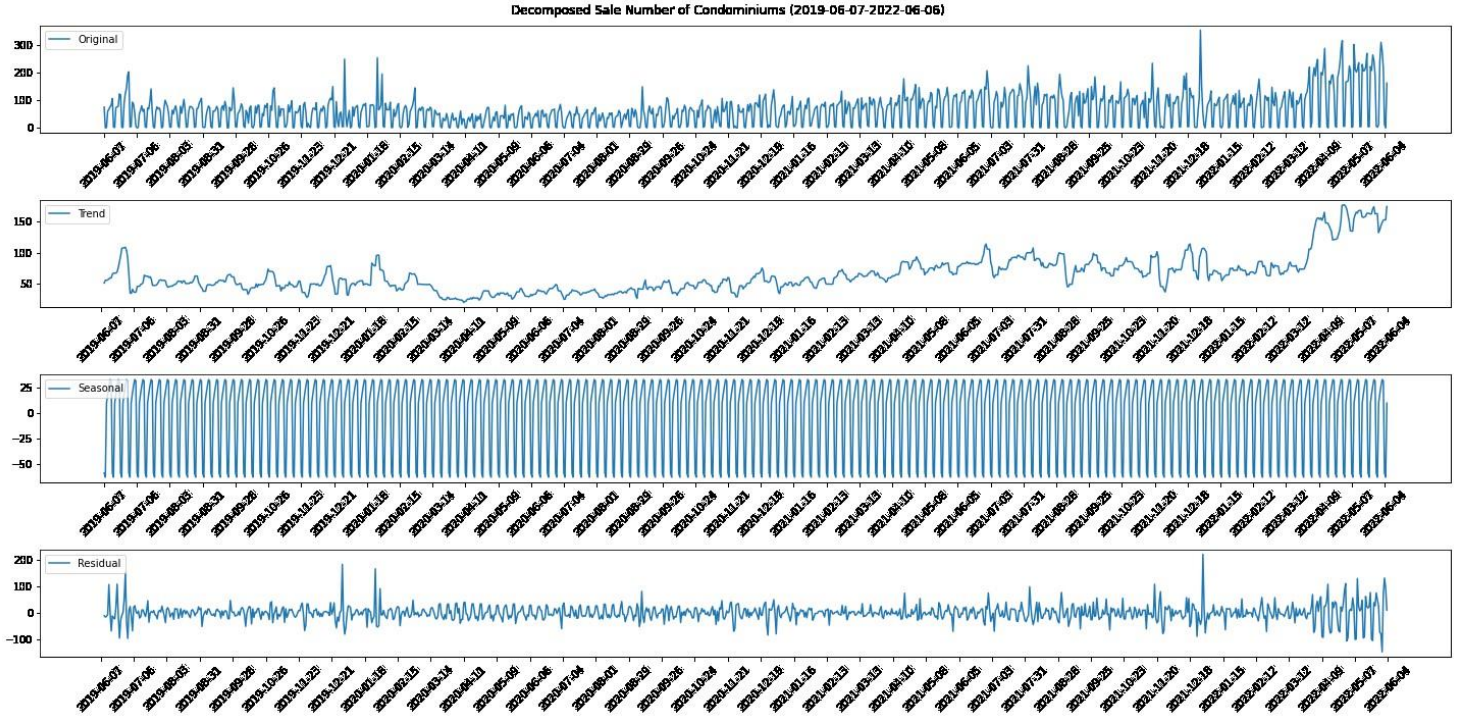


Figure 8. Trend, Seasonal, Residual Decomposition of the Sale Number of Condominiums

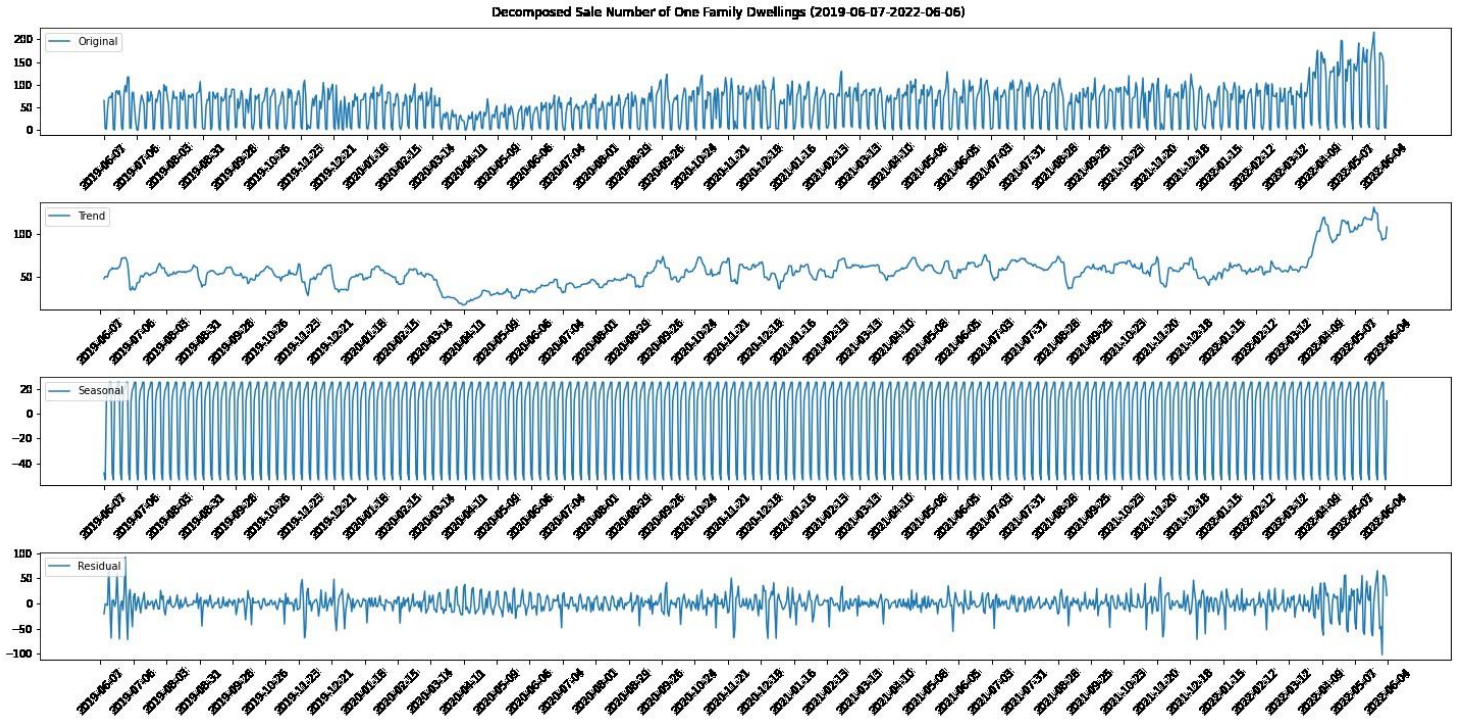


Figure 9. Trend, Seasonal, Residual Decomposition of the Sale Number of One Family Dwellings

By applying the VARIMA (p=21, d=0, q=0, s=7) model to a daily sale number vector consisting of the sale numbers of some categories of buildings (I used 5 categories here), we can see the VARIMA model gives a pretty good prediction of the future values. I believe part of the reason of such satifactory performance of VARIMA is that the sales number data is very steady across a short period of time, and the periodic feature is too obvious. There is not so much stochaticity in the series, resulting in small residuals. Note the difference from the sale price series in the later part. Fig 11 shows the coeffcients of the fitted VARIMA model. Looks like there is some periodicity in the coefficients. The positive coefficients and negative coefficients show some alternating positions.

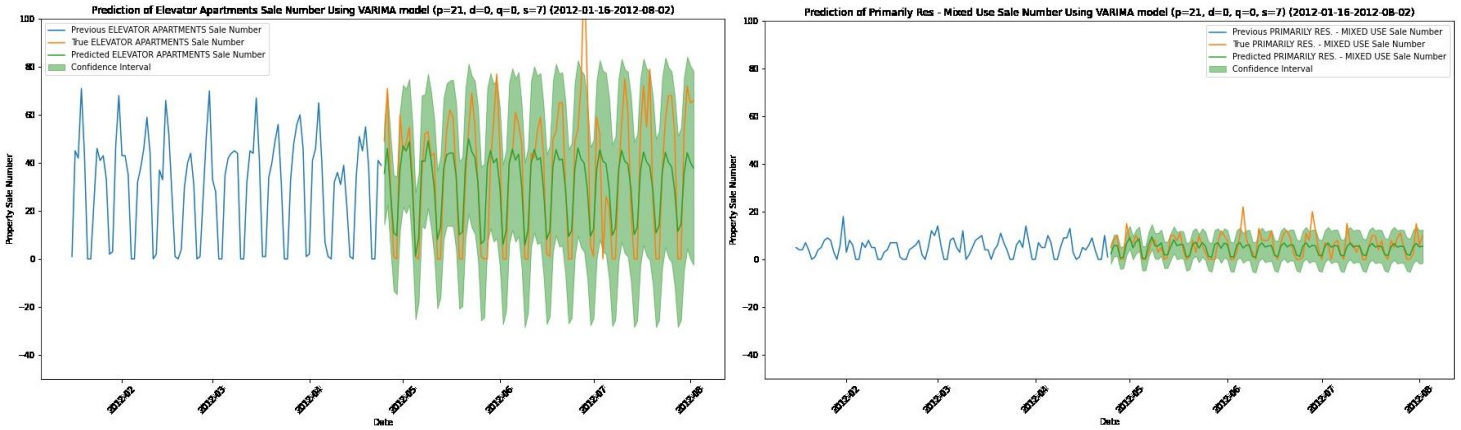


Figure 10. Use VARIMA model on Daily Sale Number Series and Predict Future Values

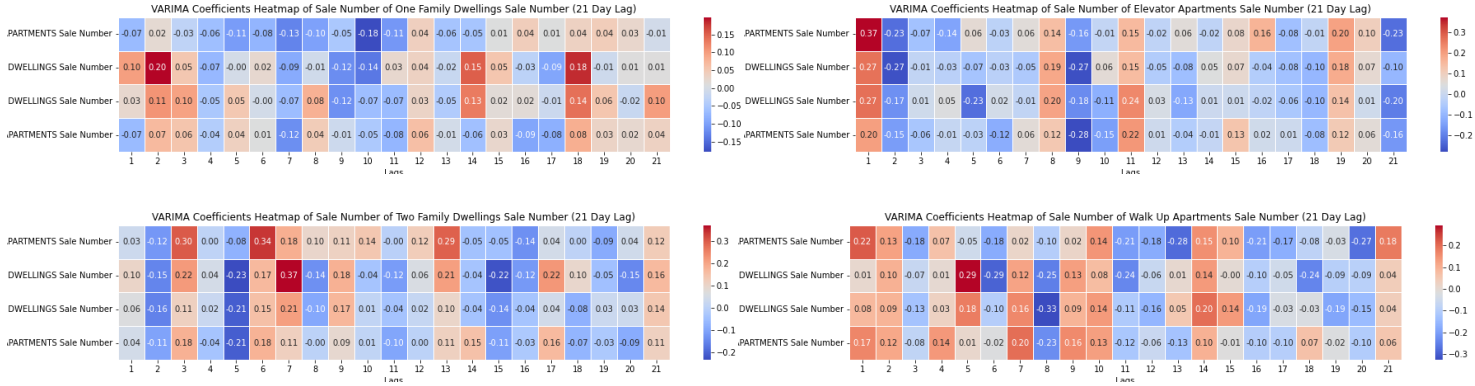


Figure 11. Some of the Coefficients of the Fitted VARIMA model

Because the VARIMA model here assumes interdependence between the sales numbers of different categories, I want to probe into the correlation between the sales number of different categories. In thes etwo heatmaps below, rows and columns are sorted in the descending order of most total sale number. We can see that for standardized sale number, those with the highest total sale numbers (all of them are residential buildings), are large correlation values, especially for one and two family dwellings. This indicates that residential buildings tends to show a collaborative trend in variation. Warehouse and industrial buildings have pretty high correlation. Hotels have very little correlation with all other categories. Office buildings is moderately correlated with residential categories. However, for standardized change rates, the heatmap looks quite messy. There isn't much we can tell, and some values are quite inexaplicable. The correlation bwteen warehouse and industrial buildings turns negative here, and the correlation between asylums and recreational facilities bcomes pretty large.

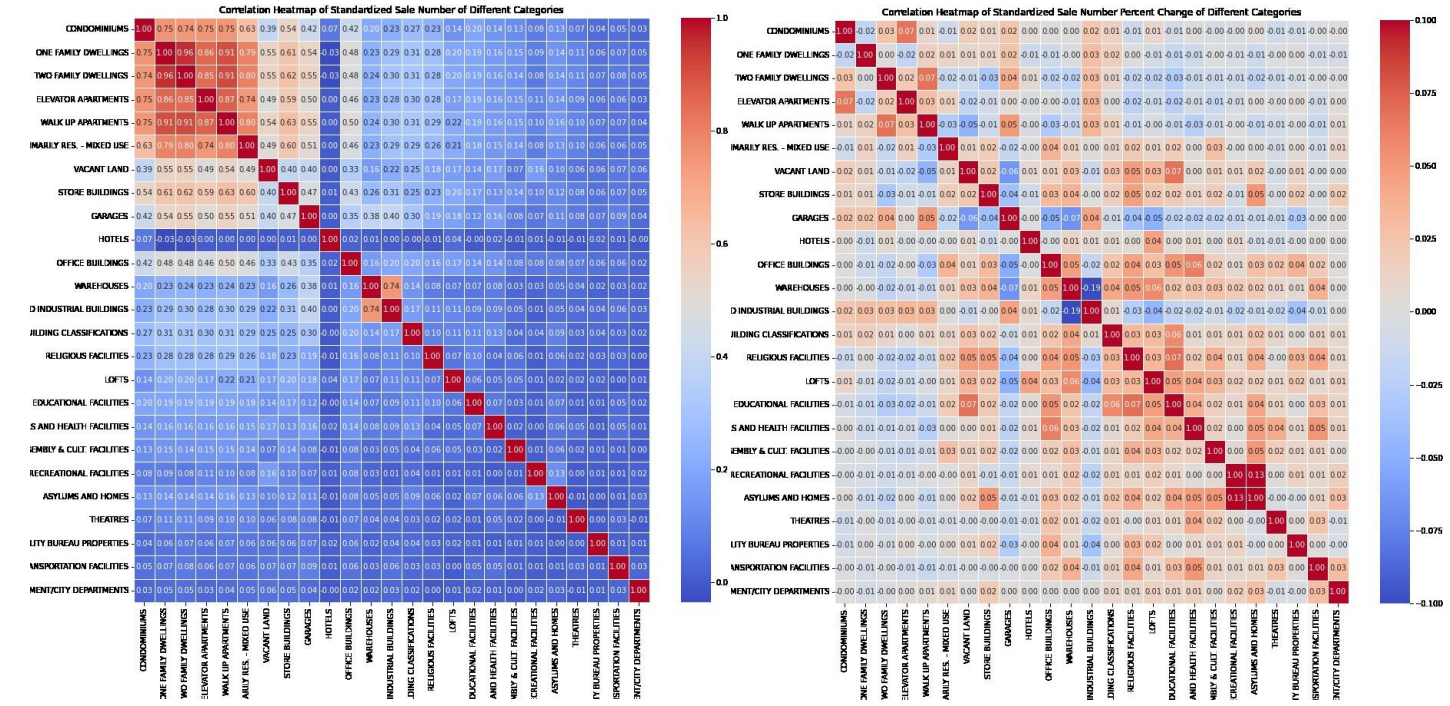


Figure 12. Correlation Heatmaps of Standardized Sale Numbers and Standardized Sale Numbers Percent Change of Different Categories

In the ACF plots below, we can see that for sale numbers, the autocorrelation is quiet high, and oscillates regularly.

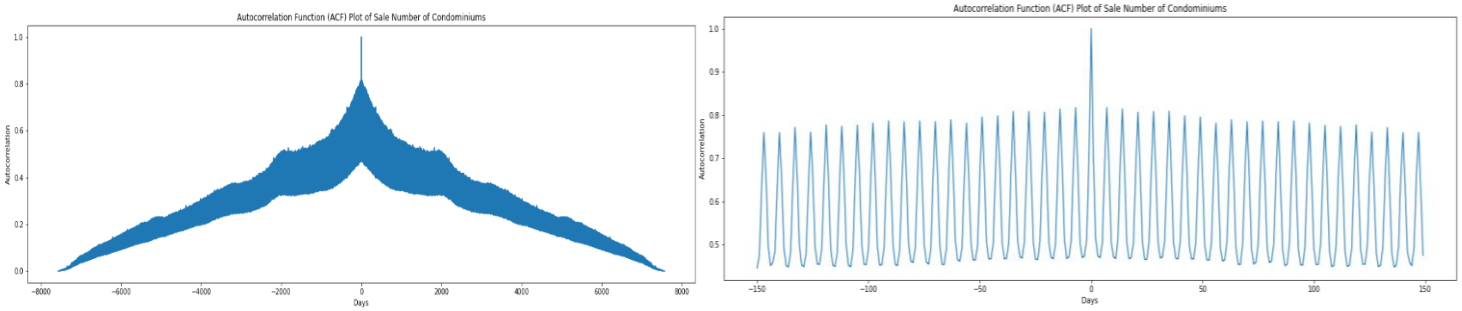


Figure 12. ACF Plot for Daily Sale Number of Condominiums

For price per gross square feet, the data is missing a huge chunk for many categories of building during a long period of time, therefore I only select two of the most complete categories. We can see that the periodic feature is not obvious.

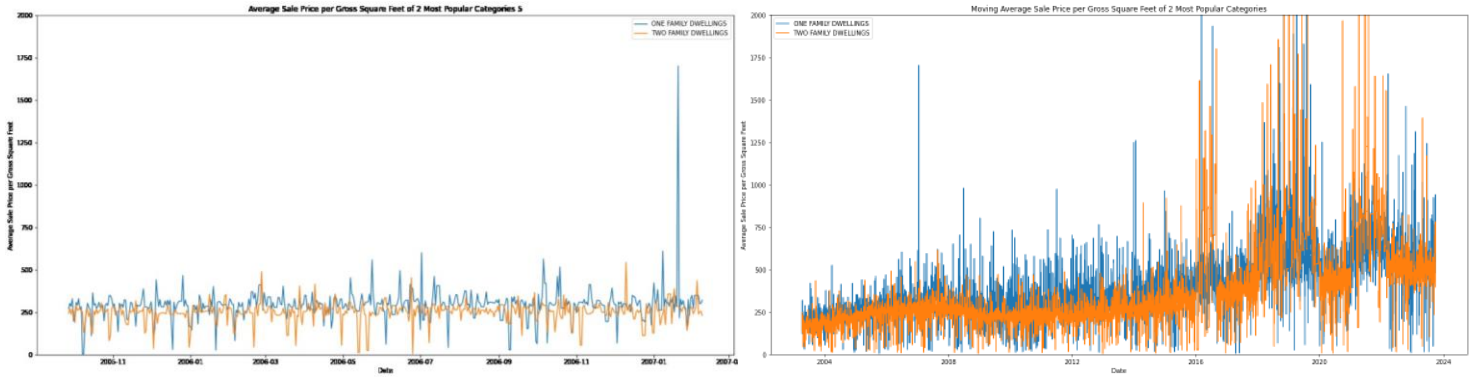


Figure 13. Property Price Per Gross Area of Every Day of Different Property Categories

We can see that though there is no visible periodic feature, Fourier Transform still gives us the same frequencies as before. The detrended series also have no peak at 0.

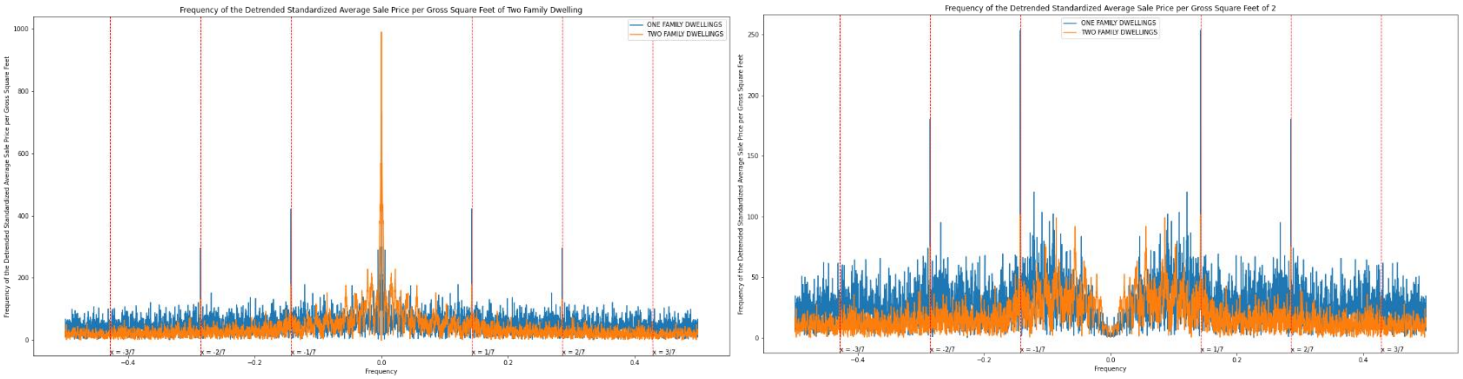


Figure 14. Magnitudes of Frequencies of the Original and Detrended Daily Price Per Gross Area

By using FFT, Filter and IFFT, we can see there seems to be a clear appoximately 2-month period in the price per gross area od two family dwellings, but no periodicity in one family dwellings. Here the fitting using FFT, Filter and IFFT, is less satifactory than the previous sale number scenario. We see a lot more noise and stochaticity here.

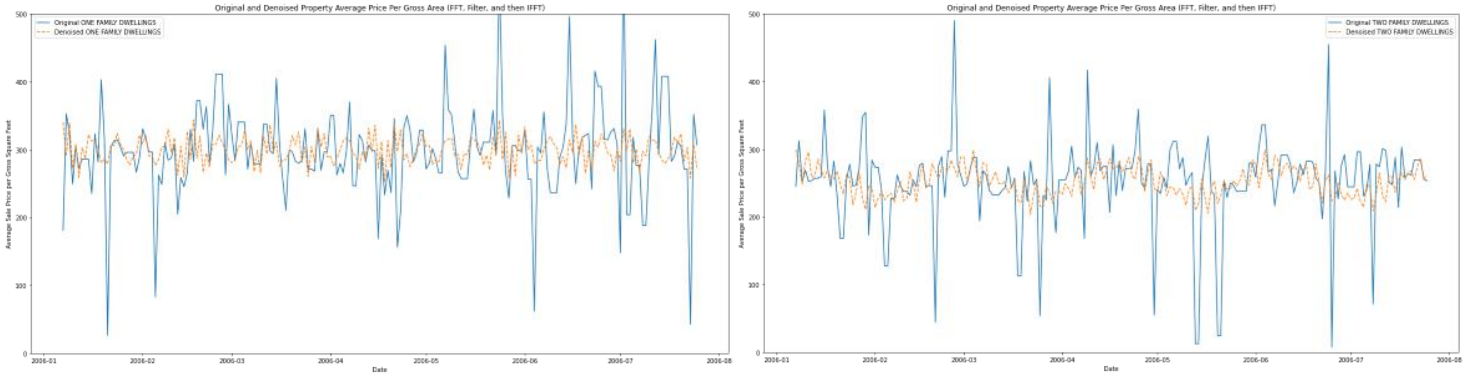


Figure 15.Orginal and Denoised Price Per Gross Square Feet Using FFT, Filter and IFFT

Use the same seasonal decomposition, we see here the two signal are very much different, both in trend and seasonality.

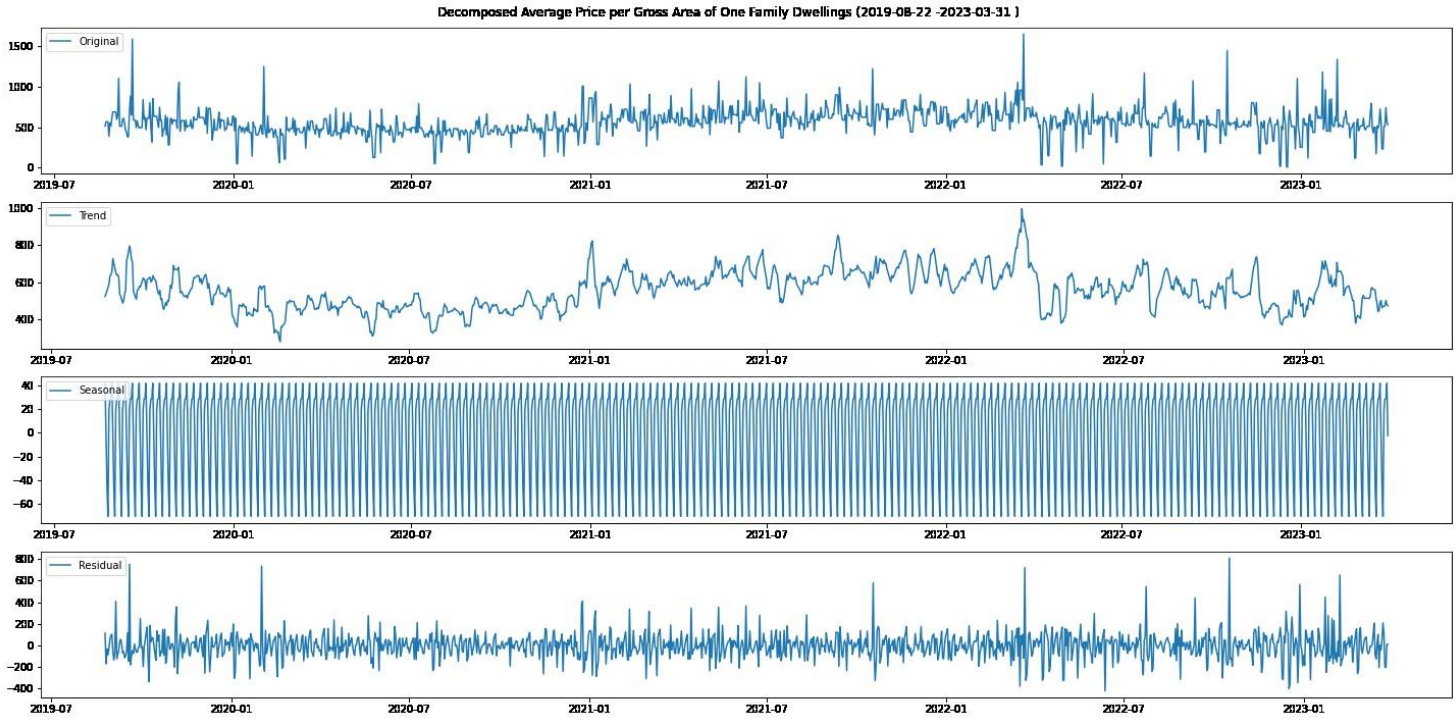


Figure 16. Trend, Seasonal, Residual Decomposition of the Price Per Gross Square Feet of One Family Dwellings

Figure 17. Trend, Seasonal, Residual Decomposition of the Price Per Gross Square Feet of One Family Dwellings

Here suing the same VARIMA model (p=21, d=0, q=0, s=7) as before, we can havea prediction of the future values. Because the periodicity is weaker, our prediction turns to a flat line quickly, and the prediction does not look like the true values.

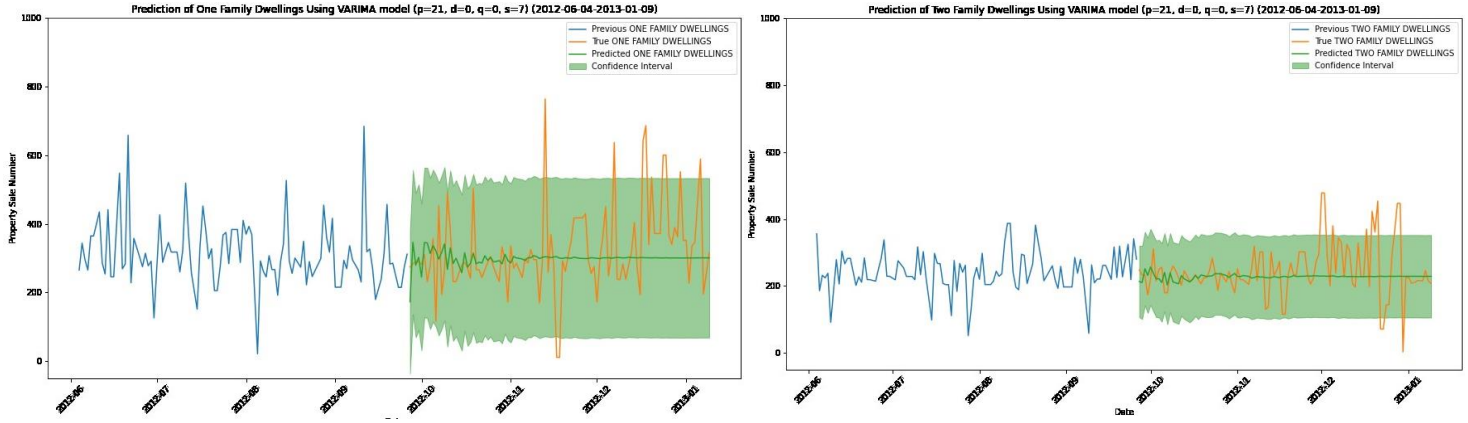


Figure 18. Use VARIMA model on Price Per Gross Square Feet Series and Predict Future Values

Here we notice soemthing interesting, though the correlation for price per gross square feet of different categories themselves are pretty weak, when we increase the window size for moving average, the correlation becomes stronger and stronger. This paritally shows that though the short term price has a lot of noise, and the trend is not obvious, in the long term, when noise is neutralized through moving average, the trend becomes more obvious. And in the long run, different categories indeed show the same collaborative variation in price.

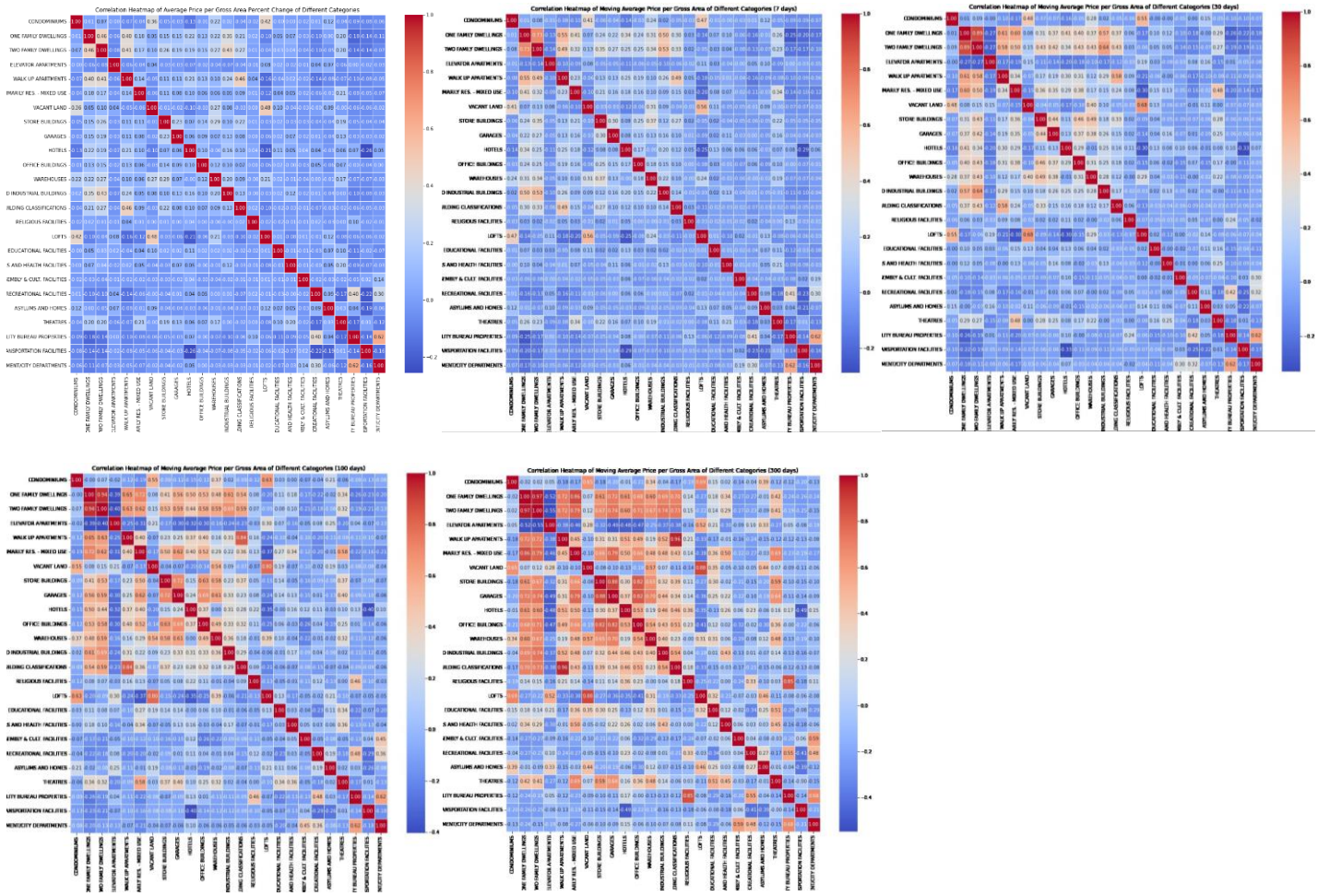
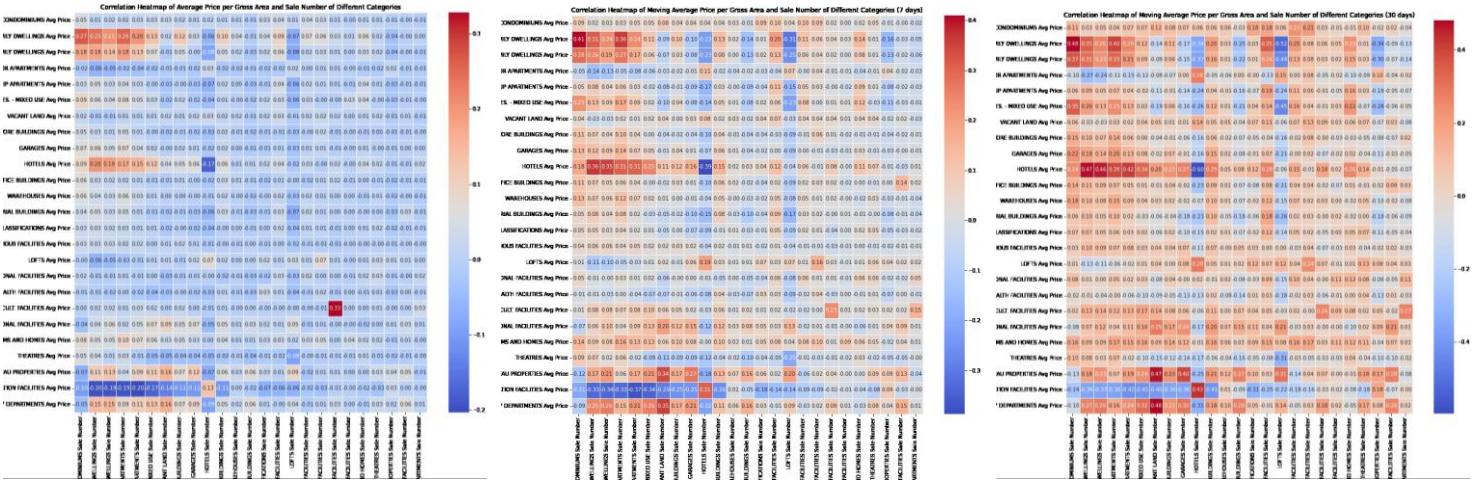


Figure 19. Correlation Heatmaps of Price Per Gross Square Feet of Different Categories of 0, 7, 30, 100, 300 Moving Average Window Size

In the following plots, we see similar patterns. Correlation between price per gross square feet and daily sale numbers strengthens as window size of moving average increases. This shows the long term positive correlation of sale numbers and sale price.



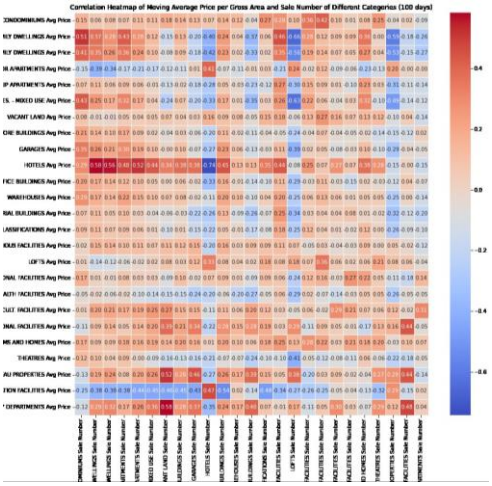


Figure 20. Correlation Heatmaps of Price Per Gross Square Feet and Daily Sale Numbers of Different Categories of 0, 7, 30, 100 Moving Average Window Size

In order to see which factors have the most impact on the price of a property, I used random forest in h2o to predict values. Feature importance in random forest is defined by the times a feature is used in splitting. Though this random forest model gives pretty big mean absolute percentage error, the feature importance could still be insightful. We can see that when predicting sale price, the area of the property is very important, and then is the category of the property. This is rather intuitive. But when predicting price per square feet, there are more features invloved in this: longitude, latitude is also important.

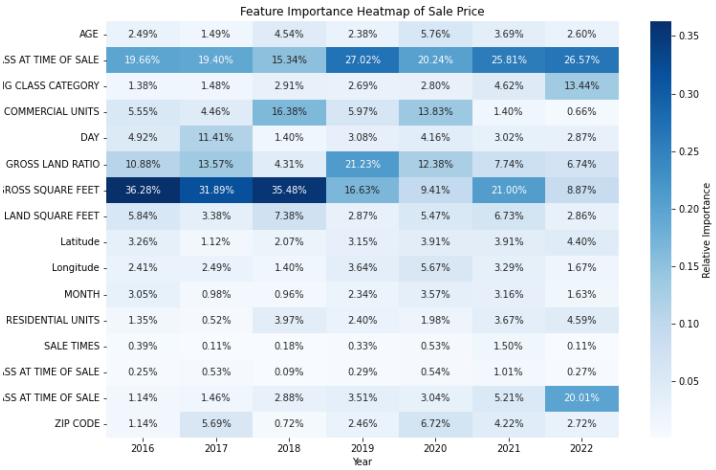


Figure 21. Feature Importance When Predicting Sale Price

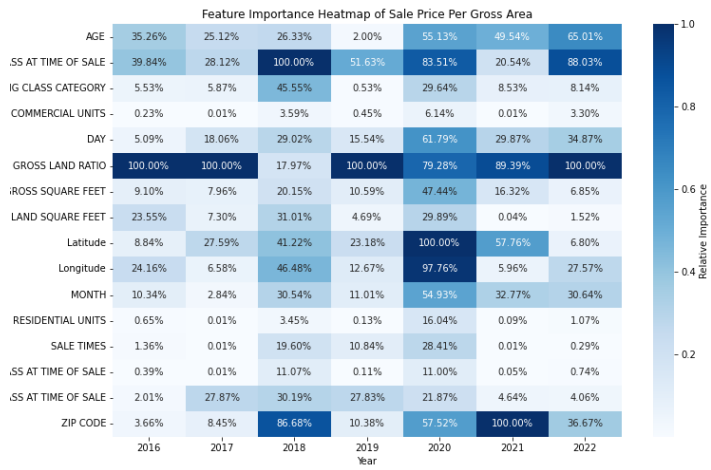


Figure 22. Feature Importance When Predicting Sale Price Per Gross Area

6. Conclusion and Future Work

In conclusion, this project employs a comprehensive approach to predict housing prices by integrating spatiotemporal methods and advanced modeling techniques. The analysis, based on a dataset spanning two decades of property sales transactions in NYC, underscores the intricate dynamics of real estate, considering factors such as time, location, and property attributes. Notable findings include the identification of periodic trends using Fast Fourier Transform, effective predictions using VARIMA models, and feature importance insights from Random Forest. The results emphasize the importance of considering both temporal and spatial dimensions in housing predictions, providing valuable information for stakeholders in the real estate market. Despite challenges such as data discrepancies, the study contributes to enhancing transparency, efficiency, and informed decision-making in the real estate industry. Future research could explore further refinements in modeling techniques and address data quality issues to advance the accuracy and applicability of housing price predictions.

In the future, I want to research more into the different methods of spatio-temporal prediction, such as statistical methods, AI

methods. I want to apply stochastic process into this problem, making it a 3 dimensional stochastic process. I would considering using smoothing, interpolation method to infer the population distribution using the observation we have, and then get the prediction of a price. Ultimately, I want to develop a algorithm and system, in which an estimate for price is given whenever the time, latitude, longitude, category, age, gross square feet is inputed. This also may require soke economics and real estate knowledge.

References

Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Exploring, modelling and predicting spatiotemporal variations in house prices. The Annals of Regional Science, 54(2), 417–436. https://doi.org/10.1007/s00168-015-0660-6

Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, 103941. https://doi.org/10.1016/j.cities.2022.103941

Wang, L., Wang, G., Yu, H., & Wang, F. (2022). Prediction and analysis of residential house price using a flexible spatiotemporal model. Journal of Applied Economics, 25(1), 503–522. https://doi.org/10.1080/15140326.2022.2045466

Xu, L., Chen, N., Chen, Z., Zhang, C., & Yu, H. (2021). Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. Earth-Science Reviews, 222, 103828. https://doi.org/10.1016/j.earscirev.2021.103828