Scalable Spatial Analytics

Pages 1–7
2022

Quantitative Analysis of the Impact of Selected Points of Interests on Housing Prices in Taipei, Taiwan

Ibtihal Alshehri, Morris Chang, Yuxin Miao

CIVENG 263

UC Berkeley - Department of Civil Engineering

December. 05. 2022

Keywords: Housing Prices, POI, Clustering Methods, Linear Model, Folium, K-Means

ABSTRACT

The housing market in Taiwan has faced significant growth over the past decade without notable premonition. There are various factors that could have contributed to this growth, including vast economic growth, population rise, and new housing policies. However, there are some factors that could impact the homogeneity of housing prices within the city, including the amenity value of the surroundings. For instance, residents' accessibility to public transportation and business centers can influence the price of housing. Researchers in the past have used various models to predict housing prices by quantifying a city's urban amenities using structural and environmental scores, but these studies were often limited due to difficulty in data collection. Therefore, Taipei City in Taiwan is and ideal location to be used as a case study, since the government publishes a wide range of data publicly, including housing sales, income, and education, etc. In addition to the publicly available data, the building of the predictive model would require collecting and analyzing urban amenity data in terms of the Points of Interest (POIs) information for each of the residential houses that is available through the Google Maps API, and projecting the result in comparison to the city's socioeconomic layout. A non-hierarchical clustering method was used to cluster the residential houses in the city by the number of different POIs in various radii surrounding the house, including 500 meters as the area within a short walkable distance, 1 kilometer as the area that is quickly reachable with a bike, scooter, or a short bus ride, and 3 kilometers which represents an area at a short car ride distance. In order to understand the effect of the nearby POIs, a linear regression model was applied to quantify the impact of various POIs on housing prices. The results have indicated that transportation POIs such as subway stations and bus stations holds a high level of significance in the effect of housing prices, and factors such as libraries, churches, and shopping mall also have a positive effect on housing prices. By projecting and comparing the housing prices with socioeconomic factors, income and activity flow have been identified as the factors that are correlated with housing prices, and aside from post-graduate degrees, education does not have a significant correlation with housing prices.

CONTENTS

Contents 1

1 Introduction 1
1.1 Related Work . . . 2

2 Data Acquisition 2
2.1 Real Estate Sales Price . . . 2
2.2 Point of Interest (POI) Data . . . 2
2.3 Education Level at District/Sub-district Level . . . 2
2.4 Average Income at District/Sub-district Level . . . 2
2.5 Population at District Level . . . 2
2.6 Telecommunication Signaling Activity Data at District Level . . . 3

3 Data Pre-processing 3

4 Visualization of Raw Data 3
4.1 City's Housing Layout . . . 3
4.2 Density per district . . . 3

5 Modeling 3
5.1 Clustering . . . 3
5.2 Linear Regression Model . . . 4
5.3 Socioeconomic Correlation . . . 5

6 Conclusion 6

7 Challenges, Caveats, and Opportunities 7

References 7

8 Appendix 7
8.1 GitHub . . . 7
8.2 Data set sources . . . 7
8.3 Additional Figures . . . 7

1 INTRODUCTION

While growth in housing prices may participate in the economic uprise and creating more business opportunities, it has substantially worsened inequality and overall well-being within nations around the world. Many households with medium to low income spend most of their earnings on housing, and the rises in residential unit prices make keeping up even harder. The ability to develop a city that is balanced in prices and the ability to invest accurately in real estate properties have become essential skills for city governments and their citizens. In this paper, our team attempts to deliver an analysis that evaluates the impact of certain points of interest (POIs), or "residential attractors", on the housing market within a city. The goal is to understand how certain public or private features, placed at certain distances, may impact the price of a residential unit, and create a geospatial predictive model that might be applied in similar urban settings. The model will be helpful in understanding how uneven distribution of resources in a city may lead to potential effects on housing prices and bring further insights for decision-makers and urban planners to identify the services that are highly attractive to residents and understand spatial factors that influence residential development in

general. In our work, we present a framework for classifying districts within cities by their attractiveness to residents and addressing the spatial dependence between these variables, all using data from various publicly open sources including, the real estate transactions data for Taipei, Taiwan, annual income tax data, education census, activity flow data based on phone signals, and POIs information from the Google Maps API. The results from the POIs analysis would be further evaluated against socioeconomic factors to understand the correlation of varying social and economic status of the population in these districts, and its impact on housing prices. This will help us assess the demand or need for certain services in relation to the socioeconomic status of the residents in each of the selected areas.

1.1 Related Work

Through literature, house pricing has been a popular topic of investigation and research in the past. Research by Lin, Yeh, and Tou has shown that socioeconomic factors within a region would affect the housing prices in the area [1], while research by Lu, Wang, and Yu has presented a model on spatiotemporal factors to predict residential house prices across cities [2]. Although these papers have shown working methods to predict residential house prices, socio-economic factors may not be the sole reason and influence, and the analysis of spatiotemporal factors would demand a large amount of reliable data to be available. This paper presents an interesting angle that could be easily followed in different urban settings to show the importance and relationship of POIs to house prices in the area.

2 DATA ACQUISITION

Sufficient and reliable data is needed in order to conduct the analysis and create the linear regression model. The following sections would discuss the origin of each of the data sets used in the model or analysis.

2.1 Real Estate Sales Price

In 2020, Taiwan's highest legislature, the Legislative Yuan passed the Third Reading of the amendment on three acts in regard to registering the actual selling prices of real estate, including The Equalization of Land Rights Act, the Land Administration Act, and the Real Estate Brokerage Management Act. This amendment required all transactions of real estate properties to be reported to the Ministry of Interior, and the information in regards to the property along with the price would be updated on a publicly available website three times a month. This data set is also available for download and uses on Taiwan's open data platform without prior permission. The scope of this paper would focus on 4077 real estate transactions out of 40,000+ transactions from the third quarter of 2020 to the third quarter of 2022.

2.2 Point of Interest (POI) Data

In order to analyze the influence and effect of the number of different kinds of Point-of-Interest(POIs) around a specific location, the data would need to be acquired from a map source that is abundant in data and often updated. This limits the option to either Google Maps API or OpenStreetMap's API Overpass. Our team decided to proceed with the Google Maps API as it is more commonly used in the selected area of study and ample in POI data. However, due to the limit in budget and credits available for the Google Maps API query requests, the number of transactions and type of POIs had to be limited to 4077 transactions, and 12 different types of POIs at three radius distances (500 meters, 1 kilometer, and 3 kilometers)for each residential properties to measure to closeness and concentration of the POIs. The POI data was queried from Google Maps API in early November 2022.

The list of Point-of-Interest(POIs) queried and the reasoning behind:

1. **Police Station/Police Office:** Potential indicator of safety and security of the neighborhood.
2. **Hospital:** Indicator of health care facilities available, access, and barriers to necessary care.
3. **Shopping mall:** Indicator of leisure activities, shopping, and restaurants available in the area.
4. **Library:** Indicator of access to information services, public facilities, and education-related resources.
5. **Bus station:** Indicator of transportation and human mobility.
6. **Subway station:** Indicator of transportation and human mobility.
7. **University:** Indicator of distance and access to higher education institutions and campus resources.
8. **Primary school:** Indicator of access and distance to education institutions and resources, also measures the number of schools within the surrounding area.
9. **Church:** Although the main religion in Taiwan is not Christianity or Catholicism, churches could serve as an indicator of access to religious facilities and services.
10. **Night club:** This could be both a positive and negative indicator, as nightclubs could represent leisure activities, but could also come with potential safety and cleanliness concerns.
11. **Supermarket:** Supermarket is a necessary store needed to purchase daily necessities and could serve as an indicator that represents ease of getting daily necessities.
12. **Park:**: Indicator of leisure space and green space available in the surrounding area.

2.3 Education Level at District/Sub-district Level

The data set on the Education level across different districts and sub-districts in Taipei is acquired from the Ministry of Interior's data uploaded on Taiwan's Open Data platform. The data set is pipe-lined and sourced from the Ministry of Education's database on education statistics, and records. The original data set is recorded in Traditional Chinese and translation of the column titled is needed. The education data used in the analysis was based on the year 2021, as it is the most updated data available.

2.4 Average Income at District/Sub-district Level

This data set on the average income across different districts and sub-districts in Taipei is acquired from the Ministry of Finance's data uploaded on Taiwan's Open Data platform. The data set is based on the previous year's Individual Income Tax Report, and it is rounded to the nearest thousand New Taiwan Dollars (NTD). The data set is recorded in Traditional Chinese and translation of the column titled is needed. The data set used for the analysis is based on 2019 since the final version of the data set for 2020 and 2021 has not been updated yet due to adjustments and delays in processing.

2.5 Population at District Level

This data is publicly available on the Taipei City Government's open data platform for download. The original data is updated on a monthly basis to show the change in the number of people living in each of the districts. In this case, the data used for analysis was based on the most up-to-date data availed, which is numbers based on October 2022.

2.6 Telecommunication Signaling Activity Data at District Level

This data is acquired through an online application on Taiwan's Ministry of Interior's Social-Economical GIS (SEGIS) platform. The data is based on the number of telecommunication signals present within each of the districts at different times of days on different days of the week in 2020. The data set is able to indicate the approximate number of people in each of the districts based on cell phone activity, and this data set that is available and used was based on activity in November 2020. Although the global pandemic occurred in early 2020, Taiwan was not significantly affected by the pandemic at this point. Therefore, their differences in the cell phone activity recorded would not be significant.

3 DATA PRE-PROCESSING

Although, the original Housing Sales Price data set acquired from Taiwan's open data platform had column titles in English, most of the data were recorded in Traditional Chinese characters. Therefore, the first step required was to translate all features and entries of the data into English. The original data set consists of 31 features columns and many of it has a larger number of missing fields or are not relevant to all entries, thus redundant columns have been dropped and most data other than address has been translated into English, as the addresses would be later transformed into pairs of longitude and latitude.

The following section would include specific actions taken to pre-process each of the columns in the Real Estates Sales Price data.

1. Transaction filter: Select transactions that include only residential properties, and exclude transactions that only consist of land
2. Date: Converted the date record in Republic of China calendar format into Common Era years.
3. Number of Land/Building/Parking Transacted: This column originally consist of three different numbers in a string format, extracting information for each of the features into separated columns.
4. Transacted Floor: Transactions with multiple floors would use the max floor number as the floor entry.
5. Type of Building: Categorized into different categories, including Apartment (Below 5F without Elevator), Apartment Building (Below 10F with Elevator), Factory Office, Office Building, Residential Apartment/Condo (11F+ with elevators), Store, Studio, Townhouse, Others.
6. Main construction Method: Categories including Brick, Brick and Stone, Prestressed Concrete, Reinforced Brick, Reinforced Concrete(RC), Reinforced Concrete and Reinforced Brick, Reinforced Concrete (Partly SRC), Reinforced Stone, Steel, Steel Constructurer (SC), Steel Reinforced Concrete (SRC), Stone, Wood, Wood and Stone

The education level, income, population, and telecommunication signal data set were more straightforward requiring less translation and pre-processing of the data. In most cases, the data set only required the translation of the district's names into English.

4 VISUALIZATION OF RAW DATA

4.1 City's Housing Layout

In order to better understand the distribution of different residential real estate around the city, visualizations of the distribution of house prices were created to show the areas with higher prices. There are several collections and groups of housing prices in the city, most of the houses in the center of the city tend to have a higher cost per square meter as shown in Figure 1, and there are minor centers all around the city with clusters of medium to high price houses. In addition, the red dots mark the houses with the highest prices in the city, and is most apparent in the center area of the city, with a few that is in areas outside the city center.

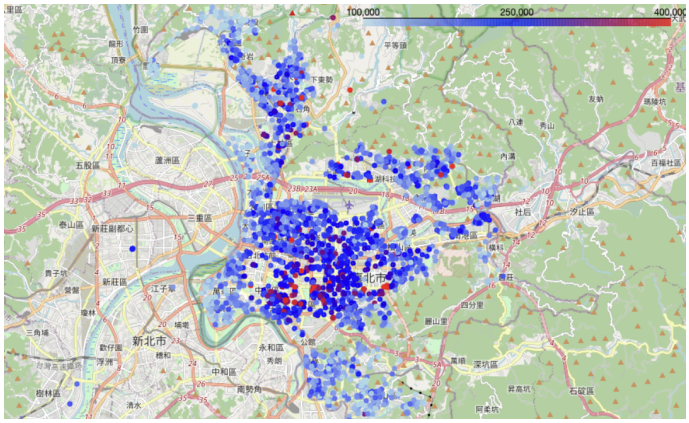

Figure 1. Residential House Prices - Cost per Square Meter

4.2 Density per district

Secondly, we visualized the socioeconmic data we obtained on a districts level as shown in Figure 11 in the appendix and a subdistrict level to relate their geographical distribution to the distribution of residentail prices. From the figure we can see a demonstration of population spread, average household income, and high education levels. An interesting commonality appears here between the three socioeconomic variables that drives our analysis further.

5 MODELING

5.1 Clustering

In order to better understand what factors contribute to housing prices, identifying common similarities between different properties may provide useful information. Therefore, the 4077 residential houses would be clustered based on the following conditions.

- Point of interest at 500 meters, 1 kilometer, and 3 kilometers with price (cost per square meter) (13 features at each distance)
- Point of interest at 500 meters, 1 kilometer, and 3 kilometers. (12 features at each distance
- Residential house properties with price (cost per square meter) (8 features in total).
- Residential house properties (7 features in total)

Point of Interest The K-means clustering method was used in the clustering process, and the number of clusters for POI clusters has been manually set as 4 since, after multiple trials, this number of clusters provides the most interpretable results. The results of the clusters were then assigned back to each of the residential houses and were plotted on Open Street Maps with labels of different colors by using the Folium in Python. The initiate results of the clustering with both the point of interest and the cost per square meter were not ideal, as the price was over-dominant in the clustering results shown in appendix Figure 1. Therefore, the price feature was removed and the residential houses were clustered

based on only points of interest. As seen in the diagram above, there

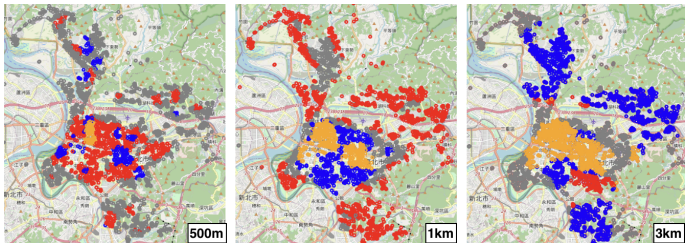

Figure 2. Clustering on POIs at 500m, 1km, and 3km

are small clusters everywhere in the city when only POIs within 500 meters of the radius are considered. However, when the radius expands to 1 kilometer, two centers of cluster 3 appear in the city center, which could be interpreted as different centers of the city. In reality, the orange cluster on the left is known as the historical district, while the orange cluster on the left is closer to the modern-day central business district. When the distance expands to 3 kilometers, it is clear that the city is separated into the inner city center, the middle layer, and the outer circle. This demonstrates that the city center clearly has the highest number of amenities and POIs on average, while the blue clusters have more on average compared to the gray clusters that are usually further away from the city. However, cluster 0 (red cluster) has an abundant number in almost every type of POIs, and this could indicate areas that would potentially have a higher housing price. The distribution of the number of different POIs is shown in Appendix Figure 2, Figure 4, and Figure 5.

Residential House Properties The number of clusters for the clustering on house properties has been manually set as 3 since, after multiple trials, this number of clusters provides the most interpretable results. Again, the initial results of the clustering with both the house properties and the cost per square meter were not ideal, as the price was over-dominant in the clustering results. Therefore, the price feature was removed and the residential houses were clustered based on the house properties alone.

The results in the figure above are not as intuitive and clear at first sight, since the 3 different clusters are spread out in different areas of the city. The only insight that could be directly observed from the visualization is cluster 1 (blue cluster) has the lowest amount of markets and points on the map. This could potentially indicate that the houses belonging to this cluster have properties that are rarer compared to other residential properties in the city.

Therefore, a closer look at the mean for each cluster and the distribution diagram, Figure and Figure 7 in the appendix, is needed to identify the differences in the clusters. As shown in the distribution diagram and the chart, the 3 clusters separate the residential houses into 3 groups that hold unique characteristics. Cluster 2 (gray cluster) is the cluster of houses that would have a lower price since it is usually older, with an average of 28 years of age, smaller in a total area of 83.94 square meters on average, and overall a lower floor. While cluster 0 (red cluster) is the cluster of houses that are in the middle of the market, with an average of 17 years of age, a smaller total area of 186.19 square meters on average, and a higher average floor level than cluster 2. Lastly, cluster 1 (red cluster) is the cluster of houses that would be highest priced, since it is usually newer, with an average of 8.32 years of age, larger in a total area of 391.25 square meters on average, and overall a higher floor level.

5.2 Linear Regression Model

The clustering results provided preliminary understanding and insights into how the number of different types of POIs would classify the residential houses in the city into different groups, and it also provided

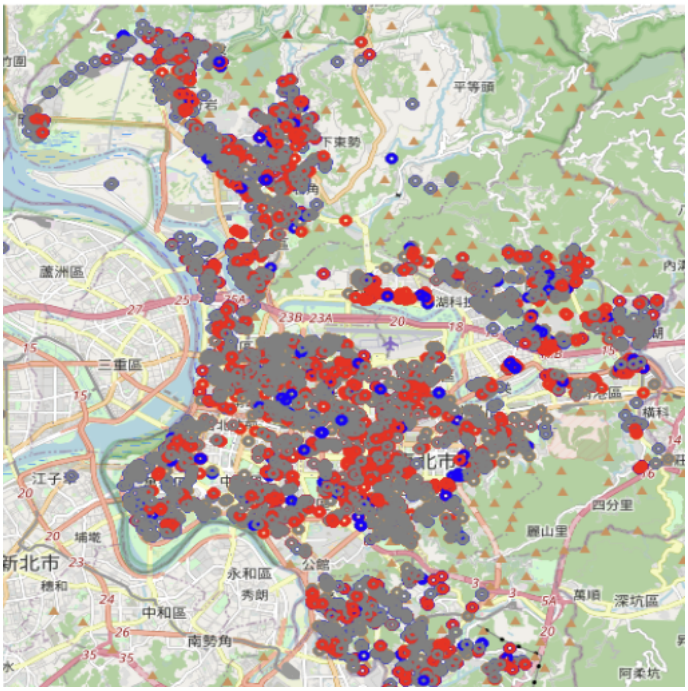

Figure 3. Clustering on Housing Properties

an overview of the layout of residential properties in the city. However, the distribution and clustering of the number of different types of POIs and house properties do not provide a clear explanation and indicator of the significance of each type of POIs. Therefore, a linear regression model would be built to identify the importance of each kind of point of interest and house properties. Similar to the clustering process, the original model includes all 36 POI features, including the three different radius buffer zones, but in order to identify the most influential factors in each of the buffer zones, the features from the same buffer radius have been grouped and a new model was created based on 12 features in every buffer zone radius.

The initial model that consist of all 36 POIs features showed 21 features with positive coefficients, 3 neutral features, and 12 features that had negative significance coefficients, as shown in Figure8 in the appendix. A large number of features were unable to provide a clear picture of how different types of POIs affect housing prices at various distances. Therefore, three additional models were built based on each of the distances defined above as shown in Figure 4 below. As indicated in the diagram, **subway stations** and **bus stations** hold a predominant role and significance in the three distances, while POIs such as libraries, shopping malls, churches, and parks also have a positive impact on the house prices in all three distances. Universities and hospitals have a coefficient close to zero, indicating that it has either a little positive effect or no effect on house prices. However, this is not the case when it comes to nightclubs, supermarkets, police stations, and primary schools. The potential reasons behind the negative coefficient for nightclubs and police stations may be the potential crowd, noise, and unsafeness that are often associated with these locations. However, the fact that supermarkets and primary schools also hold a negative coefficient was surprising and the reasoning behind such a result could be more complex. A potential explanation for this outcome is that the higher the density is for an area, it may require more schools and supermarkets, and in many cases, these areas would have lower house prices compared to areas with lower density. However, an alternative explanation is that amenities such as supermarkets are ideal in a medium distance, as houses right next to supermarkets may also face issues such as a large number of delivery

trucks, customers, and an untidy environment. In addition, the reason behind primary school's negative coefficient requires a deeper look into specific cases. However, an insight that could be drawn from the linear models built on POIs is the fact that transportation facilities such as subway and bus stations hold a significant role in determining house prices. This has inspired us to overlay the route map of Taipei's subway system, Taipei Metro, on the housing price map as shown in Figure 9 in the appendix. The darker the markers are on the map indicates houses with higher prices, and it can be seen that the darker markers often cluster around subway hubs or along subway routes. This reaffirms our understanding that house prices are largely affected by specific types of POIs and in this case transportation POIs.

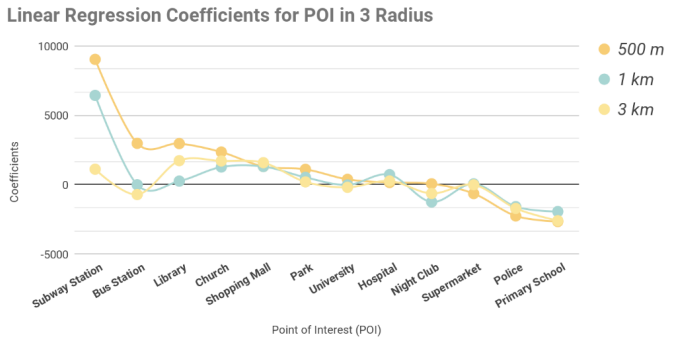

Figure 4. Linear Regression Model - 12 Features at 3 Distances

In addition to the linear regression model based on the different POIs, an additional linear regression model based on the house properties was created to show the relationship between different house properties and their effect on house prices. This would help to facilitate the understanding of the problem beyond geospatial factors. Figure 10 in the appendix shows that the type of houses would usually have a significant impact on the house prices, and townhouses have a high positive coefficient compared to other types of buildings. In addition, construction methods would also impact housing prices, for example, specific construction methods such as steel-reinforced concrete would have a higher positive coefficient compared to other methods. However, it was interesting to see factors such as the number of rooms, floors, and even the ages of the houses do not have a significant impact on housing prices as they all hold a coefficient pretty close to zero.

5.3 Socioeconomic Correlation

The previous sections of the clustering and linear regression model have provided valuable insights into the effect of different types of POIs and house properties on house prices. However, these insights are solely based on the house properties and geographic factors surroundings the houses. Although POIs do provide a certain understanding of the area, different districts and neighborhoods may have different compositions of POIs. In order to better understand the population living in each of the districts and areas and its correlation to housing prices, it would be valuable to evaluate social-economical factors such as but not limited to population, income, education, and human activity. The initial analysis of the social economic factors was done in ArcGIS, as shown in Figure 11 in the appendix, and the results showed promising insights that attracted us to dive deeper into each of the social economic factors. The following section would discuss and evaluate each of the social-economic factors and their relationship with house prices.

Population: As seen in Figure 11 in the appendix, areas with darker shades of brown are areas with a larger amount of population registered in the district. In this case, we can see the district with the highest population is the center of the city, which is also known as Daan District. The second is followed by the district in the top right, which is Neihu District, followed by Shilin District connected on its left side. In order to validate the correlation between population and house prices, the average house prices for each district would be used for comparison. According to the average house price calculated as shown in Figure 12 in the appendix, the district with the highest average house price is Daan District which matches the highest number in the population. However, after that, the order of the average housing prices in each district does not seem to align with the population in each district.

Income: Income is a factor that would determine an individual's purchasing power in their daily life, and this is also seen in the case of residential houses. As seen in Figure 5 below, the right side is a preliminary visualization of the average income in every sub-district in Taipei, and a few hot spots with the highest average income have been marked on the map. An interesting aspect that can also be identified in the diagram is how sub-districts with higher incomes tend to cluster around each other. By overlaying housing price data with the income data on the map, it is easy to recognize that houses with higher prices (marked in dark blue) are usually located in a sub-district with higher average income (marked in dark red). However, this does not signify the relationship between the two factors, as the correlation could happen in both ways. For example, when individuals gain higher income they may choose to move to houses that are more expensive, while individuals living in areas with more expensive houses may have had more resources and opportunities to start with, thus building and buying houses that are more expensive in their home area. The correlation between income and housing prices was also examined in the statistical approach, and the correlation coefficient shown in Figure 15 indicates that the average income of each sub-district has a correlation coefficient of 0.63 with average house prices in the sub-district. **Education:** The data set acquired

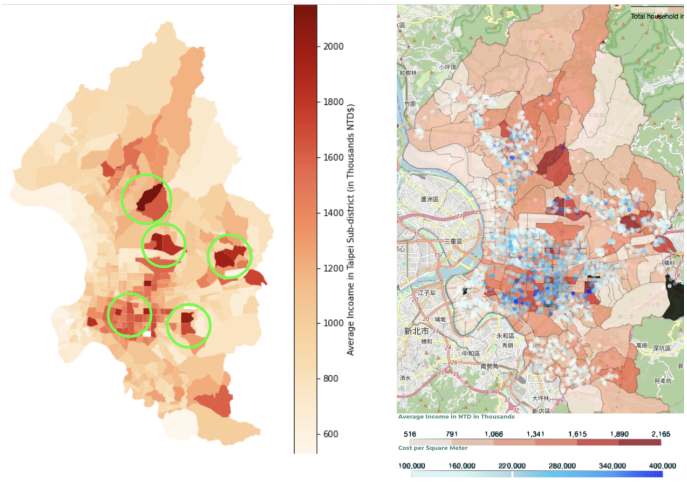

Figure 5. Income per Sub-district and Overlay of House Prices

had a total of 46 features, ranging from primary school education to doctorate degrees. The data set also distinguish if the degree was awarded or incomplete, and the number of male and females for each education level/status. Our initial analysis and visualization have shown that the difference between each sub-district is the greatest at the doctorate level, and the differences slowly dissolve after the undergraduate level. Therefore, the three top degree levels, including a doctorate, master's, and undergraduate degrees have been selected for further analysis. As shown in Figure 6 below, there are several hot spots and sub-districts that can be identified for doctorate and masters level of education, and by overlaying the housing price data on top of it, it can be seen that

the sub-districts with a larger number of people holding higher level educating degrees may correlate with higher house prices to a certain extent. However, the correlation is not as strong as income, since there are many sub-districts with a larger number of people with higher education degrees but do not have houses that are highly priced in the sub-district. However, when the correlation is later examined in statistical methods, the correlation coefficient for any three levels of education examined does not exceed 0.33 at best, as seen in Figure 16 in the appendix.

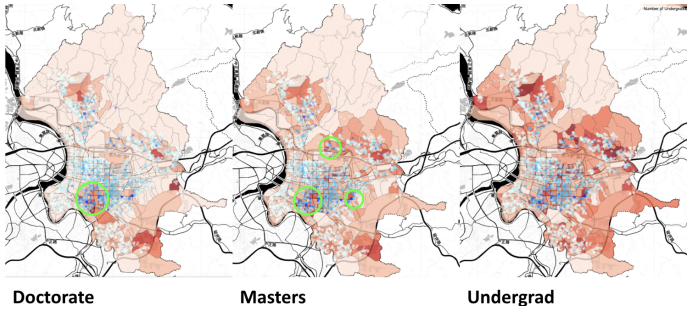

Figure 6. Education Level for Sub-district and Overlay of House Prices

Human Activity: The human activity data set is based on the average telecommunication signaling activity at each of the district levels recorded in 2020. By measuring and observing districts that have a higher inflow of people during the daytime on a workday in comparison to weekends. This would highlight districts with more people working there in the daytime on work days and leaving the district after work. Three different analysis was completed to evaluate the different activity in each of the districts, including comparing the daytime count between weekday and weekend, the nighttime count between weekday and weekends, and the daytime versus nighttime count on a workday. In order to analyze the change in activity in a mathematical approach, a ratio would be calculated for each scenario. For the first two scenarios, the ratio is calculated by dividing the workday count over the weekend, and for the third scenario, the ratio would be calculated by diving the workday daytime count against the workday night time count. The resulting ratio would be a number that is either zero, greater than one, or smaller than 1. When the number is greater than one, it would represent that the district has a higher activity count in the daytime, and when the number is smaller than one it would represent that the district has an activity count in the nighttime.

As shown in Figure 7 below, the areas marked in purple are the areas that have a higher ratio, indicating that more people are present in the area in the daytime compared to the nighttime. By overlaying the housing price over the district activity map, it can be identified that the majority of the house marked in these areas are higher priced as it is marked in darker blue markers compared to areas in brown which consist mostly of lighter blue markers that represent houses that are less expensive.

In addition, Figure 13 in the appendix represents the comparison between nighttime activities on workdays and weekends. Most districts do not have significant differences as most of the districts have more activity on a workday, there is only one district to the bottom left that has more weekend activity than workdays. Figure 14 in the appendix compares the daytime and nighttime activity on a workday, as shown in the diagram, districts on the exterior circle of the city faces would usually have less activity count in the daytime, as most people would travel to work in the city center, while districts in the city center or business districts would have more activity in the daytime compared to the night time on a workday.

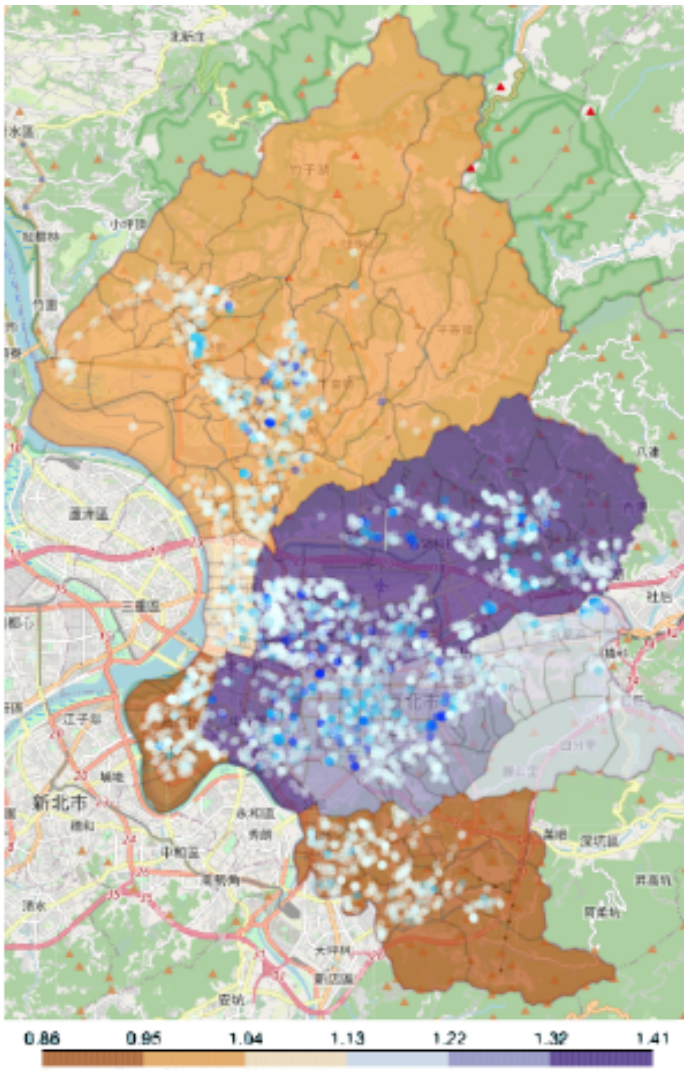

Figure 7. Activity Ratio of Workday Daytime to Weekend Daytime

6 CONCLUSION

The result of the analysis has provided a certain level of understanding and insights, but due to the significant cost that may arise from querying POI data from Google Maps API, the data entries analyzed were still limited in scale. The 4077 entries used in the analysis have been able to demonstrate the potential expandability and usefulness of the method and models. The importance coefficients provided by the linear regression model demonstrate the significance of different POI and house property factors, which could be used by city planners and city governments to improve city planning in the future. The social-economic analysis shows a strong correlation between the sub-district's average income and average housing prices in that sub-district. Although education and population do show a correlation when examined visually, in reality, the correlation coefficient is extremely low. The human activity data based on telecommunications signals also shows the inflow of people into different districts on workdays across the city have certain correlations with the housing prices in each of the districts, where the housings are more expensive in diarists where more people work and fewer people reside.

Overall, our team has concluded that the results from our analysis reaffirm the importance of Transit-oriented Development (TOD)in city planning, as shown by the high importance of transportation POIs in the models. Transit-oriented development would not only maximize the

usage of public transportation, and it could also reduce environmental pollution. In fact, when essential services are well-planned and situated across the city, it would make POIs such as hospitals, parks, and universities less important. On the other hand, housing properties do affect people's choice in purchasing houses, but it does not directly affect the prices of the houses as the POIs do.

7 CHALLENGES, CAVEATS, AND OPPORTUNITIES

Overall, the analysis and insights of this paper would provide a clear direction to policymakers on the need and method to develop a sustainable city. In fact, by allocating different public amenities and services equally access the city, it would help districts access the city to develop at an equal pace, and improve equality, access to housing, and citizens' purchase power. In addition, this would also help to reduce further urban sprawl that is due to the current high housing prices in the city. This could in return reduce the environment that is damaged by infrastructure construction, and reduce air and noise pollution that would be created by commuting.

It is worth noting a few challenges that have occurred during the production of this paper, as they have hindered and limited our analysis in different ways. The cost of for each query request to use the Google Maps API is very expensive, and due to the limited budget of this research, our analysis had to be limited to 4000 properties. However, there are also alternative solutions such as Open Street Maps may also be a valuable source to extract and query POI information.

The data sets used in this analysis are not produced in the same year, this could have minor differences and changes that would affect the analysis. However, after comparing the data with data sets from previous years, our team has determined that the changes in social-economical factors are not significant. In addition, points of interest surrounding a house may change over time, and it is difficult to obtain the changes in the POI data, it would be hard to observe and measure the changes in POI. This may become an issue when data from a long time ago is going to be analyzed, as the POI data queried nowadays would not represent the POIs when the house was transacted. Also, there are a lot of additional POIs that could be taken into consideration, such as stores, restaurants, pharmacies, etc. When more POIs are included in the analysis it would provide a more comprehensive understanding of the topic.

Furthermore, local cultural, religious, and historical factors are not included in this study, and these factors could potentially impact housing prices at unexpected levels.

REFERENCES

[1] Wei-Shong Lin, Jen-Chun Tou, Shu-Yi Lin, and Ming-Yih Yeh. Effects of socioeconomic factors on regional housing prices in the USA. *International Journal of Housing Markets and Analysis*, 7(1):30–41, February 2014.

[2] Lu Wang, Guangxing Wang, Huan Yu, and Fei Wang. Prediction and analysis of residential house price using a flexible spatiotemporal model. *Journal of Applied Economics*, 25(1):503–522, December 2022.

8 APPENDIX

8.1 GitHub

Most of the data sets used in this analysis would be available in the GitHub repository, including the various python notebook files used to produce the final results. The two data sets that would not be available on the repository are the telecommunications signaling activity data set as it requires an individual application to gain access and to use the data set. The links to the origin of the data sets would also be available below, as the updated versions of the data sets may become available for download. Github: https://github.com/Eric-Miao/2022Fall-CIVENG263N-FinalProj/

8.2 Data set sources

- Real Estate Transactions: https://plvr.land.moi.gov.tw/DownloadOpenData
- Individual Income Tax: https://data.gov.tw/en/datasets/103066
- Education: https://data.gov.tw/en/datasets/8409
- Population: https://tinyurl.com/8wdskxdd
- Telecommunications Signaling Activity: https://segis.moi.gov.tw/

8.3 Additional Figures

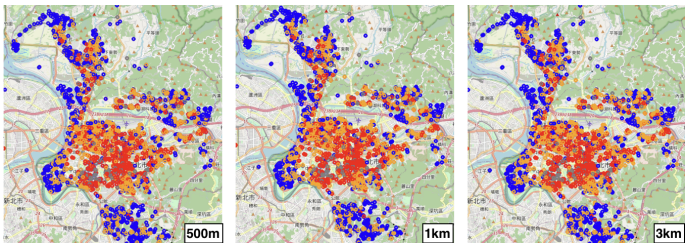

Figure 1. Clustering on POIs at 500m, 1km, and 3km with Price

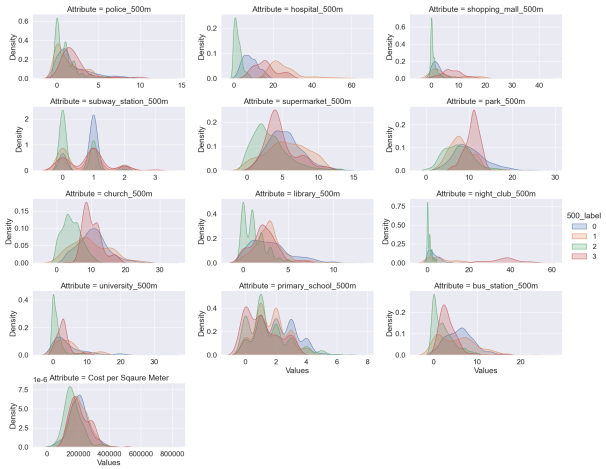

Figure 2. Cluster POI Distribution for 500 m

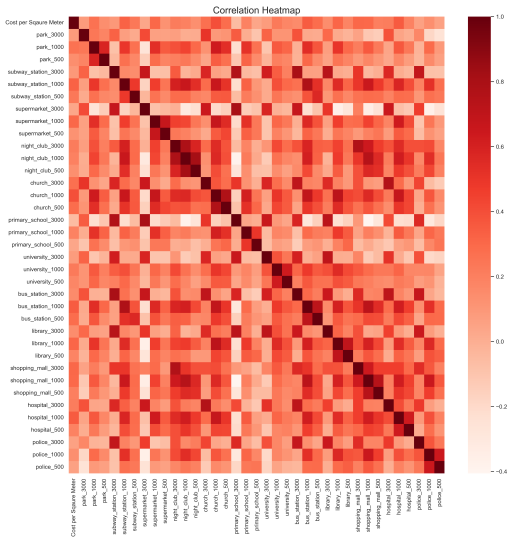

Figure 3. POI Correlation Heat Map

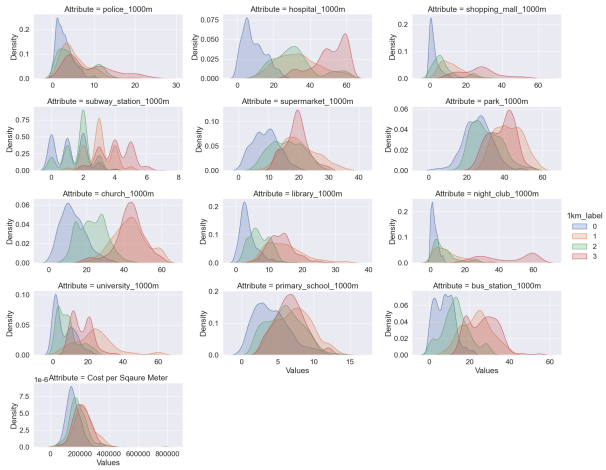

Figure 4. Cluster POI Distribution for 1 km

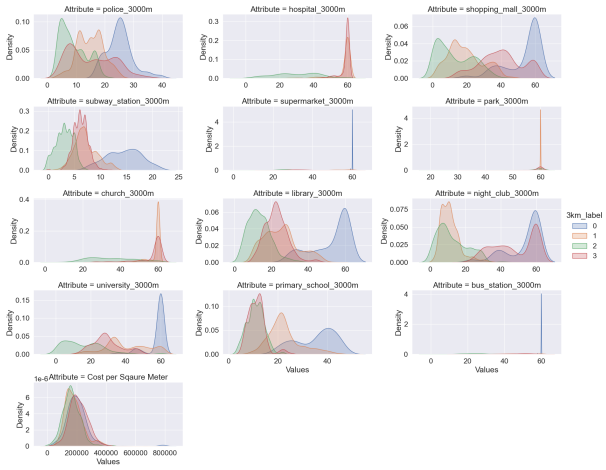

Figure 5. Cluster POI Distribution for 3 km

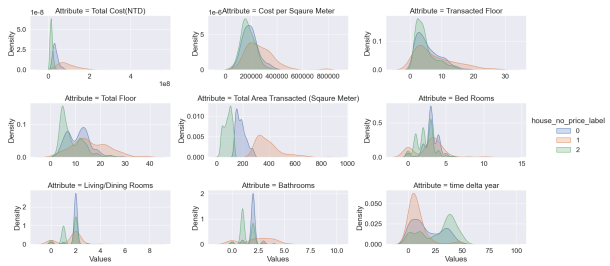

Figure 6. Cluster on House Properties

	Cluster 0	Cluster 1	Cluster 2
Total Cost	34458925	99205981	14354527
Cost m2	193140	269623	175929
Floor	5.6845	7.74	4.80
Total floor	11.318	16.051	8.4
Total Area	186.191	391.259846	83.94
Bed rooms	3.079381	2.77	2.24
Liv/Dine Rm	1.9020	1.64	1.49
Bathrooms	1.948454	2.46	1.39
Age	17	8.32	28

Figure 7. Distribution of Clustering on House Properties

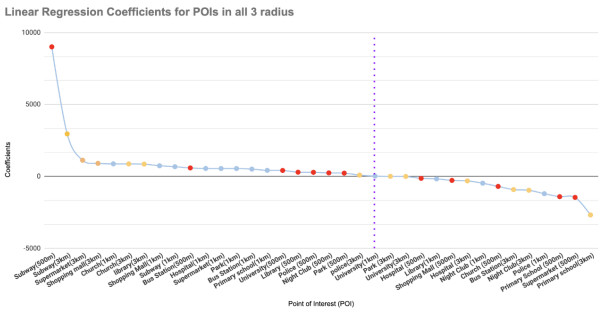

Figure 8. Linear Regression Model - All 36 features

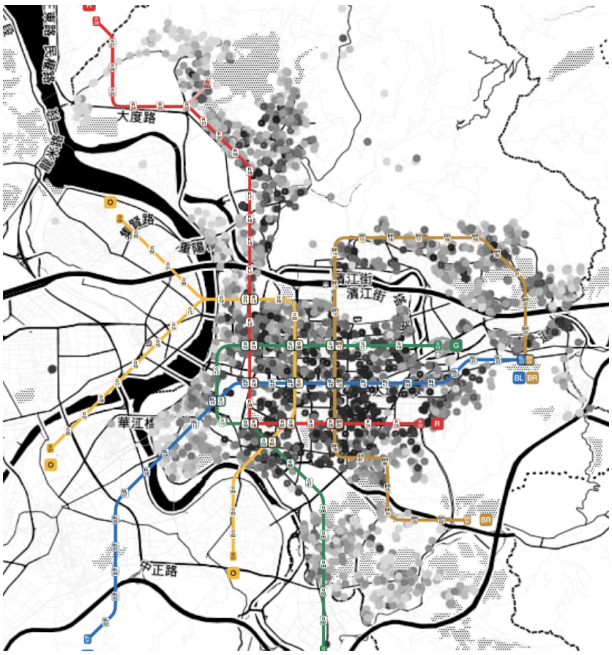

Figure 9. Overlay Taipei Metro's Route Map with Housing Prices

Urban District	Total Cost(NTD)
Daan District	38192.631753
Shilin District	31639.906017
Xinyi District	26505.168804
Zhongzheng District	26243.381543
Zhongshan District	24460.695774
Nangang District	24332.865854
Neihu District	24290.962508
Songshan District	22514.300810
Datong District	21315.409837
Beitou District	19197.829690
Wenshan District	16516.848405
Wanhua District	14218.062552

Figure 12. Average House Prices in Each District

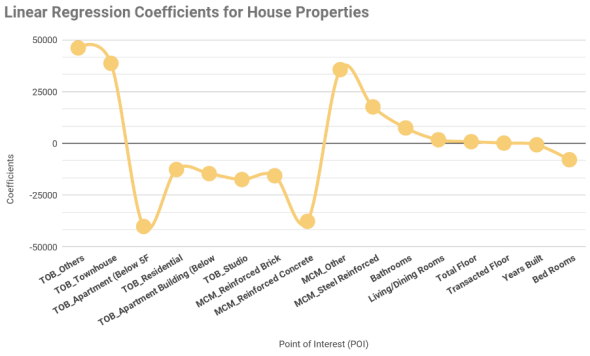

Figure 10. Linear Regression Model - House Properties

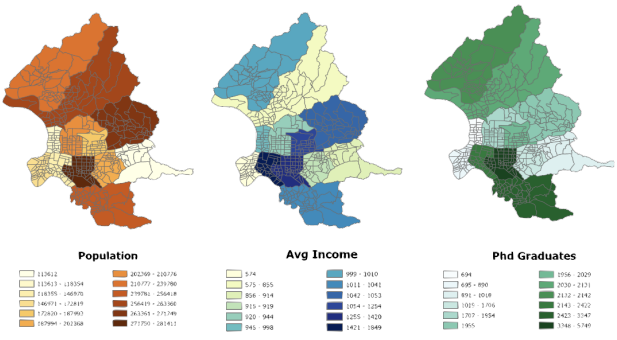

Figure 11. Preview of Social Economic Factors on ArcGIS

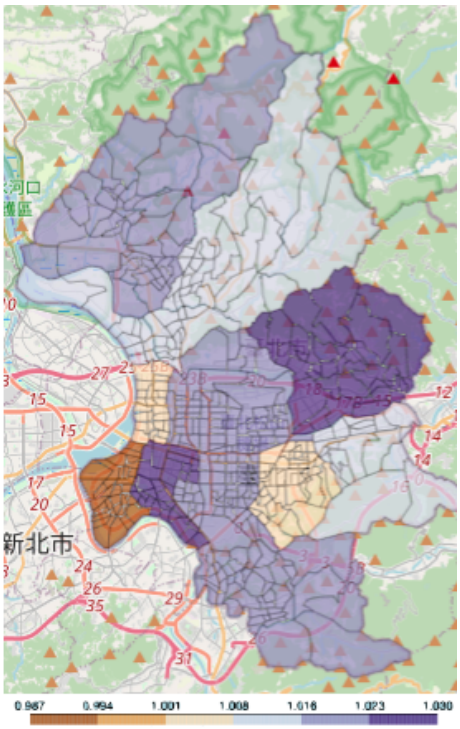

Figure 13. Activity Ratio of Workday Nigh time to Weekend Nigh time

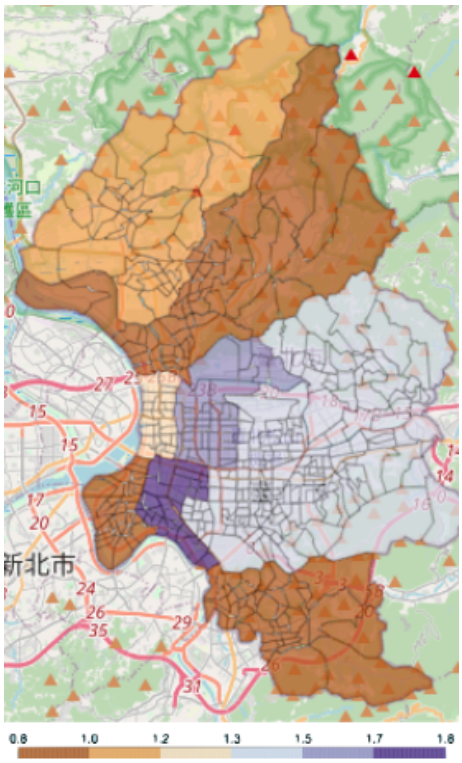

Figure 14. Activity Ratio of Workday Daytime to Workday Daytime

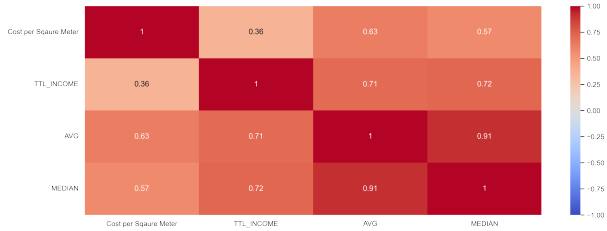

Figure 15. Correlation Heat map for Income vs. Housing Price

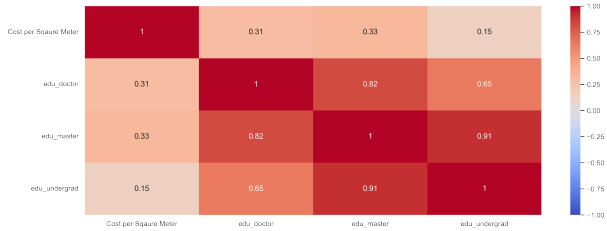

Figure 16. Correlation Heat map for Education vs. Housing Price