# Utilization of Government-Based and Non-Conventional Indicators for Property Value Prediction in the Philippines

Gabriel Isaac L. Ramolete, Dustin A. Reyes, Bryan Bramaskara, and Adrienne Heinrich

Aboitiz Data Innovation

## ABSTRACT

Property appraisal and value estimation in the Philippines are prone to human errors and bias, due to price subjectivity and the lack of knowledge on the impact of surrounding amenities to the property's value. Predictive models for property valuation routinely involve conventional features of the house/area, such as the number of bedrooms and bathrooms, floor and land area, and market prices of nearby properties. The paper suggests that alternative data should be incorporated to account for deviations in true market value and improve property value predictions in the Philippines and other developing countries with similar problems.

The study considers public data (e.g., 2022 property listings, mapping initiatives such as OpenStreetMap) and anchors socio-economic indicators from the Philippine Statistics Authority's open databases to assess its relevance to property value prediction in the Philippines. By utilizing the Department of Trade and Industry's 2021 National Competitiveness Index Rating, this research also investigates the significance of a Local Government Unit's competitiveness based on their economic dynamism, government efficiency, infrastructure, and resiliency.

Methods used to arrive at predicted prices include an exponentially smoothed forecasting of socio-economic data and geography-based feature engineering. The reliability of varying machine learning regression algorithms, from linear, SVR, tree-based, and other ensemble regressor models, were also compared. It was also hypothesized if property segmentation via clustering could improve model performance by grouping similar property listings. The paper aims to understand if including indicators measured by government entities have substantial effects in increasing model performance, as compared to conventional indicators replicable globally within developing countries facing similar issues. The researchers propose that such an approach could lead to lower error rates in Philippine appraisal and minimally biased assessments.

Keywords: property appraisal, spatial analysis, regression, city competitiveness, clustering

## 1. INTRODUCTION

Assessing the value of real estate properties can often be a problematic and iterative process for buyers and sellers. While interested parties may rely on local market information, valuations of similar properties, and the experience of professional appraisers, the number of variables to consider when determining the value of a property is often a source of contestation. Location, home size, usable space, and neighborhood comparisons are common factors considered by most professionals during the appraisal process (Naqvi, 2017; The Danh Phan, 2018; Nallathiga, Upadhyay, Karmarkar, & Acharya, 2019). Other alternative externalities have also been explored both in research and operationalized services – accessibility, public service facilities, commercial places of interest, safety, and livability have all been quantified and explored as possible additional indicators to determine a property's true market value (Wittowsky, Hoekveld, Welsch, & Steier, 2020; Zhang, Zhou, Hui, & Wen, 2018; Chen, Zhuang, & Zhang, 2020; Santos & Jiang, 2020; Buyukkaracigan, 2021).

Traditionally, methods such as the Hedonic Pricing Model, Sales Comparison Approach, the Cost Approach, the Income Capitalization Approach, the Discounted Cash Flow (DCF) Method, and the Gross Rent Multiplier Method have been used to predict property values (Adetiloye & Eke, 2014). *Modern Methods Approach in Real Estate Valuation* note that these methods are preferred in practice as sales information is easily attainable or because future income can be determined (Buyukkaracigan, 2021). The most recent editions of the Philippine Valuation Standards Manual and Malaysian Valuation Standards mention the usage of such methods as those recognized by Valuers and users of valuation (Bureau of Local Government Finance, 2018; Board of Valuers, Appraisers, Estate Agents & Property Managers, 2019).

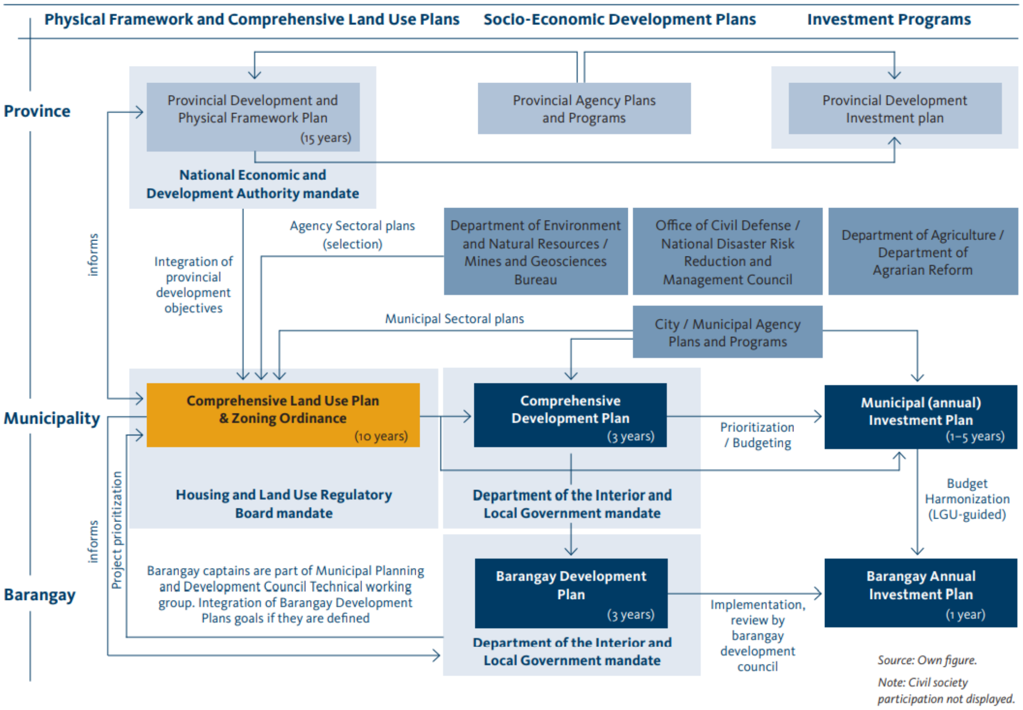
However, some market analyses over periods of time have shown that utilizing these methods can be met with difficulties (Chaphalkar & Sandbhor, 2013; Kershaw & Rossini, 1999; Adetiloye & Eke, 2014). Due to the sheer number of factors to examine, human error and unconscious bias are likely to affect appraised property prices, potentially causing valuation variation (Howard, 2004; Evans, Lausberg, & Sui Sang How, 2019; Yiu, Tang, Chiang, & Choy, 2006; Tidwell & Gallimore, 2014). Traditional methods such as the income approach also do not take non-pecuniary values into account, or may change drastically due to capitalization rates or for urban fringes (Buyukkaracigan, 2021; Tanrivermis, 2016). Other macroeconomic factors affecting property valuations and prices include the country's employment rate, inflation rate, interest rate, income of the local government unit, and poverty incidence of the area (Naqvi, 2017). Most significantly, appraisal bias can occur due to professionals' differing methods and views in the appraisal and valuation of properties. Given that property valuation is a human activity, judgment bias may occur in the form of random and systematic errors which can have a great effect on an investor's decision (Evans, Lausberg, & Sui Sang How, 2019).

In a 3rd world country such as the Philippines, the difficulty of evaluating the price of properties is exacerbated due to the presence of multiple valuation systems imposed by different government entities (Mandani Bay, 2018) (Mandani Bay, 2018). Contributing to this complexity is the lack of proper understanding by local government officials of the pricing of real estate properties in their respective localities. Achieving consistently accurate prices can be hindered by the lack of updated zonal-based fair market values and the presence of multiple valuation systems executed by local government authorities. While most land valuation standards in the Philippines adopt International Valuation Standards (IVS) published by the IVSC, the inadequacies of common valuation methods in the Philippines can still lead to undervalued properties and misinformed decisions on both the appraisers' and buyers' ends (Domingo & Fulleros, 2002).

One major factor of appraisal variation lies in the zonal valuation system in the Philippines, spearheaded by two main entities: The Bureau of Internal Revenue (BIR) and the Local Government Unit (LGU) of which the property is located. According to the Philippines' 1997 Tax

Code, the fair market value of a property is prominently assessed by the BIR. As amended by the TRAIN Law in 2017, the BIR helps LGUs assess the FMV of real properties in each zone or area upon mandatory consultation with competent appraisers (Congress of the Philippines, 2017). These values are subject to automatic adjustment every three (3) years. However, only 60% of LGUs have updated their zonal values in 2017-2020. With this, under the Local Government Code of 1991, assessors in provinces and Local Government Units (LGUs) across the Philippines are required to prepare revisions of real-property assessment and classification every three (3) years. Similarly, only 37% of LGUs have been able to submit updated schedules of market values during the same timeframe (Unciano, 2020).

Why do LGUs find it difficult to update zonal values on time? A 2018 study from the German Institute for Development Evaluation suggests that the current issues of Philippine land planning and management system can negatively affect property valuation (Lech & Gerald, 2018). In order to update roadmaps for identifying zones of regulated land use, LGUs are required to develop Comprehensive Land Use Plans (CLUPs) – a basis for handling the allocation of land resources and properties of an LGU's territory. However, accomplishing the CLUP is highly dependent on the cooperation of different agencies and their often-overlapping mandates; horizontal and vertical frictions occur when dealing with provincial development plans, budget planning, municipal budgeting, barangay development, and other frameworks to be developed in parallel. Figure 1, taken from the study, highlights the interconnectedness of CLUPs and Zoning Ordinances with respect to other plans of varying granularity and importance. Due to often outdated property valuation references, it is common that taxpayers and administrators employ their own strategies and methods of property valuation. Most valuation practices in the Philippines still depend on traditional methods to estimate the price of a property.


Physical Framework and Comprehensive Land Use Plans
Socio-Economic Development Plans
Investment Programs
Province
Provincial Development and Physical Framework Plan (15 years)
National Economic and Development Authority mandate
Provincial Agency Plans and Programs
Provincial Development Investment plan
informs
Integration of provincial development objectives
Agency Sectoral plans (selection)
Department of Environment and Natural Resources / Mines and Geosciences Bureau
Office of Civil Defense / National Disaster Risk Reduction and Management Council
Department of Agriculture / Department of Agrarian Reform
Municipal Sectoral plans
City / Municipal Agency Plans and Programs
Municipality
Comprehensive Land Use Plan & Zoning Ordinance (10 years)
Comprehensive Development Plan (3 years)
Prioritization / Budgeting
Municipal (annual) Investment Plan (1–5 years)
Housing and Land Use Regulatory Board mandate
Department of the Interior and Local Government mandate
Budget Harmonization (LGU-guided)
informs
Project prioritization
Barangay
Barangay captains are part of Municipal Planning and Development Council Technical working group. Integration of Barangay Development Plans goals if they are defined
Barangay Development Plan (3 years)
Implementation, review by barangay development council
Barangay Annual Investment Plan (1 year)
Department of the Interior and Local Government mandate

*Source: Own figure.*

*Note: Civil society participation not displayed.*

Figure 1. Overview of local government administrative planning framework, with land use planning in planning hierarchy (Lech & Gerald, 2018)

As shown, the Philippines has deep-rooted issues in its property valuation system which may make appraisals vary in accuracy and reliability. Other countries have addressed similar concerns by relying on statistical and AI-driven methods and decision support systems to aid appraisers and other parties in determining more accurate values. Studies in Dortmund, Kuala Lumpur, Guangzhou, London, and Shanghai showcase the usage of multiple regression, boosted regression, spatial lag, and geographically weighted regression models as methods to achieve price predictions with reliable accuracy (Wittowsky, Hoekveld, Welsch, & Steier, 2020; Nallathiga, Upadhyay, Karmarkar, & Acharya, 2019; Santos & Jiang, 2020; Huang, Chen, Xu, & Zhou, 2017; McCluskey, Daud, & Kamarudin, 2014). Other machine learning (ML) techniques, such as Random Forest, Support Vector Machine, Gradient Boosting Machine, LightGBM, and XGBoost, have also been utilized in identifying real estate opportunities around the world (Chou, Fleshman, & Truong, 2022; Zhao, Chetty, & Tran, 2019; Baldominos, et al., 2018; The Danh Phan, 2018). Neural networks and fuzzy logic have been used in sales prices of apartments and residential housing values, performing better than traditional methods (Chaphalkar & Sandbhor, 2013; Nguyen & Cripps, 2001; Pi-ying, 2011; Krzystanek, Lasota, & Trawinski, 2009; Lughofer, Trawinski, Trawinski, & Lasota, 2011). Unsupervised techniques have also been utilized to aid these techniques and pre-group properties with similar characteristics (Azimlu, Rahnamayan, & Makrehchi, 2021).

Compared to other countries, the Philippines has not been the subject of such experiments – a hedonic model for house prices affected by COVID-19 infected individuals (Abellana & Devaraj, 2021), and an analysis of determinants of land values in Cebu City (Agosto, 2017) only provide some insight in a Philippine context. Thus, the opportunity of using Machine Learning for property valuation beckons in developing countries like the Philippines, as it may help address the shortcomings of traditional approaches currently being utilized. In combination with using alternative data sources, the outputs of objective ML-based valuations may allow for more accurate and explainable conclusions for buyers, sellers, and appraisers to interpret (Angrick, et al., 2022; van der Hoeven, 2022; Joy, 2021; Rico-Juan & de La Paz, 2021), which is arguably of higher importance in developing countries such as the Philippines.

This paper aims to evaluate the effectiveness of utilizing commonly used ML techniques for the valuation of properties in the Philippines. In addition, as some relatively conventional indicators such as zonal values are not readily available, the paper also seeks to verify the usefulness of alternative data not commonly used by Philippine appraisers and real estate agents. This includes geolocation data sourced from free mapping initiatives like OpenStreetMap, as well as other demographic and socio-economic indicators obtained from the Philippine Statistics Authority (PSA), BIR, and other government resources. The research questions the paper seeks to address are as follows:

- RQ1: Are commonly used ML techniques found in similar property prediction publications also effective under a Philippine context?
- RQ2: Does incorporating socio-economic indicators and geolocation data provide predictive power in the estimation of property prices in areas from the Philippines?
- RQ3: Will the use of indicators measured by government entities have a substantial effect in increasing model performance related to machine learning-based property valuations?
- RQ4: How comparable is the effect of geolocation data to the inherent characteristics of the properties such as floor area and land size?
- RQ5: Can characteristics with LGU granularity still positively affect the accuracy of property price prediction?

## 2. MATERIALS AND METHODS

### 2.1. Data Sources

Four (4) main data sources were used for the study, namely:

- Property listings from Lamudi, a popular online estate listing marketplace in the Philippines
- Department of Trade and Industry's Cities and Municipalities Competitive Index (CMCI)
- Amenities and buildings listed in OpenStreetMap
- Selected socio-economic datasets created by the Philippine Statistics Authority (PSA)

The study consists of information gathered for two primary locations in the Philippines: the province of Cavite in Region IV-A, and Metro Manila, also known as the National Capital Region. These locations were chosen mainly due to their prevalence in Lamudi, a popular real estate listing website in the Philippines.

The models aim to predict the average price per square meter of a property utilizing a combination of factors sourced from these data sources. This will be derived by dividing the given price with the property's given land size. The primary motivation for this is to lessen the variation of performance metric outputs such as Mean Absolute Error (Mean AE) and Mean Absolute Percentage Error (MAPE), as prices of different properties in the Philippines do tend to flare up to huge numbers. While price alone is commonly used in a variety of property valuation papers with ML approaches, price per square meter is an alternative target variable used by other experiments (Gao, Bao, Cao, Oin, & Sellis, 2022; Xiao & Yan, 2019; Ahlfeldt, 2013; Sommervoll & Sommervoll, 2018) and is also commonly used in appraisal of mass real estate (Antipov & Pokryshevskaya, 2012; Thanasi, 2016; Beimer & Francke, 2019; Hau, 2020). In order for the study's results to be more comparable with the aforementioned papers, the land size could be multiplied back to the predicted price per square meter to get a predicted price.

## Lamudi

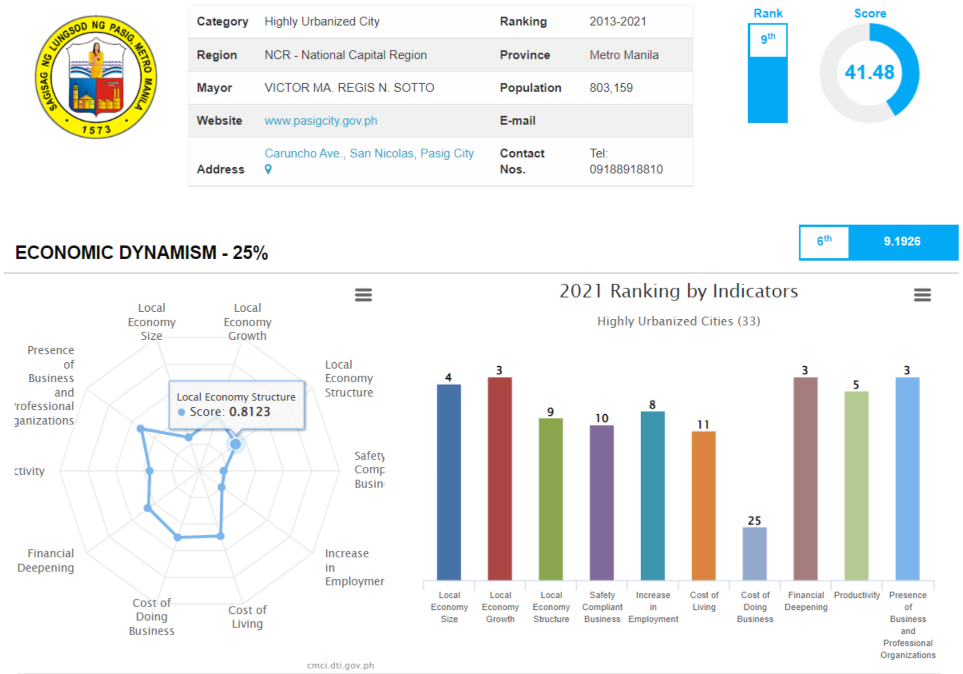
Lamudi-based property listings from Cavite and Metro Manila were collected via web-scraping in Python. Only houses were considered and scraped, as apartments, condominiums, and lots may be characterized or evaluated differently. While Lamudi is considered as a premier online marketplace in the country (Primer, 2021; Similarweb, 2022; Camella, 2022), the limitations of the scraped data include slight inaccuracies with the coordinates, unlisted amenities and furnishments, and distributions skewed to higher-end real estate developers. It is assumed that the scraped data reflects the current state of the housing market in 2022. All prices listed in Lamudi are in Philippine Peso.

| Variable Group | Description | Features |
|---|---|---|
| Location | Features pertaining to spatial characteristics of the property | Longitude, latitude, postcode, LGU, region, subdivision |
| Amenity | Amenities found within the property and its vicinity | # of AC units, balconies, decks, fences, fireplaces, fitness centers, garages, gates, grass areas, libraries/bookstores, airports, parking lots, meeting rooms, parks, pools, basketball courts, tennis courts, volleyball courts, warehouses. Additionally, if the property had security, was smoke-free, or was fully or partially furnished |
| Property Specification | Includes features detailing a property's structural specifications | # of bedrooms, # of bathrooms, floor area ($m^2$), land size ($m^2$), total rooms, property classification, car spaces |
| Price | Market price of property | Price |
| **Total Variables** | **38** | |

Table 1. Variables from Lamudi

## Cities and Municipalities Competitive Index

The Department of Industry and Trade (DTI)'s CMCI, developed by the National Competitiveness Council, is an annual ranking of the competitiveness of all provinces, cities, and municipalities in the Philippines (Department of Trade and Industry, n.d.). The overall competitiveness of an LGU every year is composed of four (4) main pillars of equal weights, namely: Economic Dynamism, Government Efficiency, Infrastructure, and Resiliency. A list of sub-indicators per pillar, each with their own score, is added to create the pillar's final score. Ranks of pillars and sub-indicators were provided. The 2021 rankings of LGUs from Cavite and Metro Manila were scraped from the site; LGU ranks instead of base scores were utilized for the models. Figure 2 exhibits pillar and sub-indicator scores of Pasig City, an LGU in Metro Manila.



| | | | |
|---|---|---|---|
| Category | Highly Urbanized City | Ranking | 2013-2021 |
| Region | NCR - National Capital Region | Province | Metro Manila |
| Mayor | VICTOR MA. REGIS N. SOTTO | Population | 803,159 |
| Website | www.pasigcity.gov.ph | E-mail | |
| Address | Caruncho Ave., San Nicolas, Pasig City | Contact Nos. | Tel: 09188918810 |

Figure 2. Sample 2021 CMCI Score and Ranking of Pasig City (Department of Trade and Industry, n.d.)

| Variable Group | Description | Features |
|---|---|---|
| Pillar Indicators | Ranks scores for the four key indicators | Economic Dynamism, Government Efficiency, Infrastructure, Resiliency |
| Economic Dynamism | Rank scores for Economic Dynamism sub-indicators | Size of the Local Economy, Growth of the Local Economy, Capacity to Generate Employment, Cost of Living, Cost of Doing Business, Financial Deepening, Productivity, Presence of Business and Professional Organizations |
| Government Efficiency | Rank scores for Government Efficiency sub-indicators | Capacity of Health Services, Capacity of Schools, Security, Business Registration Efficiency, Compliance to BPLS standards, Presence of Investment Promotions Unit, Compliance to National Directives for LGUs, Ratio of LGU collected tax to LGU revenues, Most Competitive LGU awardee, Social Protection |
| Infrastructure | Rank scores for Infrastructure sub-indicators | Existing Road Network, Distance from City/Municipality Center to Major Ports, DOT-Accredited Accommodations, Availability of Basic Utilities, Annual Investments in Infrastructure, Connection of ICT, Number of Public Transportation Vehicles, Health Infrastructure, Education Infrastructure, Number of ATMs |
| Resiliency | Rank scores for Resiliency sub-indicators | Land Use Plan, Disaster Risk Reduction Plan, Annual Disaster Drill, Early Warning System, Budget for DRRMP, Local Risk Assessments, Emergency Infrastructure, Utilities, Employed Population, Sanitary System |
| **Total Variables** | **42** | |

Table 2. Variables from DTI's CMCI 2021

## OpenStreetMap

OpenStreetMap (OSM) was used to scrape and count varying types of amenities and buildings within a walking distance of 1, 3, and 5 kilometers away from each Lamudi-scraped property. As an open-source project, the reliability of OSM is verified by millions of contributors that monitor and collaborate in real-time. OSM has been utilized by several large companies globally such as

Apple, Microsoft, and Facebook as the basis for their mapping efforts (Dickinson, 2021). Key-value tags of relevant amenity and building types were found in the project's wiki-site (OpenStreetMap, n.d.). However due to its public availability, one key limitation of OSM is the lack of actual amenities and buildings being documented in less urban areas, at least compared to Google Maps or other paid mapping services. Amenities and buildings tags used were based on indicators commonly found to be indicative of property prices in other publications (Agosto, 2017; Chen, Zhuang, & Zhang, 2020; Gao, Bao, Cao, Oin, & Sellis, 2022; Nguyen & Cripps, 2001; Huang, Chen, Xu, & Zhou, 2017; The Danh Phan, 2018; Wittowsky, Hoekveld, Welsch, & Steier, 2020; van der Hoeven, 2022).

| **Variable Group** | **Description** | **Features** |
|---|---|---|
| Neighborhood Amenities | Count of amenities within walking distance of 1, 3, and 5 kilometers (km) | # of Cafés, Fast Food, Pubs, Restaurants, Colleges, Kindergarten Facilities, Schools, Universities, Gas Stations, Parking Areas, ATMs, Banks, Clinics, Hospitals, Pharmacies, Police Stations, Townhalls, Marketplaces |
| Neighborhood Buildings | Count of buildings within walking distance of 1, 3, and 5 kilometers (km) | # of Residentials, Commercials, Industrials, Retail Stores, Supermarket, Fire Stations, Government Buildings |
| **Total Variables** | **78** | |

Table 3. Variables from OpenStreetMap

**Philippine Statistics Authority**

The study also leverages on the use of statistical data published by the Philippine Statistics Authority (PSA). These datasets include the LGU's income class, type, annual regular income, total capital expenditures, total social services expenditures, poverty level estimates, population, and population growth in five (5) and ten (10) years (National Quickstat for 2022, n.d.; Census of Population and Housing, n.d.; Statistics, n.d.). The expenditures and income variables are in Philippine Peso.

| **Variable Group** | **Description** | **Features** |
|---|---|---|
| LGU Expenditures and Income | Data regarding the LGUs' annual regular income and capital expenditures | Total capital expenditures (2021), Total social services expenditures (2021), Annual regular income (2021) |
| Population and Population Growth | Data regarding the LGU's population and growth rate in 5 and 10 years | LGU 2022 population, 5- and 10-year population growth rate |
| Poverty Indicators | Data regarding the poverty incidence of the different LGUs | LGU Poverty Incidence Rate (2021), LGU Subsistence Rate (2021) |
| **Total Variables** | **8** | |

Table 4. LGU Socio-Economic Variables

## 2.2. Methodology

The study made use of the Python programming language to conduct the data collection, processing, analysis, modeling, and evaluation. Datasets were collected and stored in CSV format. Python libraries used include but are not limited to Scikit-learn, Pandas, Numpy, OSMnx, Seaborn, Matplotlib, Yellowbrick, and Geopandas in data wrangling and modelling. Figure 3 shows an overview of the data approach throughout the study.

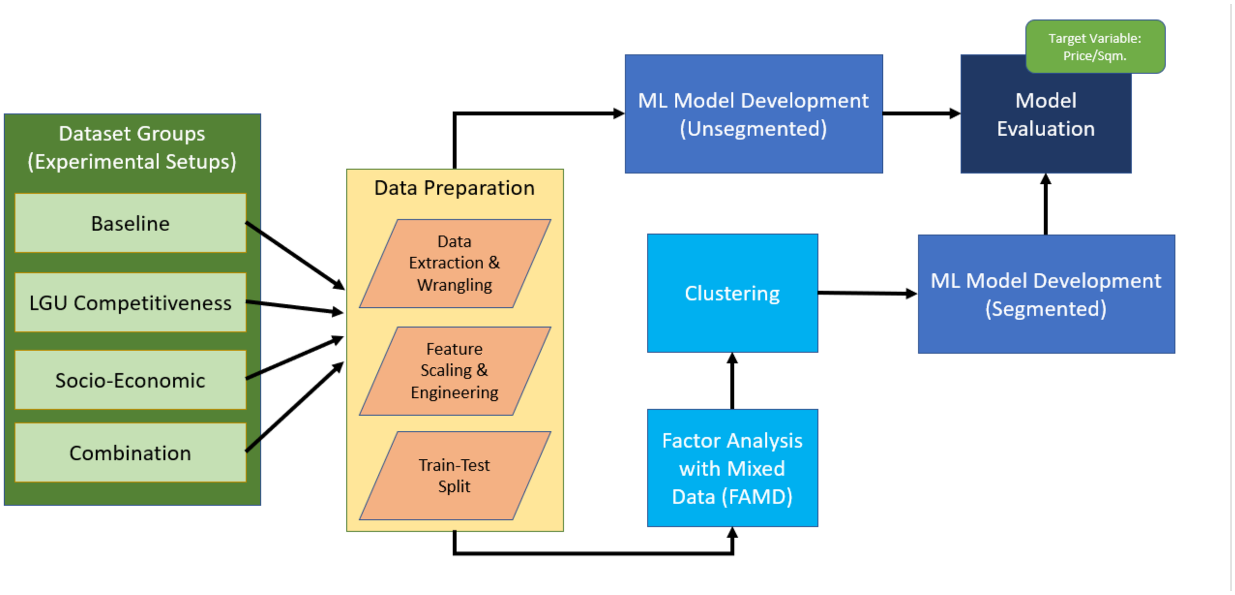

Dataset Groups (Experimental Setups)
Baseline
LGU Competitiveness
Socio-Economic
Combination
Data Preparation
Data Extraction & Wrangling
Feature Scaling & Engineering
Train-Test Split
ML Model Development (Unsegmented)
Model Evaluation
Target Variable: Price/Sqm.
Clustering
ML Model Development (Segmented)
Factor Analysis with Mixed Data (FAMD)

Figure 3. Summary of model development

### 2.2.1. Data Extraction and Preprocessing

Data for property listings in Lamudi were collected via the BeautifulSoup library in Python. Search pages for ‘House and Lot for Sale’ were filtered to LGUs within Cavite and Metro Manila. Duplicates on base price, relative location, and other property specifications were removed for a final total of 2,854 and 14,138 houses for the two locations, respectively. Some amenity objects mentioned in the specific listing pages were removed due to repetitiveness or specificity. To create the target variable, the original price of each property was divided by its land size. The two variables were kept for exploratory data analysis but removed during modelling. the base datasets of Cavite and Metro Manila were divided into 80/20 train-test splits, also preserving this ratio for each LGU.

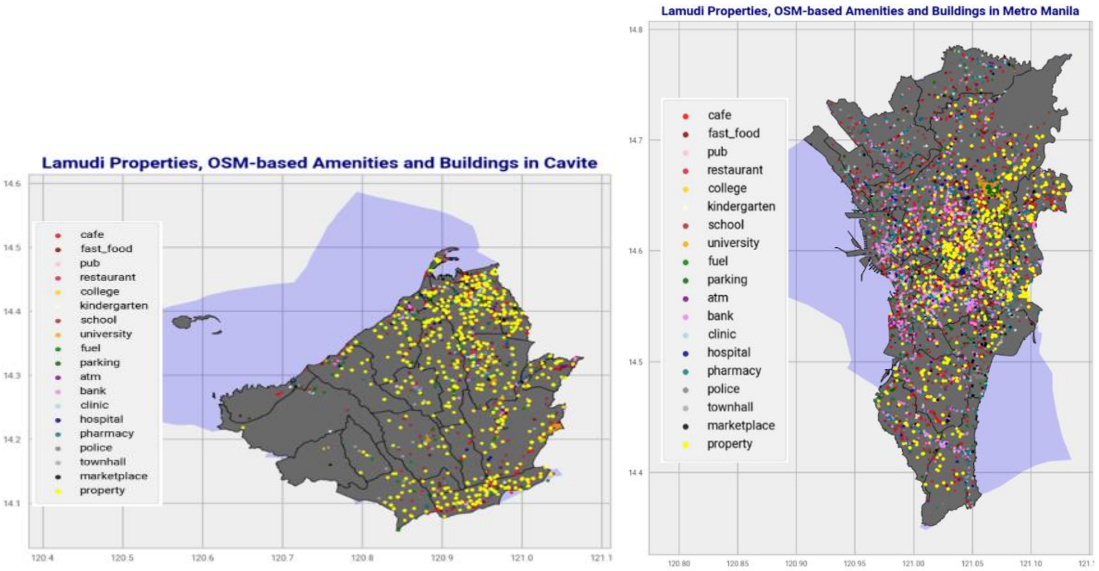
Neighborhood features were extracted via the OSMnx library to query information in the OpenStreetMap database. The two areas of focus were set as input locations wherein all amenities and buildings were extracted. These were then overlayed with the collected property listings to which the walking distance to the amenities and buildings were computed. Counting of values was done and summarized within the vicinity of 1, 3 and 5 kilometers via the K-Nearest Neighbors algorithm. It was noted that properties near the boundaries of the said areas that neighbored other land areas may lack true counts of amenities and buildings.

Reverse geocoding was done via the GeoPy Python library to extract postcode and regional data. These features were used to match the data from the properties to the LGU datasets thus all properties have LGU-level socio-economic indicators. In the Philippines, there are LGUs which have multiple postcodes designated to specific areas. This convention is particularly effective in segmenting areas that may have different demographics and jurisdiction of local authorities. From the dataset, this convention was considered via data wrangling wherein the postcodes were aggregated to its LGU.

Philippine Statistics Authority data which did not have 2021 nor 2022 values were forecasted using Holt-Winters' method (Chatfield, 1978), due to a lack of observed trends or seasonal variations.

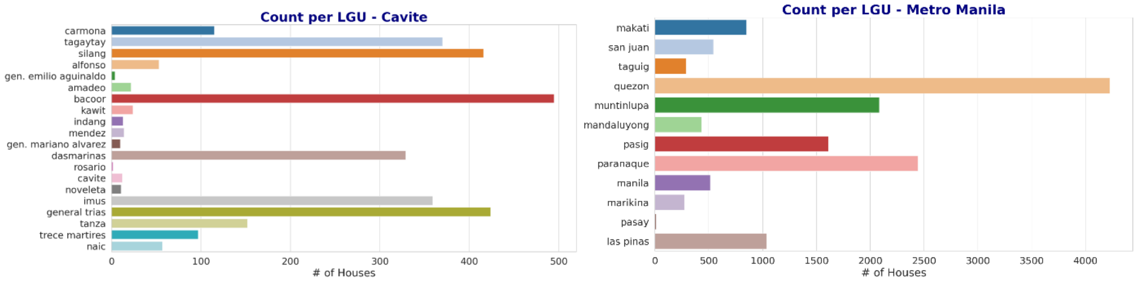
### 2.2.2. Exploratory Data Analysis

Exploratory data analysis was performed separately on the two locations to gain further understanding on the properties scraped in Lamudi and identify potential features to be removed before the modelling phase, in the hopes of increasing general model performance. The analysis was done on the whole dataset before splitting. Figures 4a and 4b show plots of Cavite and Metro Manila with the Lamudi-scraped property listings and the OSM-extracted amenities and buildings in those locations. Denoted on the two maps are the Lamudi locations in yellow.



Figures 4a and 4b: Visualizations of OSM-extracted amenities and buildings with property listings in Cavite (a) and Metro Manila (b); the legend denotes the different entities on the map

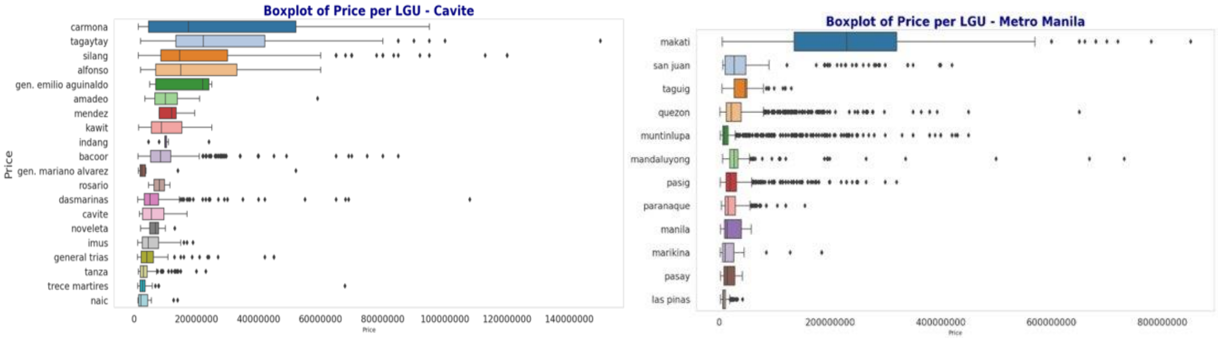
The province of Cavite contains 16 municipalities and 7 cities; of those, 20 LGUs are present in the Lamudi-scraped dataset. The majority of the 2,854 houses in the province are found in Bacoor, General Trias, Imus, Silang, and Tagaytay, which are either component cities or highly populated municipalities. Metro Manila contains 16 cities and 1 municipality; of those only 12 cities are present. The majority of the 14,138 houses are found in Quezon City, with Paranaque and Muntinlupa also having a sizable number of real estate properties.



Figures 5a and 5b. Counts of houses scraped in (a) Cavite and (b) Metro Manila

Despite these distributions with respect to the number of houses present in the datasets, the average prices per LGU in both areas do not follow a similar trend. In Cavite, the LGUs of Carmona, Tagaytay, Silang, and Alfonso have the most expensive houses on average. Component cities which do not feature as highly in Figure 6a, such as Bacoor and Dasmarinas, do have many outlier houses, which may suggest a lack of representative scraped.

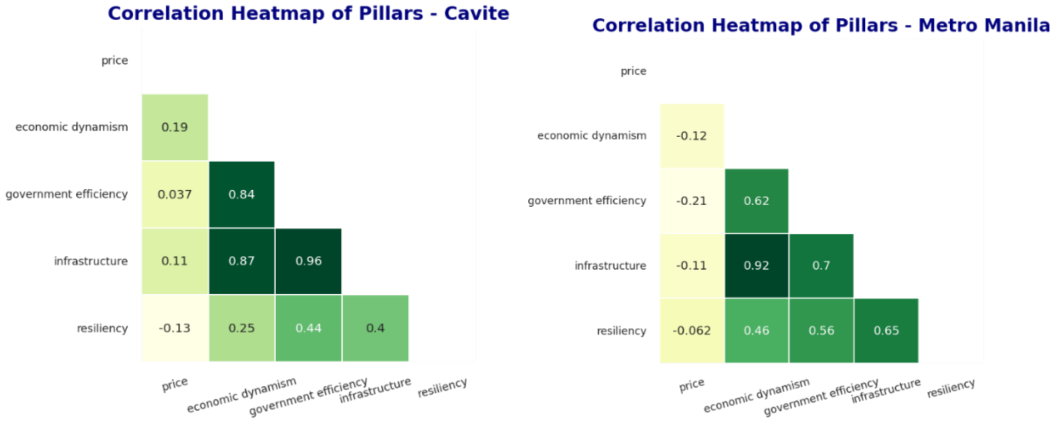
In Metro Manila, the cities of Makati, San Juan, and Taguig contain the most expensive houses. This could be corroborated by the presence of business districts and commercial areas among some barangays within these LGUs. For Makati, as found in Figure 6b, it should be noted that the city is home to many exclusive subdivisions and gated communities, such as Forbes Park and Dasmarinas Village. This is further validated by analyzing the boxplot graphs of house prices per postcode in Metro Manila found in Appendix A.



Figures 6a and 6b. Boxplots of Lamudi-Scraped House Prices per LGU in (a) Cavite and (b) Metro Manila

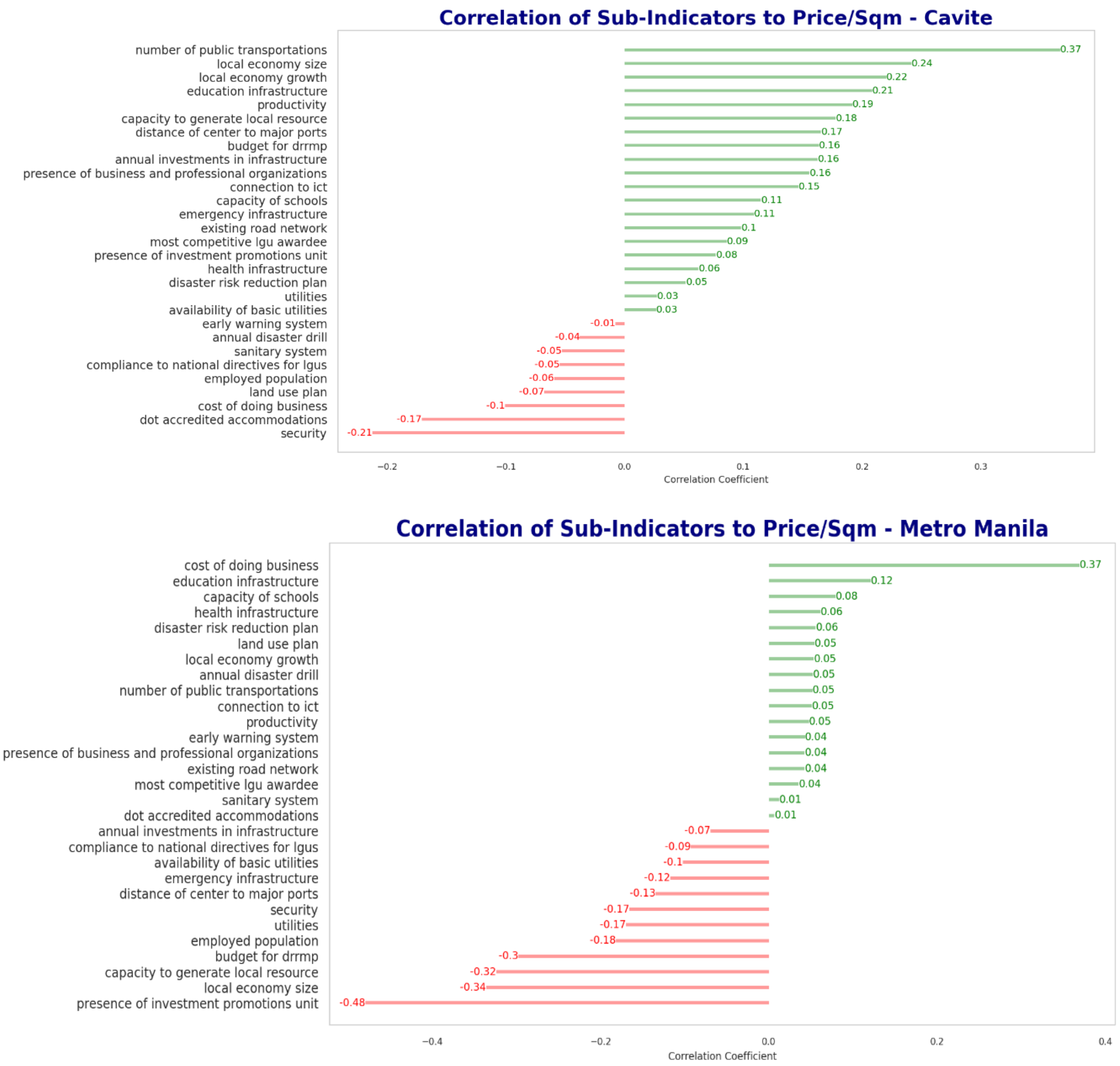
#### 2.2.2.1. CMCI Pillars and Sub-Indicators

To recall, ranks of each LGU were scraped and provided for modelling purposes – as such, the lower the rank and LGU has for a pillar or sub-indicator, the better that LGU performs in that aspect. With this, correlation analysis was performed on both locations.



Figures 7a and 7b. Correlation Heatmaps of 2021 CMCI Key Pillars for houses in (a) Cavite and (b) Metro Manila

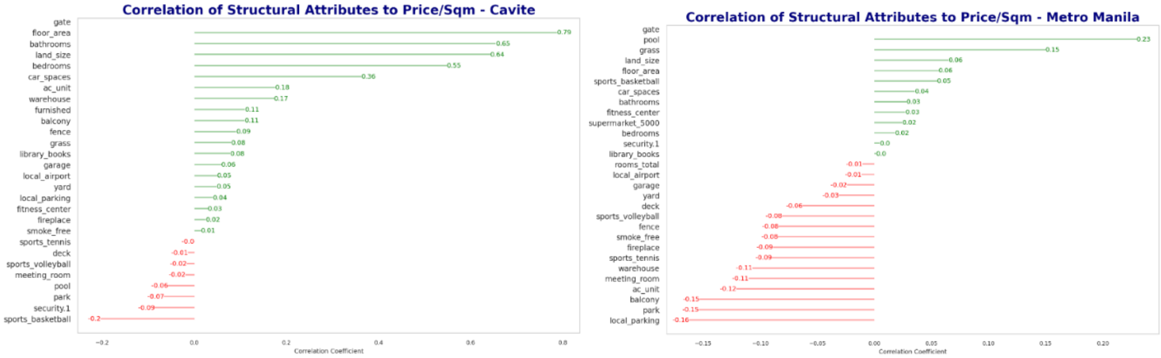
As found in Figure 7a, while pillars show positively correlated relationships to each other, they all have relatively more negligible effects with house prices in Cavite. Only better-performing LGUs for Resiliency increase property prices in Cavite. Diving deeper into the sub-indicators in Figure 8a, it can be seen that ‘Number of Public Transportations’, ‘Local Economy Growth’, and ‘Local Economy Size’ are slightly important indicators in increasing the target variable in Cavite. According to ranks, if there are less options for public transportation, or if the economy of the LGU seems to stagnate, then there are higher chances of having more expensive houses. In Figure 7b, better-performing LGUs in Metro Manila only slightly increase prices. Other than ‘Cost of Doing Business’ and ‘Presence of Investment Promotions Unit’, other sub-indicators do not have as much effect in a more urbanized area such as Metro Manila. In both cases, a number of sub-indicators could be removed from the features to be modelled.



Figures 8a and 8b. Correlation of 2021 CMCI Sub-Indicators to Price/Sqm. in (a) Cavite and (b) Metro Manila

### 2.2.2.2. Lamudi Amenities and Structural Attributes

In Cavite, some structural attributes and Lamudi-based amenities play bigger roles in influencing the price/sq. meter of a real estate property. Standard indicators such as ‘Floor Area’, ‘# of Bathrooms’, ‘Land Size’, and ‘# of Bedrooms positively influence the target variable. Unlike other publications where pools or other furnishments help contribute, the presence of gates and the amount of car spaces available are considered more important, as found in Figure 9a.



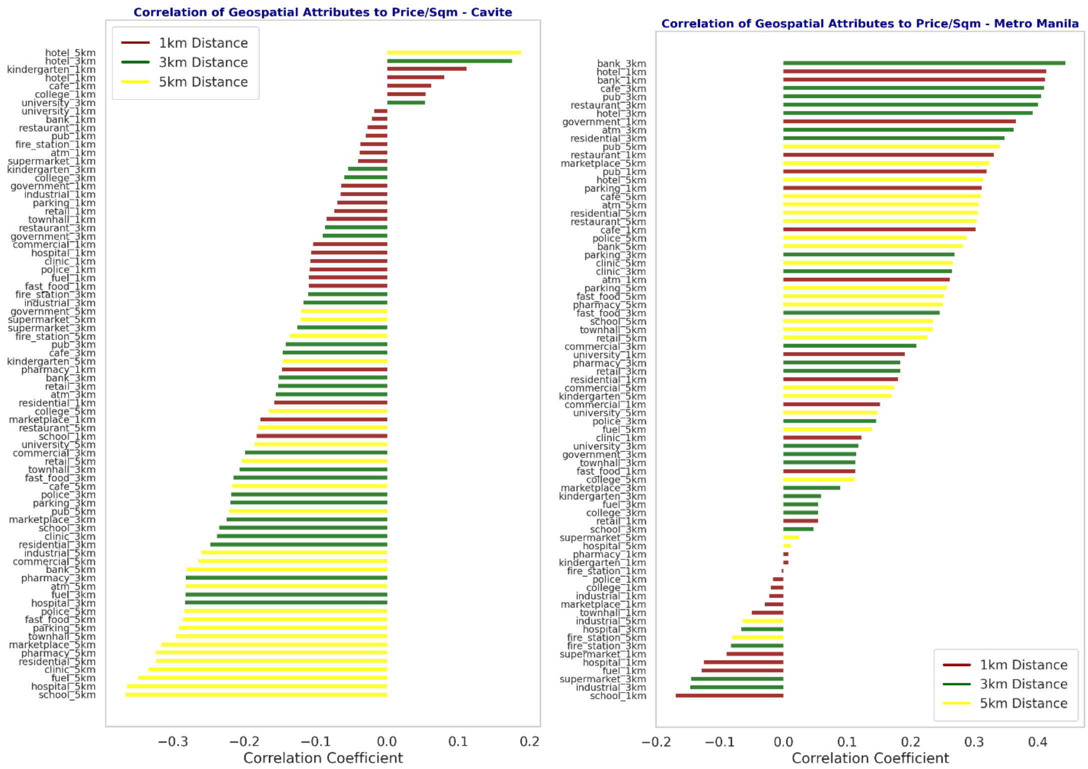
Figures 9a and 9b. Correlation of Structural Attributes to Price/Sqm of Real Estate Properties in (a) Cavite and (b) Metro Manila

Despite this, it was found that in Metro Manila, with the exception of the 'gate' variable, most features do not outright influence the price/sq. meter of properties. No negative nor positively correlated variable passes a 0.3 coefficient, with standard indicators previously mentioned are as close to being not correlated at all. With this, it can be expected that Metro Manila models may perform worse than Cavite, as more intricate factors may need to be considered.

### 2.2.2.3. Geospatial Attributes from OSM

The characteristics of correlations of the OSM-based geospatial attributes to the target variable greatly differ from Cavite and Metro Manila, as found in Figures 10a and 10b. In Cavite, only a few of these attributes positively correlate; notably all hotel-based attributes are in this region. This just may be accredited as an unexplained variance, or the lack of properly documented amenities in Cavite. Other than this, it is apparent that the farther the geospatial attribute is in Cavite, the more likely it negatively affects the target variable – Cavite homes may be in gated communities with which most amenities may not be walkable to.

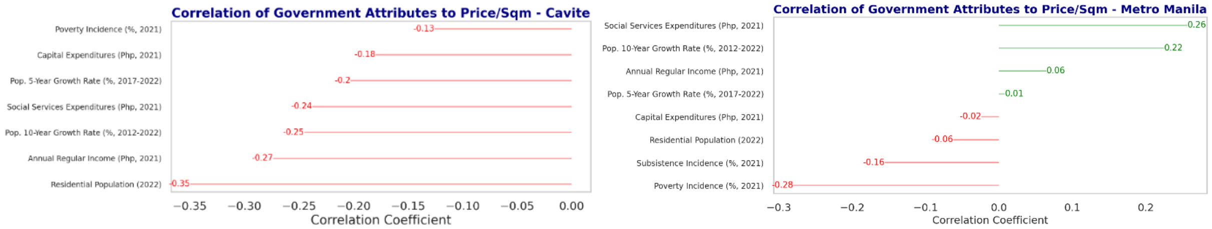
Compared to Cavite, geospatial attributes in Metro Manila are more positively correlated with the target variable, evidenced by coefficients reaching a 0.3-0.4 threshold. Distance between amenities means much more in this highly urbanized context, with 1-km and 3-km variables dominating both ends of the correlation spectrum. Banks within 1-km and 3-km are relatively more correlated than other geospatial variables; the presence of banks or other similar financial institutions are likely well placed as to cater houses with inhabitants of higher income or who are more likely to transact.



Figures 10a and 10b. Correlation Analysis on OSM-based Geospatial Attributes to Price/Sq. Meters of Real Estate Properties in (a) Cavite and (b) Metro Manila

### 2.2.2.4. Socio-Economic Indicators from other Government Sources

Socio-economic indicators from other government sources do not seem to have much correlation with the target variable. With both locations as found in Figures 11a and 11b, increasing poverty incidence slightly negatively correlates with average price/sq. meter; lower prices may be positioned to entice buyers who may not be able to afford more expensive homes. Population in Cavite seems to lower average prices, which may also be correlated with the annual regular income – the higher the population a Cavite LGU has, the lower their average income may be. Social services expenditures have the most positive correlation with the target variable.



Figures 11a and 11b. Correlation Analysis on Socio-Economic Government Attributes to Price/Sq. Meters of Real Estate Properties in (a) Cavite and (b) Metro Manila

### 2.2.3. Experimental Setup for Machine Learning Models

The study implemented two main experimental setups: a non-segmented approach where all houses under a single location were considered in training and testing models, and a segmented approach where unsupervised learning techniques were utilized to segment houses with similar characteristics. Separate baseline models with commonly used features in other scientific publications were set up for the two investigated locations. The baselines were compared with models with combinations of different data inputs to see if prices could be more accurately predicted. Model performances of clusters in each segmented experiment were evaluated against the baselines and non-segmented counterparts.

| Approach | Experiment | Datasets Used |
|---|---|---|
| Non-Segmented | Baseline | Lamudi + OSM |
| | LGU Competitiveness | Lamudi + OSM + CMCI |
| | Socio-Economic | Lamudi + OSM + Government |
| | Combination | Lamudi + OSM + CMCI + Government |
| Segmented | Segmented Baseline | Lamudi + OSM |
| | LGU Competitiveness | Lamudi + OSM + CMCI |
| | Socio-Economic | Lamudi + OSM + Government |
| | Combination | Lamudi + OSM + CMCI + Government |

Table 5. Experimental Setups

As previously mentioned, the base datasets of Cavite and Metro Manila were divided into 80/20 train-test splits, also preserving this ratio for each LGU. These were utilized throughout all iterations of the two experimental setups.

### 2.2.4. Feature Design and Selection

A summary of variables extracted and engineered, as detailed in Tables 1-4, can be found in Table 6. Their data sources would be the basis of differentiating the experiments mentioned in Table 5.

| Variables | Data Source |
|---|---|
| Location | Lamudi, OSM |
| Lamudi Amenities | Lamudi |
| Property Specification | Lamudi |
| Pillar Indicators | CMCI |
| Economic Dynamism Sub-Indicators | CMCI |
| Government Efficiency Sub-Indicators | CMCI |
| Infrastructure Sub-Indicators | CMCI |
| Resiliency Sub-Indicators | CMCI |
| Neighborhood Amenities | OSM |
| Neighborhood Buildings | OSM |
| LGU Expenditures and Income | PSA |
| Population and Population Growth | PSA |
| Poverty Indicators | PSA |
| **Target Variable – Price/sq.m** | **Lamudi** |

Table 6. Variables considered in the study

As 160+ variables were initially available, a set of feature selection processes were conducted before modelling to improve model performance. Pearson and Spearman correlation analysis were done to identify numerical variables with no correlation, which were either dropped or kept note of during modelling. One hot encoding was performed on relevant categorical variables such as 'price conditions', 'income class', 'property classification', 'LGU Type', and 'postcode'. As doing so would further increase the dimensionality of the inputs, variance thresholding and mutual information regression methods were used to decrease the final number of columns used for the machine learning models.

To prepare for property segmentation, a **Factor Analysis of Mixed Data (FAMD)** method was used to create principal components usable for clustering, as a variety of qualitative and quantitative features were present. Guided by rules of thumb (Cangelosi & Goriely, 2007), a 90% variance threshold was used to determine the optimal number of principal components.

### 2.2.5. Machine Learning Modeling

For each experiment, a comparative analysis was conducted on sets of linear, tree-based, and deep learning models. Two clustering algorithms were utilized for property segmentation. Random states for each model used were saved for replicability. A summary of the model development and comparison of segmented and non-segmented approaches is found in Figure 12, while the list of models used for prediction and clustering is found in Table 6. The development of the models was conducted solely on the training sets.

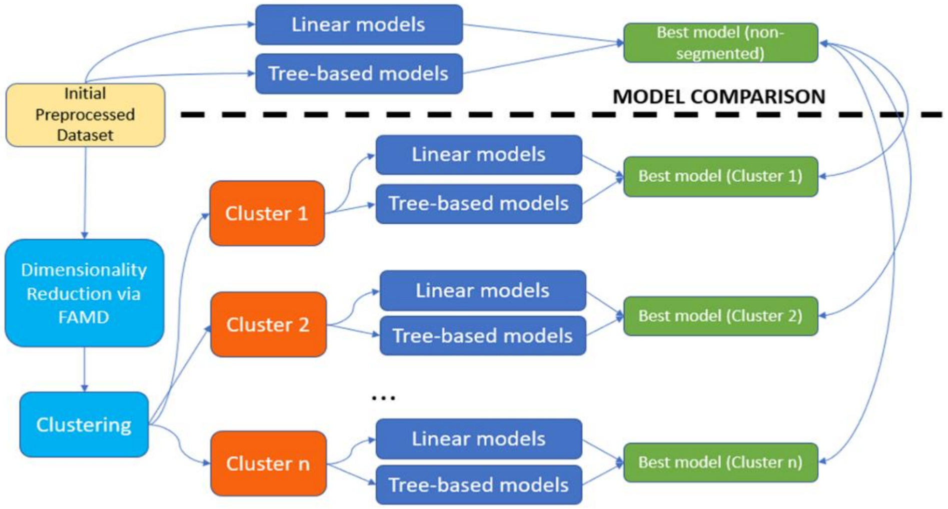


Figure 12. Model development and comparison of segmented and non-segmented approaches

| Model Group | Model Name |
|---|---|
| Linear Models | Ordinary Least Squares (OLS)<br>Ridge Regression<br>Lasso Regression |
| Tree-Based Models | Decision Tree Regressor<br>AdaBoost estimator on Decision Tree Regressor<br>Gradient Boosting Machine Regressor<br>Random Forest Regressor<br>Extremely Randomized Trees Regressor<br>Bagging estimator on Support Vector Regression<br>Stacking Regressor<br>XGBoost Regressor<br>LightGBM Regressor |
| Clustering Algorithms | K-Means<br>Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) |

Table 6. Machine learning models utilized by the study

### Linear Models

Linear models assume that the $x$ (input) and $y$ (output) have a linear relationship. Three of these methods were used as they are easy to interpret compared to tree-based and other non-linear ML methods. While other spatial linear models were utilized in some publications (Wittowsky, Hoekveld, Welsch, & Steier, 2020; Chen, Zhuang, & Zhang, 2020; Santos & Jiang, 2020; Huang, Chen, Xu, & Zhou, 2017; Sario, 2019; Cellmer, Cichulska, & Belej, 2020), the study aimed to focus more on classical ML techniques and argues that the usage of OSM-based data points and CMCI may account for some spatial bias.

**Ordinary Least Squares (OLS)** is a linear method that estimates coefficients of a perceived linear relationship between a dependent variable and at least one independent variable. In **Lasso Regression**, the linear model is penalized for the sum of absolute values of the weights. **Ridge regression** penalizes the model for the sum of squared value of the weights. Outputs not only tend to have smaller absolute values, but also often penalize the extremes of the weights.

### Tree-Based Models

Tree-based models utilize tree-like structures for deciding target variable classes or values and may be useful when input and output variables do not exhibit linear relationships. While commonly used for classification problems, regression trees can obtain numerical values in their terminal nodes by selecting splits that minimize the sum of squared deviations from the mean. Ensemble methods can also be used to produce optimal predictive results through considering weighted scores from sets of weaker classifiers. The models mentioned below were picked mainly due to their prevalence in other ML-based price prediction papers.

**Decision Trees** are basic non-parametric models with hierarchical tree structures used as a data mining method for developing classification systems or predictions. Paths from the tree's root node branch out into internal nodes, which further split via characteristics such as entropy, Gini index, and information gain. Scikit-learn's Decision Tree module utilizes a modified version of CART (Classification and Regression Trees) to develop decision trees. **Random Forest** regressors averages outputs from a number of decision trees of various samples in a training dataset to improve predictive power and avoid overfitting (Ho T. , 1995). In extension, **Extremely Randomized Trees** algorithms is similar to Random Forest, but does not undergo bootstrapping procedures and, as the name suggests, undergoes random splits instead of the more 'optimized' splits the random forest algorithms utilize (Geurts, Ernst, & Wehenkel, 2006). **AdaBoost**, or Adaptive Boosting, is a meta-algorithm and ensemble learning method, which adapts performances of other learning algorithms into a weighted sum (Freund & Robert, 1995). For the study, an AdaBoost estimator was utilized on a Decision Tree regressor. **AdaBoost**, or Adaptive Boosting, is a meta-algorithm and ensemble learning method, which adapts performances of other learning algorithms into a weighted sum (Freund & Robert, 1995). For the study, an AdaBoost estimator was utilized on a Decision Tree regressor.

**Gradient Boosting Machine** regressors utilize ensembles of weaker prediction classifiers to optimize arbitrary differentiable loss functions (Freidman, 2001). Extreme gradient boosting, also known as **XGBoost**, is a scalable implementation of gradient boosted algorithms, taking on a Newton-Raphson root-finding method for accurate results (Chen & Carlos, 2016). Despite its complexity and relatively difficult interpretability, it is favored among other boosted tree procedures due to its shrewd penalization of trees and lead node shrinking. **Light Gradient Boosting Machine (LightGBM)** is another gradient boosting decision tree-based framework which uses Gradient-Based One-Side Sampling and Exclusive Feature Bundling to speed up training processes of standard gradient algorithms (Ke, et al., 2017).

**Bagging** meta-estimators aggregate the predictions of many versions of an initial estimator on random subsets of training data to form a final optimized result. For the study, a bagging estimator was used on a Support Vector Regressor. **Stacking regressors** are another set of methods for combining estimators to reduce biases. An overall estimator cross-validates stacked predictions from a variety of individual estimators (Breiman, 1996). A Gradient Boosting regressor, Random Forest regressor, and an Extremely Randomized Trees regressor were utilized in a stacking approach for modelling, with another Random Forest regressor acting as the final estimator.

**Clustering Algorithms**

After restructuring the datasets using FAMD, a set of unsupervised techniques were used to create property segmentations with which individual models could be tested. The **K-Means** technique is a centroid-based algorithm which partitions $n$ observations into $k$ clusters, with each observation belonging to the cluster with the nearest cluster centroid or mean (Likas, Vlassis, & Verbeek, 2003). The **Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)** technique is a hierarchical-based algorithm known to be effective over large databases, exploiting the idea that data points have different importances across the data space, or 'noisy' points (Zhang, Ramakrishnan, & Livny, 1996). **K-Means** was used for the Cavite dataset, while **BIRCH** was used for the Metro Manila dataset.

## 2.3. Model Evaluation

Together with the 20% split test data, hyperparameter tuning using Scikit-learn's GridSearchCV function with five (5) folds was used to optimize and cross-validate the performances of all models. Common assessment metrics found in other similar publications were utilized to evaluate the performance of each machine learning model. **Mean Absolute Percentage Error (MAPE)** measures the accuracy of forecasting methods by computing the mean ratio between the actual value $A_t$ and forecasted value $F_t$. As it is usually used for measuring regression model quality (de Myttenaere, Golden, Le Grand, & Rossi, 2016; McKenzie, 2011), it will likely be a better indication of true model performance as compared to **Mean Absolute Error (Mean AE)**, which outputs the mean of taking the absolute differences between $A_t$ and $F_t$. However, the Mean AE and **Median Absolute Error (Median AE)**, which gives the median of the same procedure, will give more practical interpretations when read through the unit of measurements used: Philippine Peso / square meters. The $\boldsymbol{R}^2$ **score**, while not sufficient alone in judging a model's regression performance nor for non-linear relationships (Dunn, 2021), can help determine the effectiveness of linear models or be utilized as slight comparisons between other similar publications.

Separate metrics were used for determining the optimal number of clusters used for property segmentation in conjunction with the **elbow method,** a popular heuristic based on the intuition that diminishing returns of a metric are not worth additional expenses. The **inertia** or within-cluster sum-of-squares error method measures the squared average distance between all cluster centroids (Chavent, 1998). The **Calinski-Harabasz** index evaluates the goodness of cluster splits by determining the ratio of the sum of between-cluster dispersion and within-cluster dispersion (Calinski & Harabasz, 1974; Wang & Xu, 2019). The **Silhouette Score** utilizes cohesion and separation of points to clusters, dividing the difference of the mean intra-cluster distance and mean nearest-cluster distance over the greater of the aforementioned values (Rousseeuw, 1987). Utilizing these methods, it was found that four (4) clusters were optimal numbers for both Cavite and Metro Manila contexts. Figure 13 shows results after test runs from 2-12 clusters, while Figures 14a and 14b showcase scatterplots of the first two principal components after re-running the optimal number of clusters in both areas of analysis.

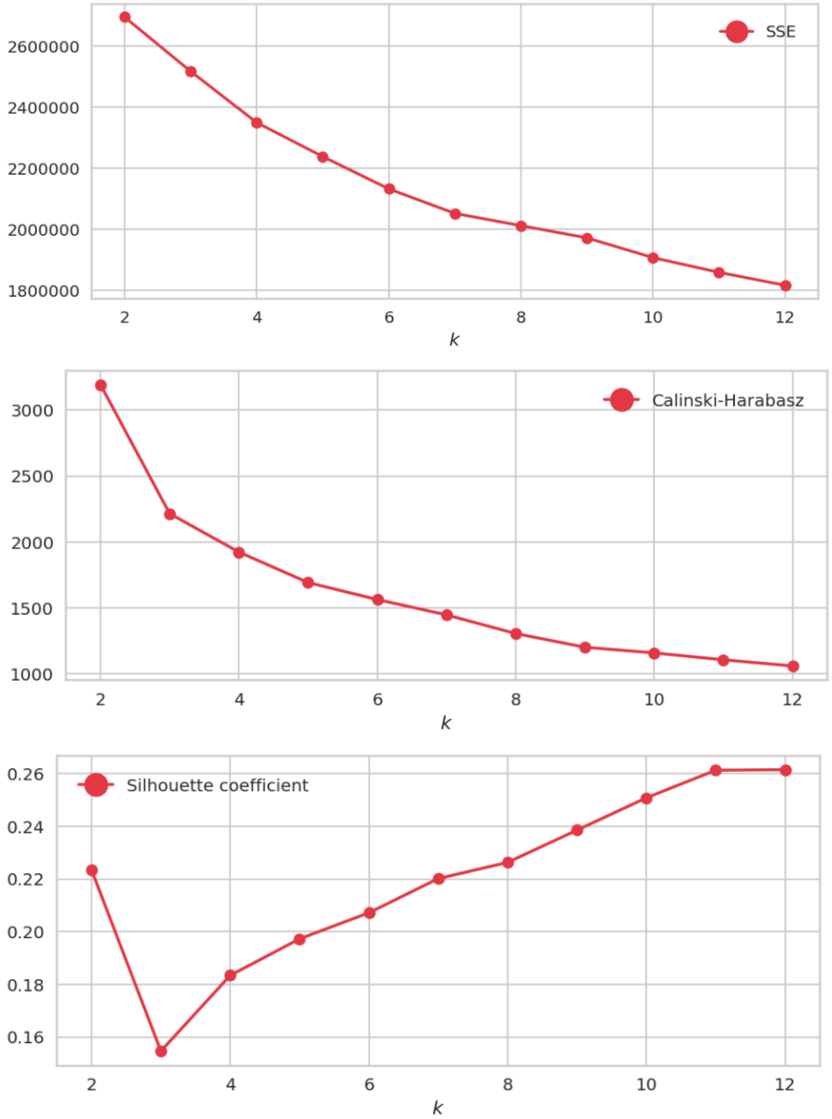
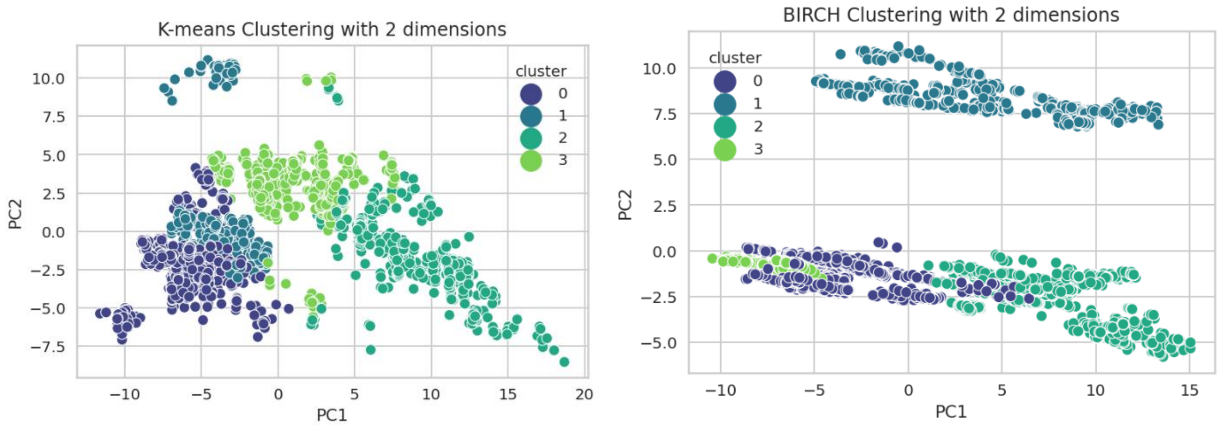


Figure 13. Clustering performance metrics on a FAMD-employed Metro Manila dataset



Figures 14a and 14b. Cluster visualizations on two principal components created during FAMD procedure for (a) Cavite and (b) Metro Manila

## 3. RESULTS AND DISCUSSION

This section discusses our results and findings with regards to the conducted experiments and the five research questions introduced in Section 1.

### RQ1: Are commonly used ML techniques found in similar property prediction publications also effective under a Philippine context?

Our findings indicate that utilizing ML techniques under a Philippine context can be effective depending on the location. As the papers found over the globe differ in currency and in temporal valuation since they were published in different years, the models' performances were compared in terms of MAPE and $R^2$ score. The best Cavite models, both from the non-segmented and segmented approaches, were tree-based models that had MAPEs ranging from 18-22% and $R^2$ scores more broadly from 0.58-0.84. Within the region, regression models on Kuala Lumpur had MAPEs ranging from 11.3-20.9% and $R^2$ scores from 0.74-0.91 (McCluskey, Daud, & Kamarudin, 2014). A Hong Kong study utilizing three ML algorithms outputted MAPEs ranging from 32-54%, but higher $R^2$ scores of 0.83-0.90 (Ho, Tang, & Wong, 2020). Other studies on Shanghai and Xi'an utilize other performance metrics to highlight their results, but provide best $R^2$ scores of 0.70 and 0.89, respectively (Xue, Ju , Li , Zhou, & Liu, 2020). Outside Asia, studies mostly provide RMSE or Mean AE as primary metrics of comparison, which again only leave us with $R^2$ scores as an unreliable basis. A study in Santiago, Chile tested Random Forest, SVM, Linear Regression, and Neural Network models which had scores ranging from 0.74-0.96 (Masias, Crespo, Valle, & Crespo, 2016). A Dortmund study which utilized OLS and Spatial Lag models had adjusted R2 scores of 0.35-0.60 (Wittowsky, Hoekveld, Welsch, & Steier, 2020). A spatial analysis on London real estate prices achieved an R2 score of 0.7116 (Santos & Jiang, 2020). While R2 scores generally aren't the best basis for comparison, it could be said that tree-based machine learning techniques do generally perform better than linear models, and that Cavite's initial models are up to research standards.In contrast, Metro Manila's best models could only output MAPEs of 50-59% and R2 scores of 0.71-0.87. It could be said that initial ML approaches were not sufficient for Metro Manila partially due to the granularity of alternative data utilized or made available. While other aforementioned papers had estimated indicators available by distance, the granularity of a good chunk of Metro Manila's used indicators were only at an LGU or municipality level. It should also be noted that the approach to utilize such a granularity was only enabled by utilizing a cluster of cities and municipalities, while other papers have utilized a single city and its suburban extensions.

With this, it could be argued that the performances of both Cavite and Metro Manila, and in extension other locations in the Philippines, could improve with the availability and utilization of indicators at a distance level of granularity.

### RQ2: Does incorporating socio-economic indicators and geolocation data provide predictive power in the estimation of property prices in areas from the Philippines?

Our results show that socio-economic indicators were considered less important features compared to other data features such as geolocation, area competitiveness and house characteristics. It is hypothesized that utilizing socioeconomic indicators might seem either too outdated or future looking. In the Philippines, data gathering varies per how large the population is. Census data, for example, are taken every five years while the Family Income and Expenditure

Survey (FIES) is taken every three years. The real-estate market is fast-paced and can be regarded as volatile given a particular time period and the socio-economic data used in the study might be irrelevant to property price valuations.

Based on feature importance, geolocation data provide better predictive power when it comes to the task of property price prediction. This is attributed to the “neighborhood effect” wherein houses within a vicinity might have valuations or prices similar to the property in observation. The difference would then be the inherent characteristics of the houses which provided even better feature importance to the model for this task.

**RQ3: Will the use of indicators measured by government entities have a substantial effect in increasing model performance related to machine learning-based property valuations?**

The data setups which included government-based data (i.e., CMCI and socio-economic datasets) beat their baseline models for the non-segmented approach. Cavite’s best baseline model had a 22.10% MAPE, to which it was beaten by the best LGU Competitiveness, Socio-Economic, and Combination models with a 20.41%, 21.59%, and 20.70%, respectively. Metro Manila’s best baseline model had a 58.86% MAPE, a Php 144,697 Mean AE, and a Php 41,266 Median AE. The LGU Competitiveness data setup partially improved on this with a 57.15% MAPE and a Php 140,932 Mean AE but had a larger Median AE of Php 41,471. Similarly, the best Socio-Economic setup showcased a 58.22% MAPE and Php 37,692 Median AE, but worsened with a Php 149,307 Mean AE. The best Combination setup had a MAPE of 54.66% and Median AE of Php 38,218, but worsened with a Php 145,320 Mean AE.

The best segmented approaches mostly improved on their best non-segmented counterparts, especially in the subpar Metro Manila setups, which is discussed in detail in Section 3.2.

**RQ4: How comparable is the effect of geolocation data to the inherent characteristics of the properties such as floor area and land size?**

Both the property specification and geolocation data were highly significant in predicting the target variable, based on the best performing models of Cavite and Metro Manila. Compared between the two areas, the best performing model of Cavite utilized location data more often than the best performing model in Metro Manila. Across both areas, the characteristics of the houses still mostly contribute largely to the predictions. The inherent characteristics of the properties still affects the prices of the properties in these house listings because real-estate sites such as Lamudi (from where the house listings were from) mostly highlights these physical characteristics of the properties. Location and neighborhood play an important role in house prices when buyers are particularly focused on quality of life and income opportunities.

**RQ5: Can characteristics with LGU granularity still positively affect the accuracy of property price prediction?**

A more granular data as compared to LGU, for instance barangay level would be the best approach to have in this case as data with LGU Competitiveness has better impact with best performing model obtained in Cavite with non-segmented format with LGU Competitiveness index data setup as compared to Socioeconomics (mixed of LGU and city levels) data format. In addition, the most significant features for prediction in the best-performing models were property specification and location variables, which were at a granularity lower than LGU. That being said, LGU granularity-based variables did improve model performance albeit not with a substantial effect as what may initially be expected.

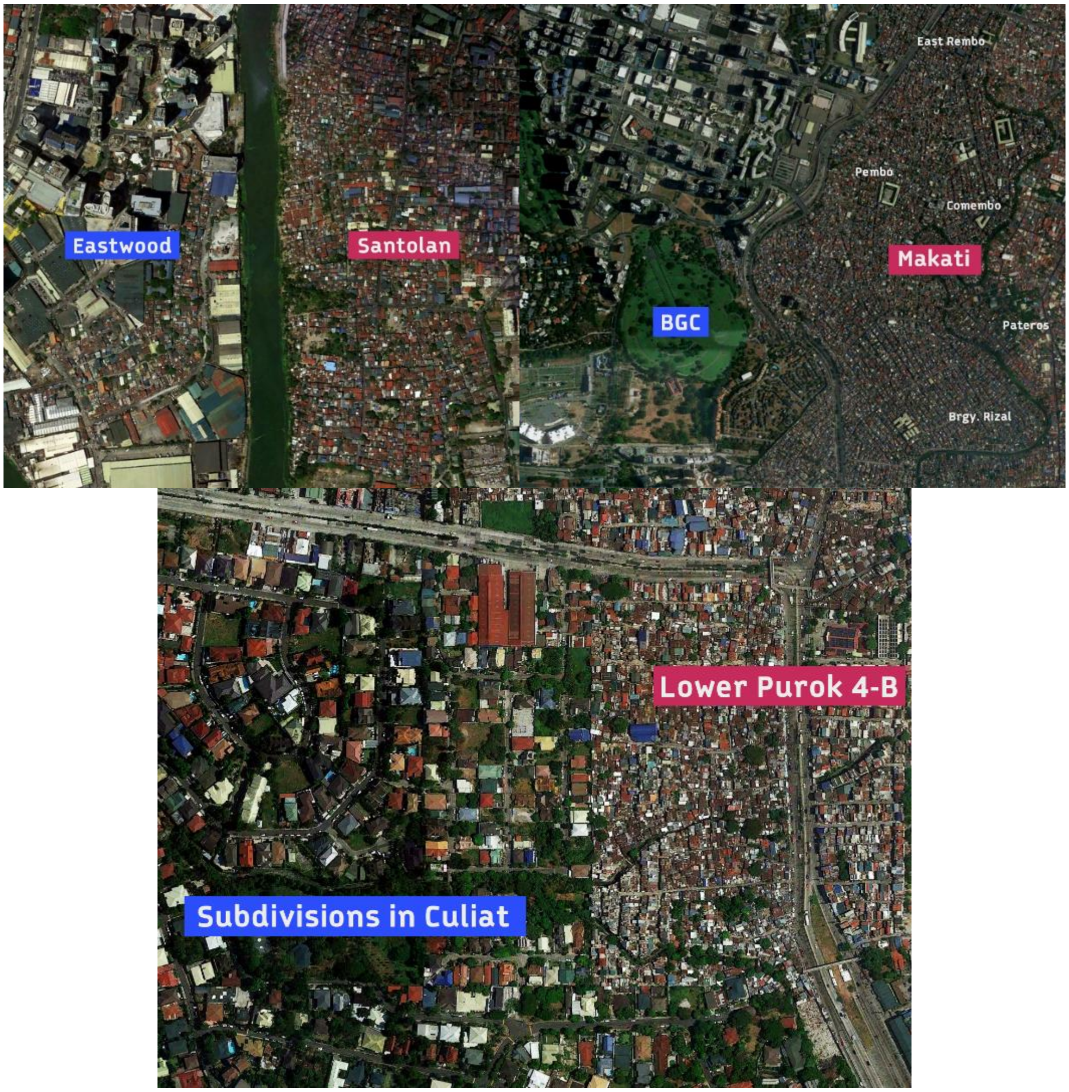
### 3.1. Non-Segmented Approach

The different data/experimental setup yielded results that vary accordingly to the area of analysis in the Philippines. These data setups were fed into different varying machine learning algorithms with the specific goals for the two approaches the study has observed. For the non-segmented set-up, the study aimed to identify the best performing model when applied with incremental addition of data. The results were summarized in Table 7 wherein the best performing models per data setup according to MAPE for Cavite and Metro Manila were highlighted in green. As observed in Table 7, the AdaBoost algorithm performed most reliably as compared to other algorithms. Linear models, regardless of the location or data setup, did not outperform most if not all tree-based models; a lack of spatial consideration found in other spatial econometric models could explain the under-performance.

In Cavite, the AdaBoost algorithm achieved MAPE values in 20-21% when used with data from CMCI and the government, with other tree-based models generally ranging from 22-32%. In comparison with the best Cavite Baseline model MAPE of 22.10%, all inclusions of alternative data improved model performance, with LGU Competitiveness, Socio-Economic, and Combination experiments showcasing MAPEs of 20.41%, 21.59%, and 20.70% respectively. Interestingly, the LGU Competitiveness performed the best out of all models, which may mean that the socio-economic data utilized for Cavite is not granular enough to significantly increase performance. Only the AdaBoost models were analogous to property prediction models in Kuala Lumpur (McCluskey, Daud, & Kamarudin, 2014). The Mean AE and Median AE values look to be acceptable for certain ranges of land sizes but may unnecessarily increase prices for LGUs or areas with lower average land values.

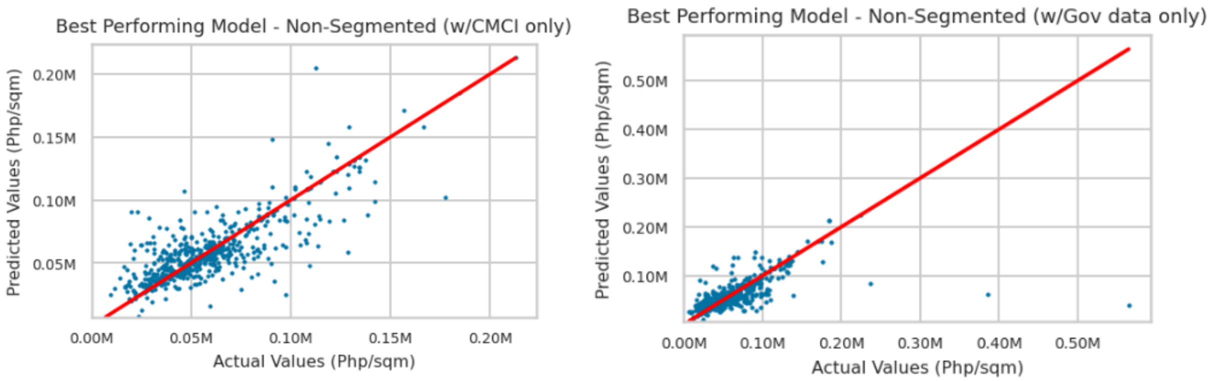
Metro Manila's models performed considerably worse compared to Cavite's and other efforts in similar publications for places outside the Philippines. As shown in Table 7, the best MAPE values in Metro Manila could only reach 54-58% with other tree-based models usually range from 60-75%. In comparison with the best Metro Manila Baseline model MAPE of 58.86%, all inclusions of alternative data improved model performance as well, with LGU Competitiveness, Socio-Economic, and Combination Experiments showcasing MAPEs of 57.15%, 58.22%, and 54.66% respectively. Its Mean AE and Median AE values also exhibited exponential increases, pricing out most potential buyers on average if used.

Metro Manila's subpar performance could be attributed to the lack of more granular data. Scraped data regarding walking distances of 1-, 3-, 5-kilometers to amenities and government buildings were not enough to estimate this variance. Despite its highly urbanized setting, zonal values and average income may highly vary across barangays and populated areas within a single LGU found in Metro Manila. Images from a 2020 article discussing geospatial divides in Metro Manila display the disproportions visually, as found in the differences in Eastwood and Santolan (Figure 15a, Marikina and eastern Quezon City), Pembo, Brgy. Rizal, and Bonifacio Global City (Figure 15b, Taguig and Makati), Culiat and Lower Puroks 4-8 (Figure 15c, central Quezon City) (Commoner, 2020). Recent and trustworthy socio-economic data which may capture these disparities are only readily available at an LGU level. While utilizing the advantages of satellite imaging may help, applying predicted property value to rooftops is a great barrier to overcome.

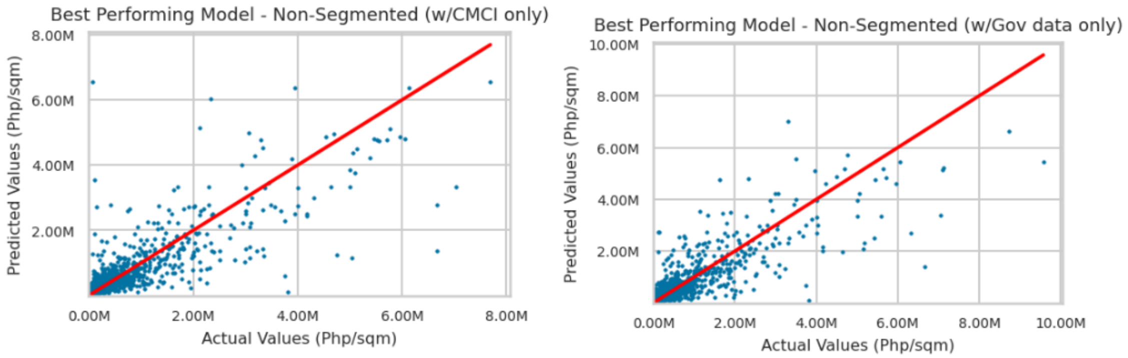


Figures 15a, 15b, and 15c. Satellite images of certain residential and commercial areas across Metro Manila (Commoner, 2020)

A post-model analysis was implemented on the best-performing models of each experimental set-up to understand where improvements could be made, as seen in Figures 16-18. From the graphs, it is seen that results may be skewed by a few significantly misclassified houses. While the majority of houses have good predictions, as verified by their close distances to the red lines in Figures 16-17, a number of them have predictions off by Php 20,000-90,000/sq. meter. As with Figure 18, the models looked to have overfit on the training data – this may explain why a simpler model such as AdaBoost on a Decision Tree regressor performed better in most cases.



Figures 16a and 16b. Non-Segmented Approach – Cavite (Predicted vs Actual)



Figures 17a and 17b. Non-Segmented Approach – Metro Manila (Predicted vs Actual)

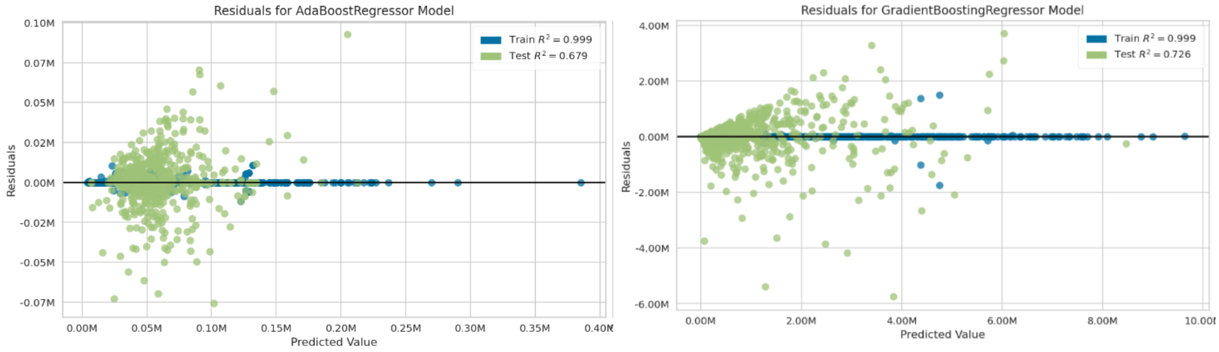


Figure 18. Sample Residual Plot for best-performing non-segmented models for Cavite (a) and Metro Manila (b)

| Non-Segmented Approach | | | Cavite | | | | Metro Manila | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Data Setup** | **Model Type** | **Algorithm** | **MAPE (%)** | **Mean AE** | **Median AE** | **$R^2$ Score** | **MAPE (%)** | **Mean AE** | **Median AE** | **$R^2$ Score** |
| Baseline | Linear | OLS | 46.62% | 20,864 | 14,531 | 0.26 | 376.90% | 349,579 | 182,378 | 0.19 |
| | | Ridge | 46.30% | 20,605 | 14,438 | 0.27 | 376.80% | 349,552 | 182,351 | 0.19 |
| | | Lasso | 46.30% | 20,605 | 14,445 | 0.27 | 376.90% | 349,578 | 182,390 | 0.19 |
| | Tree-Based | Decision Tree | 33.05% | 15,744 | 6,344 | 0.33 | 68.31% | 195,851 | 43,892 | 0.54 |
| | | **AdaBoost** | **22.10%** | **10,774** | **5,000** | **0.58** | 64.88% | 147,023 | 41,800 | 0.75 |
| | | GBM | 27.75% | 11,610 | 5,164 | 0.55 | 63.25% | 165,954 | 40,344 | 0.69 |
| | | RF | 26.33% | 11,545 | 5,705 | 0.60 | 59.90% | 155,228 | 41,382 | 0.73 |
| | | **ERT** | 27.43% | 11,771 | 5,921 | 0.56 | **58.86%** | **144,697** | **41,266** | **0.76** |
| | | Bagging | 44.58% | 19,944 | 11,161 | 0.21 | 130.85% | 241,310 | 71,970 | 0.06 |
| | | Stacking | 26.84% | 12,281 | 5,619 | 0.54 | 59.78% | 150,028 | 41,927 | 0.75 |
| | | XGBoost | 26.59% | 11,892 | 5,426 | 0.54 | 65.57% | 162,875 | 42,361 | 0.68 |
| | | LightGBM | 29.21% | 13,253 | 7,438 | 0.26 | 67.98% | 144,912 | 46,253 | 0.78 |
| LGU Competitiveness | Linear | OLS | 39.28% | 17,450 | 13,510 | 0.38 | 60.71% | 187,381 | 52,050 | 0.52 |
| | | Ridge | 32.88% | 15,283 | 11,114 | 0.50 | 74.29% | 216,570 | 58,593 | 0.37 |
| | | Lasso | 37.01% | 17,010 | 11,510 | 0.37 | 70.52% | 195,810 | 55,690 | 0.42 |
| | Tree-Based | Decision Tree | 33.08% | 16,493 | 8,333 | 0.02 | 67.52% | 176,700 | 45,045 | 0.55 |
| | | **AdaBoost** | **20.41%** | **9,630** | **5,380** | **0.68** | 65.51% | 133,924 | 41,666 | 0.76 |
| | | GBM | 23.72% | 10,766 | 5,470 | 0.61 | 61.39% | 144,719 | 39,471 | 0.73 |
| | | RF | 29.12% | 13,078 | 5,953 | 0.28 | 68.32% | 167,941 | 45,264 | 0.65 |
| | | ERT | 23.19% | 10,695 | 5,606 | 0.64 | 60.65% | 135,536 | 38,033 | 0.75 |
| | | Bagging | 34.97% | 16,293 | 11,493 | 0.41 | 128.00% | 237,270 | 68,443 | 0.19 |
| | | **Stacking** | 26.05% | 12,032 | 7,937 | 0.66 | **57.15%** | **140,932** | **41,471** | **0.74** |
| | | XGBoost | 23.80% | 10,896 | 5,570 | 0.63 | 76.47% | 171,120 | 52,263 | 0.66 |
| | | LightGBM | 23.48% | 10,954 | 6,675 | 0.69 | 62.76% | 134,298 | 42,407 | 0.77 |
| Socio-Economic | Linear | OLS | 35.26% | 16,822 | 12,703 | 0.45 | 62.30% | 188,818 | 51,657 | 0.56 |
| | | Ridge | 33.27% | 15,856 | 10,917 | 0.49 | 64.54% | 189,990 | 52,234 | 0.53 |
| | | Lasso | 35.33% | 16,671 | 11,588 | 0.44 | 66.23% | 192,390 | 53,541 | 0.47 |
| | Tree-Based | Decision Tree | 29.04% | 14,635 | 8,648 | 0.46 | 66.55% | 184,665 | 43,489 | 0.55 |
| | | **AdaBoost** | **21.59%** | **10,241** | **5,633** | **0.70** | 65.81% | 144,497 | 41,779 | 0.74 |
| | | **GBM** | 23.52% | 10,851 | 6,343 | 0.71 | **58.22%** | **149,307** | **37,692** | **0.71** |
| | | RF | 28.25% | 14,149 | 7,868 | 0.49 | 64.36% | 179,674 | 41,659 | 0.59 |
| | | ERT | 25.81% | 11,884 | 7,898 | 0.68 | 61.37% | 145,617 | 38,603 | 0.74 |
| | | Bagging | 32.84% | 14,836 | 10,034 | 0.51 | 136% | 245,210 | 68,398 | 0.16 |
| | | Stacking | 27.15% | 12,577 | 8,658 | 0.66 | 62.70% | 164,017 | 44,992 | 0.67 |
| | | XGBoost | 22.79% | 10,499 | 6,790 | 0.73 | 59.13% | 145,880 | 39,499 | 0.74 |
| | | LightGBM | 25.14% | 11,778 | 7,804 | 0.66 | 60.58% | 143,012 | 41,630 | 0.77 |
| Combination | Linear | OLS | 29.34% | 15,094 | 10,590 | 0.51 | 60.41% | 177,371 | 52,050 | 0.52 |
| | | Ridge | 31.11% | 14,559 | 9,936 | 0.55 | 73.29% | 216,570 | 58,593 | 0.37 |
| | | Lasso | 36.39% | 16,918 | 11,676 | 0.42 | 71.52% | 195,810 | 55,690 | 0.42 |
| | Tree-Based | Decision Tree | 31.07% | 14,913 | 8,830 | 0.43 | 60.06% | 169,665 | 41,666 | 0.65 |
| | | **AdaBoost** | **20.70%** | **9,846** | **5,977** | **0.74** | 64.89% | 145,554 | 41,585 | 0.73 |
| | | **GBM** | 23.75% | 11,022 | 6,632 | 0.68 | **54.66%** | **145,320** | **38,218** | **0.72** |
| | | RF | 28.24% | 14,130 | 8,507 | 0.49 | 62.35% | 174,510 | 42,182 | 0.58 |
| | | ERT | 26.14% | 12,006 | 7,991 | 0.66 | 59.15% | 141,927 | 37,492 | 0.74 |
| | | Bagging | 32.32% | 14,978 | 10,296 | 0.50 | N/A | N/A | N/A | N/A |
| | | Stacking | 26.89% | 12,090 | 7,466 | 0.66 | N/A | N/A | N/A | N/A |
| | | XGBoost | 23.69% | 10,637 | 6,764 | 0.72 | 61.48% | 147,729 | 39,658 | 0.74 |
| | | LightGBM | 25.57% | 11,808 | 7,933 | 0.66 | 63.16% | 137,118 | 41,086 | 0.79 |

Table 7. Non-Segmented Approach – Summary of Results (Best Model Marked in Green for each Data Setup and Area)

Feature importances of the best-performing non-segmented models for Cavite and Metro Manila are shown in Tables 8 and 9. Property specification characteristics feature dominantly in both tables; ‘Floor Area’, ‘# of Bedrooms’, and ‘# of Car Spaces’ do show up on both tables. Different location attributes sourced from OpenStreetMap also populate the most important feature list. Interestingly despite utilizing data setups which contain alternative data, features from these alternative sources were not considered as highly influential features. Only ‘LGU Cost of Doing Business’ from CMCI is found in Table 8. The lack of CMCI and Socio-Economic indicators found in these tables may indicate that their granularity is not enough to greatly impact the model’s overall decision. An intricacy further than an LGU level is likely needed to truly separate different houses. Nonetheless, the presence of such alternative data does improve model performance slightly, which in theory may save hundred-thousands to millions of Philippine pesos in incorrect valuations of properties.

| Feature | Category | Feature Importance |
|---|---|---|
| Floor Area | Property Specification | 0.24623 |
| LGU Cost of Doing Business | CMCI | 0.16773 |
| # of basketball courts | Property Specification | 0.10473 |
| # of government buildings w/in 5 kms | Location | 0.03438 |
| # of bedrooms | Property Specification | 0.03091 |
| # of pools | Property Specification | 0.02920 |
| # of Schools w/in 1km | Location | 0.02462 |
| Having Postcode of 1231 | Location | 0.02459 |
| # of Car Spaces | Property Specification | 0.02402 |
| # of Grass patches | Property Specification | 0.02289 |

Table 8. Top 10 Important Features for Best-Performing Model – Metro Manila (GBM – Combination)

| Feature | Category | Feature Importance |
|---|---|---|
| Floor Area | Property Specification | 0.30592 |
| # of bathrooms | Property Specification | 0.03991 |
| # of residential buildings w/in 5kms | Location | 0.03690 |
| # of bedrooms | Property Specification | 0.03494 |
| # of pubs w/in 3kms | Location | 0.03275 |
| # of schools w/in 5kms | Property Specification | 0.02407 |
| # of car spaces | Property Specification | 0.02268 |
| # of fences | Location | 0.02211 |
| # of fuel stations w/in 1kms | Location | 0.02206 |
| presence of local airport | Location | 0.01725 |

Table 9. Top 10 Important Features for Best-Performing Model – Cavite (AdaBoost – LGU Competitiveness)

### 3.2. Segmented Approach

Segmentation generally provides substantial reduction in minimizing the Mean AE and MAPE. Comparing the best-performing non-segmented models in Table 10 to the best-performing segmented models per cluster in Table 11, this is particularly noticeable in the data setups wherein only the city competitiveness index and only the socio-economic variables were used. Highlighted in blue in Table 11 are clusters which performed better than their best non-segmented model counterpart of the same data setup and highlighted in green are metrics which on average beat the same metric of the best non-segmented model counterpart of the same data setup.

In Cavite, 3 out of 4 clusters beat the MAPE of their non-segmented data setup counterpart, namely in Baseline, LGU Competitiveness, and Socio-Economic experiments. The Combination experiment only had 1 out of 4 clusters beat their counterpart's MAPE of 20.70%. AdaBoost models were still dominant as best-performing models in Cavite but Stacking and XGBoost regressors were also notable inclusions. In Metro Manila, 3 out of 4 clusters beat the MAPE for Baseline model, while 2 out of 4 clusters did so for LGU Competitiveness, Socio-Economic, and Combination data setups. Performances in some clusters were significantly better, reaching 32-48% MAPEs. Interestingly, ERT, Stacking, and GBM regressors did not appear as commonly in Table 11, as compared to the ones found in Metro Manila for Table 10.

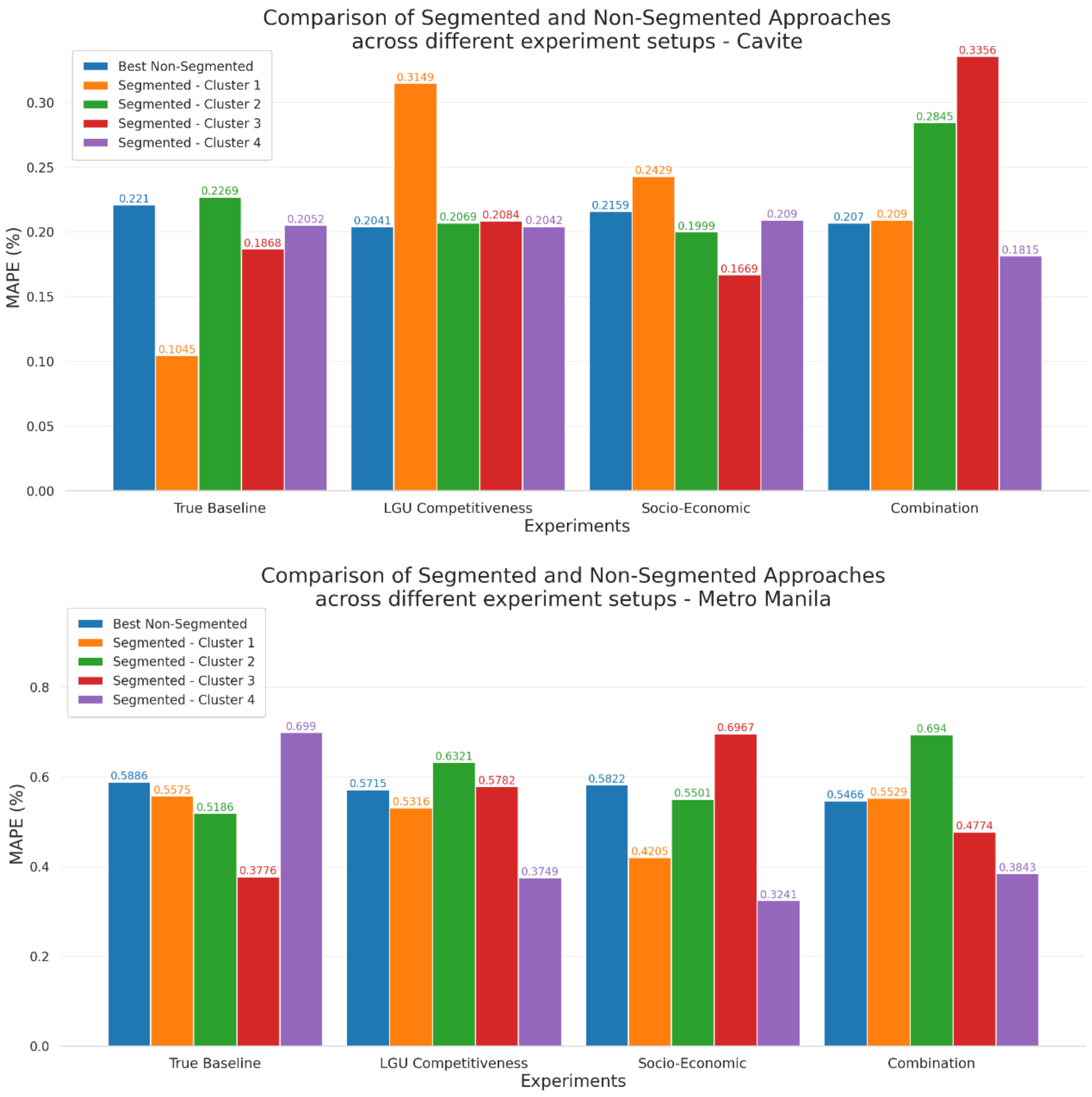
| Data Setup | Cavite | | | | | Metro Manila | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Best Model | MAPE (%) | Mean AE | Med AE | $R^2$ Score | Best Model | MAPE (%) | Mean AE | Med AE | $R^2$ Score |
| **Baseline** | AdaBoost | 22.10% | 10,774 | 5,000 | 0.58 | ERT | 58.86% | 144,697 | 41,266 | 0.76 |
| **LGU Comp.** | AdaBoost | 20.41% | 9,630 | 5,380 | 0.68 | Stacking | 57.15% | 140,932 | 41,471 | 0.74 |
| **Socio-Econ.** | AdaBoost | 21.59% | 10,241 | 5,633 | 0.70 | GBM | 58.22% | 149,307 | 37,692 | 0.71 |
| **Combi.** | AdaBoost | 20.70% | 9,846 | 5,977 | 0.74 | GBM | 54.66% | 145,320 | 38,218 | 0.72 |

Table 10. Best-Performing Non-Segmented Models for all Data Setups

| Segmented | | Cavite | | | | | Metro Manila | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Setup | Cluster # | Best Model | MAPE (%) | Mean AE | Median AE | R2 Score | Best Model | MAPE (%) | Mean AE | Median AE | R2 Score |
| Baseline | 1 | Stacking | 10.45% | 5,140 | 3,057 | 0.85 | ERT | 55.75% | 203,501 | 56,719 | 0.75 |
| | 2 | AdaBoost | 22.69% | 11,083 | 4,737 | 0.70 | AdaBoost | 51.86% | 77,070 | 32,311 | 0.61 |
| | 3 | AdaBoost | 18.68% | 8,784 | 5,086 | 0.57 | AdaBoost | 37.76% | 77,070 | 16,412 | 0.78 |
| | 4 | ERT | 20.52% | 11,129 | 6,694 | 0.84 | AdaBoost | 69.90% | 116,232 | 36,626 | 0.44 |
| | Ave. | -- | 18.09% | 9,034 | 4,893 | -- | -- | 53.82% | 118,468 | 35,267 | -- |
| LGU Comp. | 1 | GBM | 31.49% | 12,101 | 6,843 | 0.39 | AdaBoost | 53.16% | 95,408 | 31,146 | 0.59 |
| | 2 | AdaBoost | 20.69% | 9,365 | 4,175 | 0.69 | AdaBoost | 63.21% | 125,525 | 42,140 | 0.54 |
| | 3 | AdaBoost | 20.84% | 8,606 | 5,696 | 0.77 | RF | 57.82% | 208,530 | 49,661 | 0.84 |
| | 4 | XGBoost | 20.42% | 14,684 | 6,906 | 0.43 | RF | 37.49% | 81,209 | 15,914 | 0.76 |
| | Ave. | -- | 23.36% | 11,189 | 5,905 | -- | -- | 52.92% | 127,918 | 34,715 | -- |
| Socio-Economic | 1 | GBM | 21.94% | 10,846 | 4,338 | 0.72 | AdaBoost | 42.05% | 109,740 | 29,444 | 0.44 |
| | 2 | AdaBoost | 24.29% | 16,686 | 8,999 | 0.36 | AdaBoost | 55.01% | 100,278 | 41,065 | 0.54 |
| | 3 | AdaBoost | 19.99% | 7,013 | 4,887 | 0.56 | LGBM | 69.67% | 266,615 | 83,741 | 0.81 |
| | 4 | AdaBoost | 16.69% | 9,288 | 5,363 | 0.67 | AdaBoost | 32.41% | 86,025 | 15,622 | 0.75 |
| | Ave. | -- | 20.73% | 10,958 | 5,897 | -- | -- | 49.79% | 140,665 | 42,468 | -- |
| Combi. | 1 | AdaBoost | 20.90% | 8,680 | 4,834 | 0.85 | AdaBoost | 55.29% | 89,924 | 31,868 | 0.60 |
| | 2 | XGBoost | 28.45% | 18,717 | 7,244 | 0.24 | ERT | 69.70% | 126,574 | 52,290 | 0.64 |
| | 3 | LGBM | 33.56% | 12,368 | 7,381 | 0.35 | ERT | 47.74% | 191,309 | 46,016 | 0.87 |
| | 4 | AdaBoost | 18.15% | 8,996 | 5,045 | 0.66 | Stacking | 38.43% | 79,230 | 15,288 | 0.76 |
| | Ave. | -- | 25.27% | 12,190 | 6,126 | -- | -- | 52.79% | 121,759 | 36,366 | -- |

Table 11. Segmented Approach – Summary of Results (Highlighted in blue are clusters which performed better than their best non-segmented model counterpart of the same data setup, while those highlighted in green are metrics which on average beat the same metric of the best non-segmented model counterpart of the same data setup)

Notably, no experiment in both locations had all clusters beat the MAPE of its best non-segmented model. For individual cluster performances, only Cavite's Cluster 4 and Metro Manila's Clusters 1 and 4 performed better at least thrice as compared to their non-segmented counterparts. While individual clusters may beat their best non-segmented data setup counterparts, combining all four clusters' performances as a whole may not always cast improvements, as seen also in Table 11 and in Figures 22a and 22b. Cavite's best non-segmented models were only beaten on average in the Baseline (18.09%) and Socio-Economic (20.47%) experiments. Metro Manila's best non-segmented models performed considerably better across all experiments, with average MAPEs of 53.82%, 52.92%, 49.79%, and 52.71% for Baseline, LGU Competitiveness, Socio-Economic, and Combination experiments respectively. On average, Mean AE and Median AEs did not perform better for Cavite; Metro Manila fared better with all but the Median AE in the Socio-Economic data setup not performing better.



Figures 22a and 22b. Comparison of Best Performing Segmented and Non-Segmented Approaches across different experiment setups for (a) Cavite and (b) Metro Manila

# 4. CONCLUSION

## 4.1. Summary

The main objective of this paper is to evaluate the effectiveness of utilizing commonly used ML techniques for the valuation of properties in the Philippines. With this, the paper also sought out to verify the effectiveness of using alternative data not commonly used by similar publications nor Philippine appraisers, to account for the lack of some conventional indicators readily available in the Philippine context.

The study considered two experimental set-ups consisting of a variety of ML models and combinations of data sources as inputs – a non-segmented approach considering all house listings during modelling, and a segmented approach where data points were clustered according to their structural attributes. Individual models were tested for each cluster for the latter set-up. Linear and tree-based methods were compared in finding the best models for each setup. The data sources included geolocation data from OpenStreetMap, rankings from Department of Trade and Industry's Cities and Municipalities Competitiveness Index, and other socio-economic indicators obtained from the Philippine Statistics Authority (PSA), BIR, and other government resources. House property listings from Cavite and Metro Manila, were scraped from a popular Philippine real estate listing site and were used in separate models.

For the non-segmented approach, a data setup composed of Lamudi, OSM, and CMCI features performed the best for the Cavite area with a 20.41% MAPE, Php 9,630/sqm Mean AE, Php 5,380/sqm Median AE, and 68% $R^2$ score. The Metro Manila models performed considerably worse – the best setup consisted of Lamudi, OSM, CMCI, and Socio-Economic features with a 54.66% MAPE, Php 149,307/sqm Mean AE, Php 38,218/sqm Median AE, and 72% $R^2$ score. A lack of finer granularity for government-based indicators in Metro Manila is assumed to be the main cause of poor performance, as the area's highly urbanized setting has highly varying contexts within LGUs, i.e., in barangays and populated areas. Despite having alternative data improve modelling performance, property specification and location features found in Lamudi and OSM still dominated feature importances for both locations. Performing property segmentation via K-Means and BIRCH Clustering slightly improved model performances for all data setups in Metro Manila, showing as much as 5-15% increases in MAPE for some clusters, and partially for the Cavite datasets, only improving in the Baseline and Socio-Economic experimental setups.

Commonly used ML techniques found in similar property prediction publications were found to be partially effective under a Philippine context, as Cavite's were competitive against Kuala Lumpur ML models with 11.3-20.9% MAPE and beat Hong Kong ML models of 32-54% MAPE. The Metro Manila models' relative failures can be attributed to a lack of granularity in the data, which suggests that both Cavite and Metro Manila's models can improve even more substantially. LGU granularity-based socio-economic indicators and other government-based features were found to be less predictive than the traditional property data and alternative OSM-based location data, but still improved model performances.

## 4.2. Recommendations

This paper hopes to spark discussion and further research on more objective and transparent approaches regarding Philippine property valuation, as well as push for updated and readily available data for appraisers, home buyers, and home sellers to utilize during their property valuation process in the country. In order to achieve performances that can match the standards needed for operationalization, the following are recommended for further study: comparison with spatial econometric models; finer granularity of government-based indicators, such as those in the barangay level; the usage of mapping initiatives with more documentation of amenities and buildings, such as a Google Maps API; utilizing scores instead of ranks for CMCI-related indicators; comparison and possible integration with procedures found in automated valuation machines (AVMs); further hyperparameter tuning and usage of deep learning techniques; usage of satellite imagery for capturing in detail the granular features of a location; an increase in and variety of data points found in the specified locations; and obtaining a prediction interval of forecasted prices along with the base predicted target variable.

## 5. REFERENCES

Abellana, J. A., & Devaraj, M. (2021). Hedonic Modeling for Predicting House Prices during CoVid19 Pandemic in the Philippines. *3rd International Conference on Management Science and Industrial Engineering* (pp. 21-26). Association for Computing Machinery.

Adetiloye, K., & Eke, P. (2014). A Review of Real Estate Valuation and Optimal Pricing Techniques. *Asian Economic and Financial Review*, 1878-1893.

Agosto, A. B. (2017). Determinants of Land Values in Cebu City, Philippines. *International Conference on Business and Economy.*

Ahlfeldt, G. M. (2013). If We Build it, Will They Pay? Predicting Property Price Effects of Transport Innovations. *Environment and Planning A: Economy and Space, 45(8)*, 1977-1994.

Angrick, S., Bals, B., Niko, H., Kleissl, M., Schmidt, J., Vanja, D., . . . Friedrich, T. (2022). Towards Explainable Real Estate Valuation via Evolutionary Algorithms. *Genetic and Evolutionary Computation Conference* (pp. 1130-1138). Association of Computing Machinery.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39(2)*, 1772-1778.

Azimlu, F., Rahnamayan, S., & Makrehchi, M. (2021). House price prediction using clustering and genetic programming along with conducting a comparative study. *Genetic and Evolutionary Computation Conference Companion*, (pp. 1809-1816).

Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernardez, O., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences, 8(11)*.

Beimer, J., & Francke, M. (2019). Out-of-Sample House Price Prediction by Hedonic Price Models and Machine Learning Algorithms. *Real Estate Research Quarterly, 18(2)*, 13-20.

Board of Valuers, Appraisers, Estate Agents & Property Managers. (2019). *Malaysian Valuation Standards, 6th ed.*

Breiman, L. (1996). Stacked Regressions. *Machine Learning, 24*, 49-64.

Bureau of Local Government Finance. (2018). *Philippine Valuation Standards.*

Buyukkaracigan, N. (2021). *Modern Methods Approach in Real Estate Valuation.* Iksad Publications.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3(1)*, 1-27.

Camella. (2022, March 2). *The Best Real Estate Websites In The Philippines.* Retrieved from Camella PH: https://www.camella.com.ph/real-estate-websites-philippines/

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*.

Cellmer, R., Cichulska, A., & Belej, M. (2020). Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *International Journal of Geo-Information*.

*Census of Population and Housing*. (n.d.). Retrieved from Philippine Statistics Authority: https://psa.gov.ph/population-and-housing

Chaphalkar, N., & Sandbhor, S. (2013). Use of Artificial Intelligence in Real Property Valuation. *International Journal of Engineering and Technology (IJET)*, 2334-2337.

Chatfield, C. (1978). The Holt-Winters Forecasting Procedure. *Applied Statistics, 27(3)*, 264-279.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters, 19(11)*, 989-996.

Chen, S., Zhuang, D., & Zhang, H. (2020). GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China. *Complexity*.

Chen, T., & Carlos, G. (2016). XGBoost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). New York: Association of Computing Machinery.

Chou, J.-S., Fleshman, D.-B., & Truong, D.-N. (2022). Comparison of machine learning models to provide preliminary forecasts of real estate prices. *Journal of Housing and the Built Environment*.

Commoner. (2020, June 24). *The Divide in Our Cities.* Retrieved from Medium: https://mediacommoner.medium.com/the-divide-in-our-cities-bff743e1584

Congress of the Philippines. (2017). *Republic Act No. 10963.*

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing, 192*, 38-48.

Department of Trade and Industry. (n.d.). Retrieved from Cities & Municipalities Competitiveness Index: https://cmci.dti.gov.ph/

Dickinson, C. (2021, February 19). *Inside the 'Wikipedia of Maps,' Tensions Grow Over Corporate Influence.* Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2021-02-19/openstreetmap-charts-a-controversial-new-direction

Domingo, E. V., & Fulleros, R. F. (2002). *Real estate price index: a model for the Philippines.* Bank of International Settlements.

Dunn, K. (2021, April 6). *Avoid R-squared to judge regression model performance.* Retrieved from Towards Data Science: https://towardsdatascience.com/avoid-r-squared-to-judge-regression-model-performance-5c2bc53c8e2e

Evans, K., Lausberg, C., & Sui Sang How, J. (2019). Reducing Property Appraisal Bias with Decision Support Systems: An Experimental Investigation in the South African Property Market. *Journal of African Real Estate Research, 4(1)*, 108-138.

Freidman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29(5)*, 1189-1232.

Freund, Y., & Robert, S. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory* (pp. 23-37). Berlin: Springer.

Gao, G., Bao, Z., Cao, J., Oin, A., & Sellis, T. (2022). Location-Centered House Price Prediction: A Multi-Task Learning Approach. *ACM Transactions on Intelligent Systems and Technology, 13(2)*, 1-25.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3-42.

Hau, K.-C. (2020). House Prices in the Peripheries of Mass Rapid Transit Stations Using the Contingent Valuation Method. *Sustainability, 12(20)*.

Ho, T. (1995). Random Decision Forests. *3rd International Conference on Document Analaysis and Recognition*, (pp. 278-282). Montreal.

Ho, W. K., Tang, B.-S., & Wong, S. (2020). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 1-23.

Howard, C. (2004). *Is There Assessor Bias in the Real Estate Market?* Illinois Wesleyan University.

Huang, Z., Chen, R., Xu, D., & Zhou, W. (2017). Spatial and hedonic analysis of housing prices in Shanghai. *Habitat International, 67*, 69-78.

Joy, C. (2021, May 1). *Explainable AI for Property Valuation.* Retrieved from Medium: https://medium.com/clear-capital-engineering/explainable-ai-for-property-valuation-cc110381106b

Ke, G., Meng, Q., Finley, T., Wang, T., Chan, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting. *International Conference on Neural Information Processing Systems* (pp. 3149-3157). New York: Association of Computing Machinery.

Kershaw, P., & Rossini, P. (1999). Using Neural Networks to Estimate Constant Quality. *Pacific-Rim Real Estate Society Conference.*

Krzystanek, M., Lasota, T., & Trawinski, B. (2009). Comparative Analysis of Evolutionary Fuzzy Models for Premises Valuation Using KEEL. *International Conference on Computational Collective Intelligence*, (pp. 838-849).

Lech, M., & Gerald, L. (2018). *Current issues of the Philippine land use planning and management system.* German Institute for Development Evaluation.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36(2)*, 451-461.

Lughofer, E., Trawinski, B., Trawinski, K., & Lasota, T. (2011). On-Line Valuation of Residential Premises with Evolving Fuzzy Models. *International Conference on Hybrid Artificial Intelligence Systems*, (pp. 107-115).

Mandani Bay. (2018). *An expert's manual for real estate valuation in the philippines*. Retrieved from https://www.mandanibay.com/blog/experts-manual-real-estate-valuation-philippines/

Masias, V., Crespo, F., Valle, M., & Crespo, R. (2016). Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile. In C. Berger, *Lectures on Modelling and Simulation* (pp. 98-105). AMSE.

McCluskey, W. J., Daud, D., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction, 19(2)*, 152-167.

McKenzie, J. (2011). Mean absolute percentage error and bias in economic forecasting. *Economics Letters, 113(3)*, 259-262.

Nallathiga, R., Upadhyay, A., Karmarkar, P., & Acharya, K. (2019). Tenure-Wise Determinants of Residential Property Value: An Application of Hedonic Pricing Model in Balewadi, Pune, India. *Theoretical and Empirical Researches in Urban Management, 14(4)*, 70-85.

Naqvi, S. (2017). *THE IMPACT OF MACROECONOMIC FACTORS ON THE REAL ESTATE PRICES IN USA.* North Carolina, Wilmington.

*National Quickstat for 2022*. (n.d.). Retrieved from Philippine Statistics Authority: https://psa.gov.ph/statistics/quickstat/national-quickstat/all/*

Nguyen, N., & Cripps, A. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research, American Real Estate Society, 22(3)*, 313-336.

OpenStreetMap. (n.d.). *Map Features*. Retrieved from OpenStreetMap: https://wiki.openstreetmap.org/wiki/Map_features

Pi-ying, L. (2011). Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City. *Journal of Modern Accounting and Auditing, 7(10)*, 1080-1089.

Primer. (2021, June 19). *8 Real Estate Websites and Apps You Can Rely On in the Philippines*. Retrieved from Primer PH: https://primer.com.ph/tips-guides/2021/06/19/list-6-real-estate-websites-and-apps-you-can-rely-on/

Remo, A. (2021, December 3 ). *The Future of PH Real Estate*. Retrieved from Inquirer: https://business.inquirer.net/335602/the-future-of-ph-real-estate

Rico-Juan, J., & de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Santos, L. J., & Jiang, R. (2020). *Spatial Analysis of House Price Determinants: A Greater London Case Study.* University College London.

Sario, M. S. (2019). A Spatial Econometric Model for Household Electricity Consumption in the Philippines. *14th National Convention on Statistics.*

Scikit-Learn. (n.d.). *Ensemble methods.* Retrieved from Scikit-Learn: https://scikit-learn.org/stable/modules/ensemble.html#bagging

Similarweb. (2022). *Top Real Estate Websites in Philippines Ranking Analysis for July 2022*. Retrieved from Similarweb: https://www.similarweb.com/top-websites/philippines/category/business-and-consumer-services/real-estate/

Sommervoll, D., & Sommervoll, A. (2018). *Learning from man or machine: Spatial aggregation and house price prediction.* Norwegian University of Life Sciences.

*Statistics*. (n.d.). Retrieved from Bureau of Local Government Finance: https://blgf.gov.ph/lgu-fiscal-data/

Tanrivermis, H. (2016). *Real Estate Valuation Principles.* Ankara.

Thanasi, M. (2016). Hedonic appraisal of apartments in Tirana. *International Journal of Housing Markets and Analysis, 9(2)*, 239-255.

The Danh Phan. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE*, 35-42.

Tidwell, O., & Gallimore, P. (2014). The influence of a decision support tool on real estate valuations. *Journal of Property Research, 31(1)*, 45-63.

Unciano, R. C. (2020, February 25). *Reforming the real-property valuation system in the Philippines.* Retrieved from Business Mirror Philippines: https://businessmirror.com.ph/2020/02/25/reforming-the-real-property-valuation-system-in-the-philippines/

van der Hoeven, D. (2022). *Appraiser-based automated valuation: A case study of valuing buy-to-let properties in the Netherlands.* University of Groningen.

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering.* IOP Publishing Ltd.

Wittowsky, D., Hoekveld, J., Welsch, J., & Steier, M. (2020). Residential housing prices: impact of housing characteristics, accessibility and neighbouring apartments – a case study of Dortmund, Germany. *Urban, Planning and Transport Research*, 44-70.

Xiao, L., & Yan, T. (2019). Prediction of House Price Based on RBF Neural Network Algorithms of Principal Component Analysis. *International Conference on Intelligent Informatics and Biomedical Sciences* (pp. 315-319). IEEE.

Xue, C., Ju , Y., Li , S., Zhou, Q., & Liu, Q. (2020). Research on Accurate House Price Analysis by Using GIS Technology and Transport Accessibility: A Case Study of Xi'an, China. *mdpi*.

Yiu, C., Tang, B., Chiang, Y., & Choy, L. T. (2006). Alternative Theories of Appraisal Bias. *Journal of Real Estate Literature, 14(3)*, 321-344.

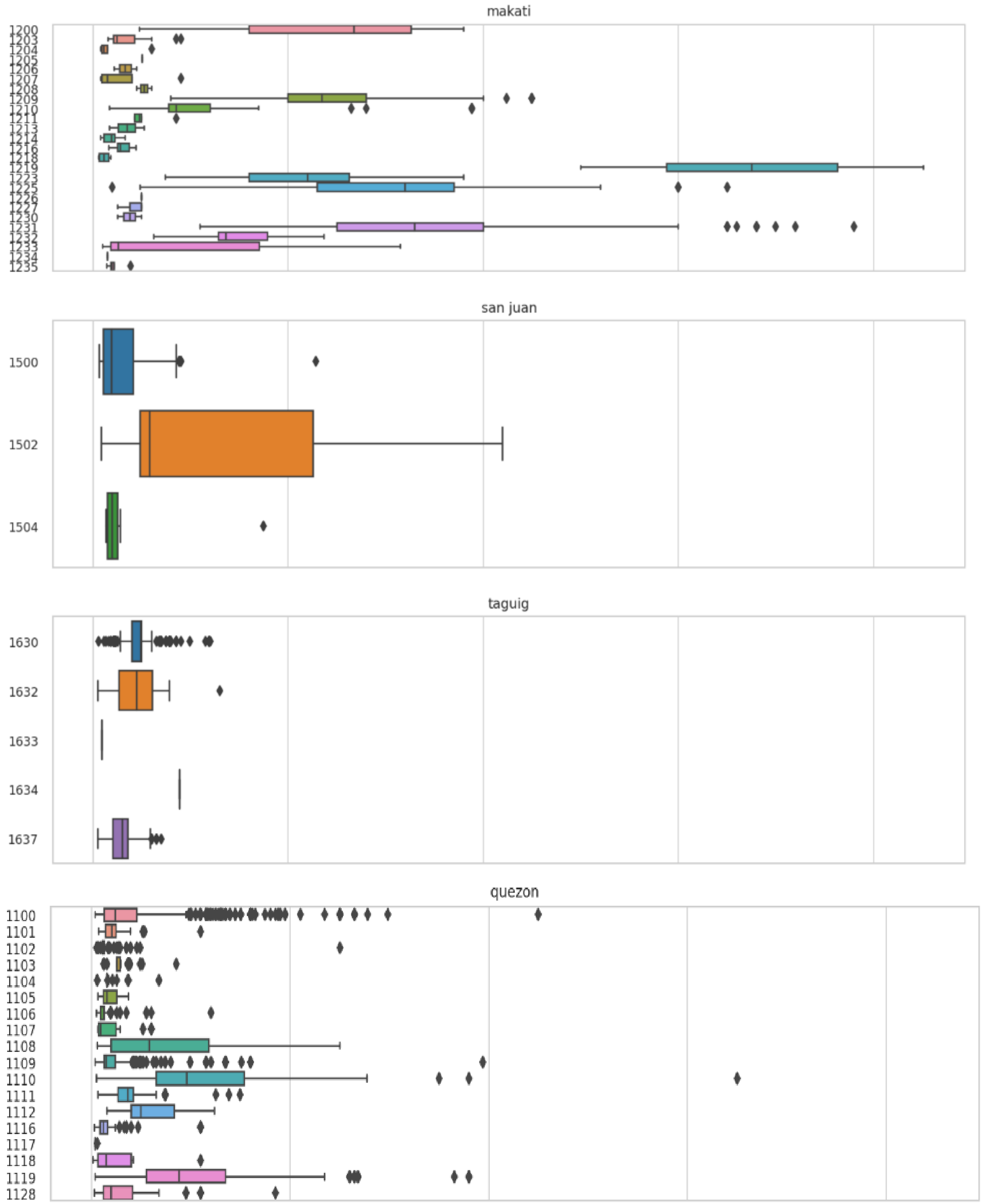
Zhang, L., Zhou, J., Hui, E., & Wen, H. (2018). The effects of a shopping mall on housing prices: A case study in Hangzhou. *International Journal of Strategic Property Management*, 65-80.

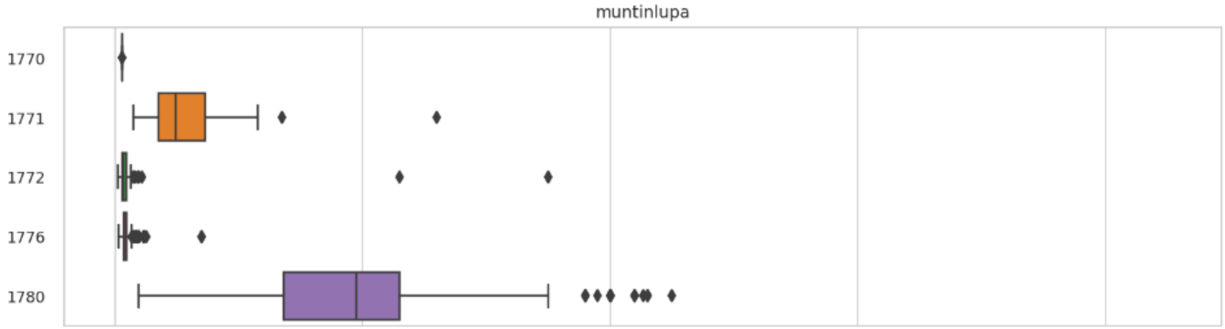
Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record, 25(2)*, 103-114.

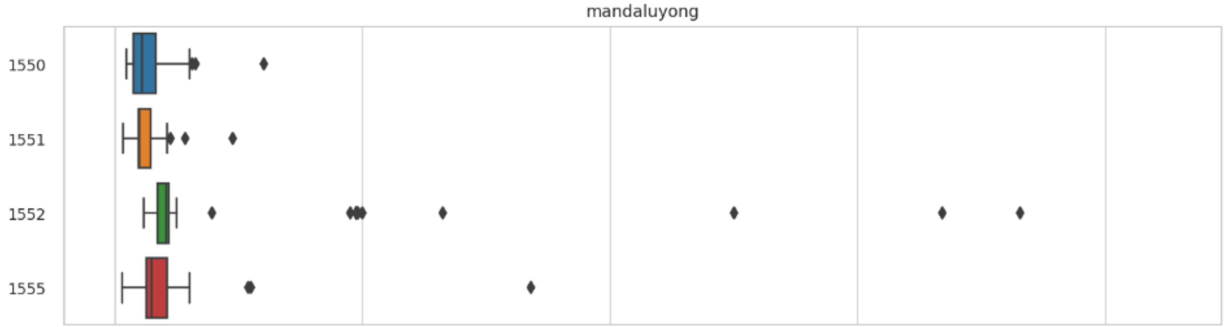
Zhao, Y., Chetty, G., & Tran, D. (2019). Deep Learning with XGBoost for Real Estate Appraisal. *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, (pp. 1396-1401).

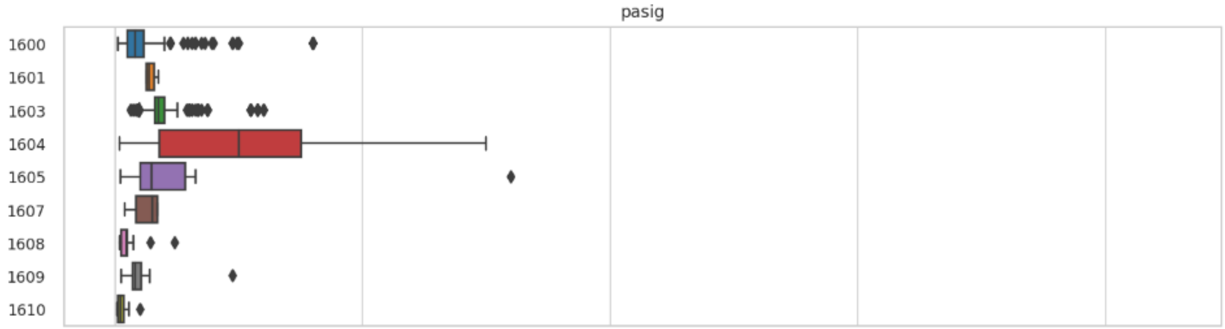
## 6. APPENDIX

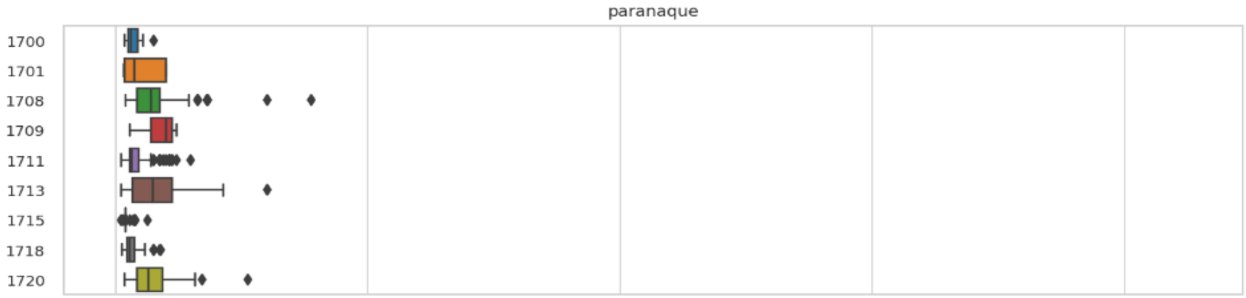
Appendix A. Boxplot Graphs of Lamudi-Scraped House Prices per Postcode per LGU in Metro Manila

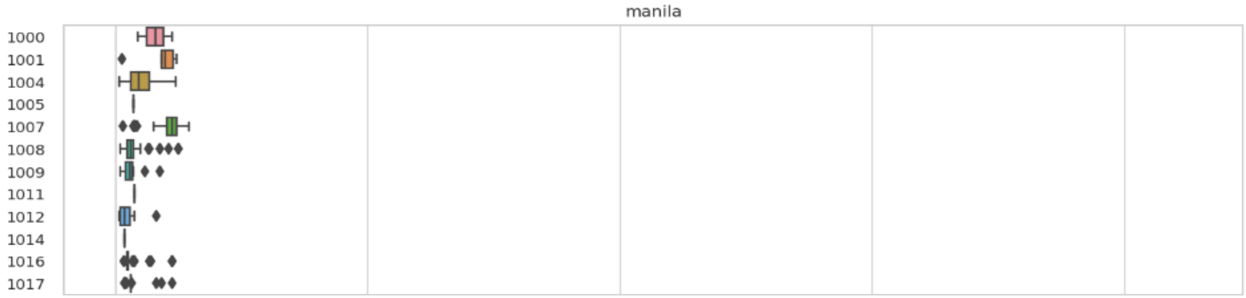


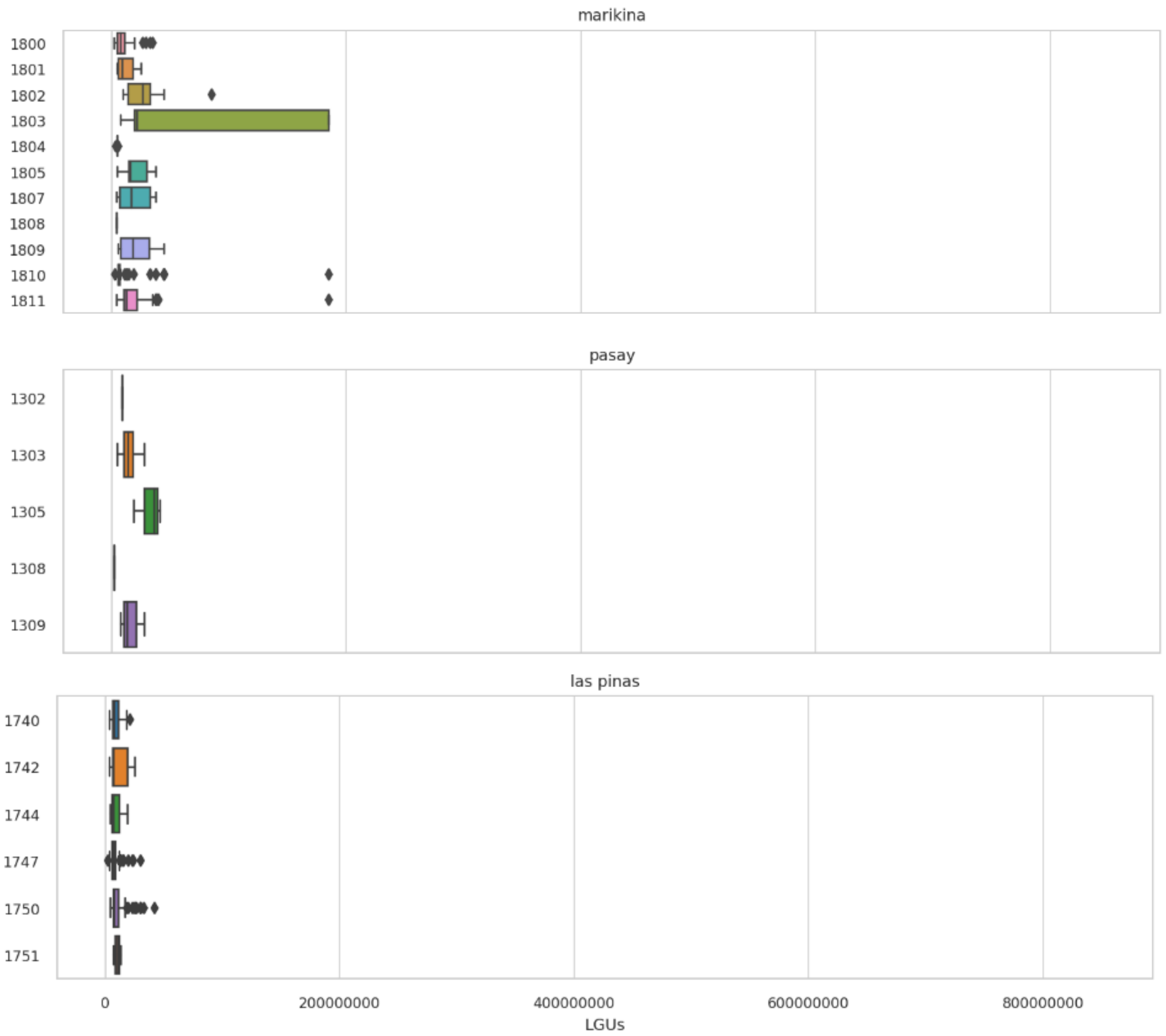

muntinlupa
1770
1771
1772
1776
1780


mandaluyong
1550
1551
1552
1555

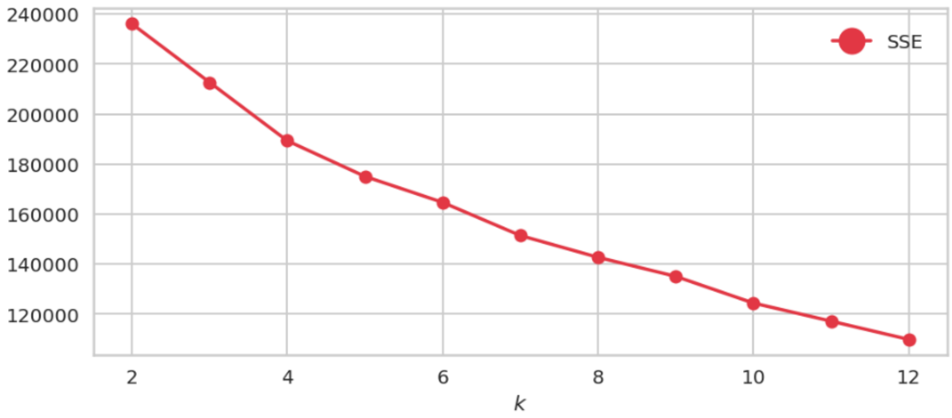

pasig
1600
1601
1603
1604
1605
1607
1608
1609
1610

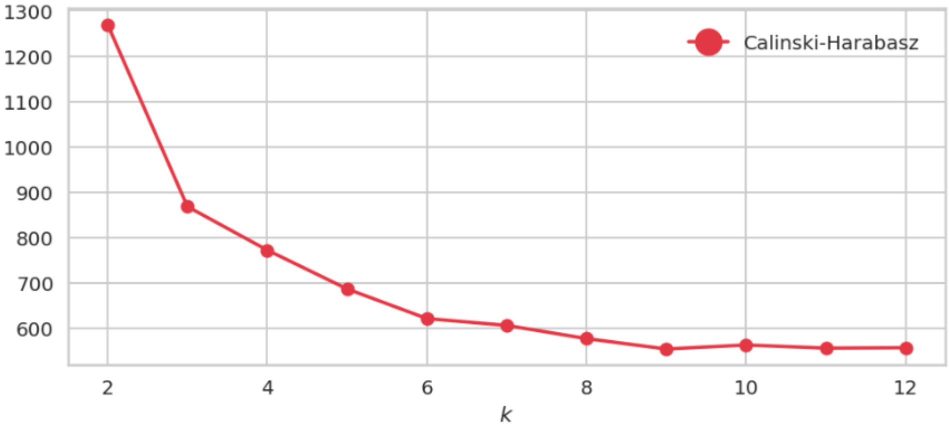

paranaque
1700
1701
1708
1709
1711
1713
1715
1718
1720

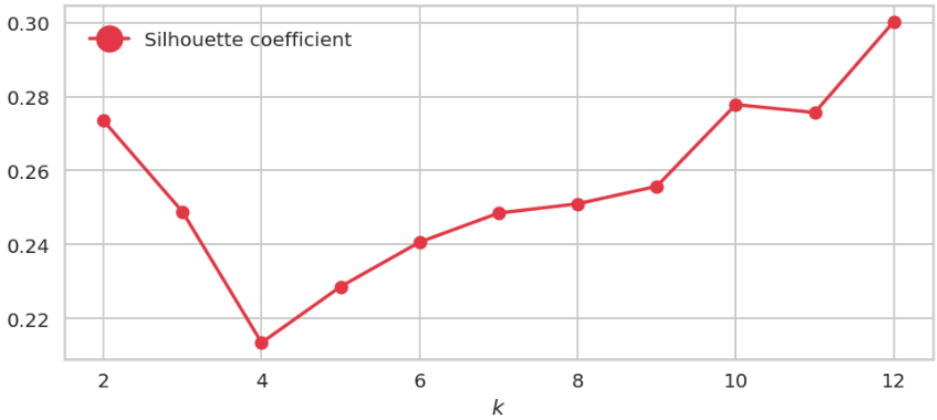

manila
1000
1001
1004
1005
1007
1008
1009
1011
1012
1014
1016
1017



Appendix B. Clustering performance metrics on a FAMD-employed on Cavite




Calinski-Harabasz
1300
1200
1100
1000
900
800
700
600
2
4
6
8
10
12
k


Silhouette coefficient
0.30
0.28
0.26
0.24
0.22
2
4
6
8
10
12
k