University of Warsaw
Faculty of Economic Sciences

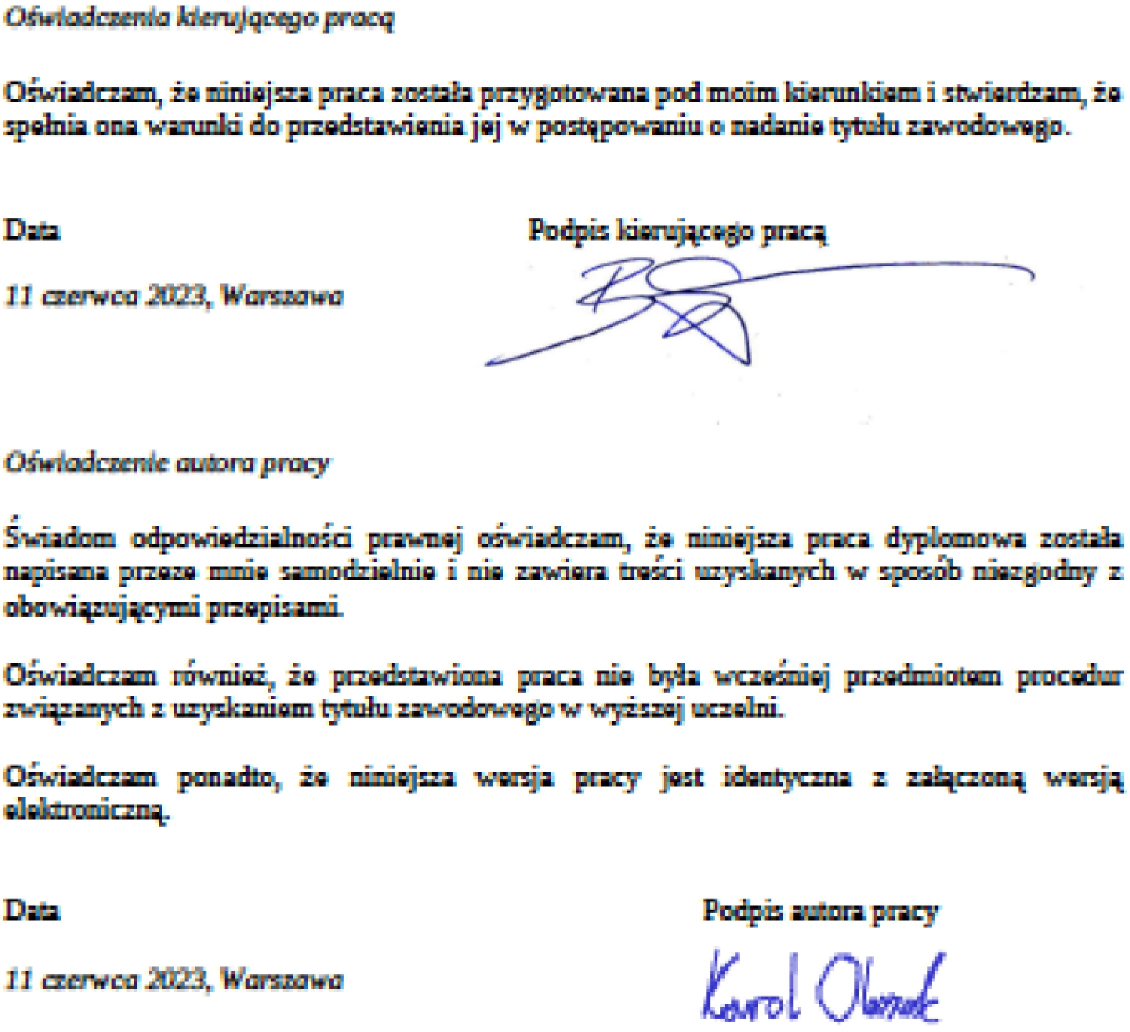
Karol Oleszek
Student no. 443069

Data Science Approach to Real Estate Pricing: An Empirical Study of Public Listing Data from 5 OECD Countries

Master's thesis
in ECONOMICS

Supervisor:
dr. Bartłomiej Dessoulavy-Śliwiński
Department of Management and Information Technology

Warsaw, June 2023



Oświadczenia kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data **Podpis kierującego pracą**

11 czerwca 2023, Warszawa

Oświadczenie autora pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data **Podpis autora pracy**

11 czerwca 2023, Warszawa

Karol [illegible]

Serdecznie dziękuję Panu Promotorowi,
dr Bartłomiejowi Dessoulavy-Śliwińskiemu,
za pomoc w pisaniu pracy magisterskiej,
poświęcony czas oraz cenne i celne uwagi.

Abstract

This thesis focuses on the gathering of real estate listings data from 5 OECD countries, construction of a housing pricing model, and its detailed diagnostics. The real estate market plays an important role in the broader economy, and a better understanding of its dynamics can provide valuable insights into wealth distribution, consumer spending, and financial stability. By analyzing the spatial dynamics of real estate prices, this thesis contributes to the field of economics and informs evidence-based pricing models construction. The dataset collected in this study will be made available for scientific purposes, facilitating future research and collaborations in the field of real estate economics. Hypothesis relating to the location based property pricing models performance are evaluated. Overall, this thesis demonstrates the potential for research to yield significant insights into the functioning of real estate markets.

Keywords

real estate, listings, spatial model, OECD countries, economics, dataset, property pricing model

rynek nieruchomości, ogłoszenia, model przestrzenny, kraje OECD, ekonomia, zbiór danych, model wyceny nieruchomości

Thesis domain (Socrates-Erasmus subject area codes)

14.3 Economics

Subject classification

Tytuł pracy w języku polskim

Wycena Nieruchomości z Wykorzystaniem Metod Data Science: Badanie Empiryczne oparte o publiczne ogłoszenia z 5 krajów OECD

Contents

Introduction

Historically, the real estate sector has been perceived as a traditional industry, often resistant to the adoption of new technologies. Its operations and practices have been slow to change, largely maintaining the conventional methods of functioning.

Simultaneously, this sector has played a pivotal role in generating massive wealth and creating jobs, spanning both affluent and developing economies. This industry, by virtue of its operations, contributes significantly to economic development and employment opportunities.

Given the sector's importance, it has drawn the attention of researchers from a multitude of fields such as economics, geography, urban studies, and computer science. These experts have been intrigued by the potential of the real estate market and have sought to explore its intricacies and potential opportunities for growth and innovation.

In open market economies, real estate assets are frequently bought and sold, representing some of the most significant monetary transactions. These exchanges, excluding those on the financial instruments markets, are among the highest-priced transactions recorded, further underscoring the sector's economic impact.

This has led to an exponential increase in the importance of accurate real estate market price information. The necessity of precise valuation has resulted in a proliferation of pricing models and methodologies aimed at more effectively determining property values.

This study makes a significant contribution to the field of real estate in two major ways. Firstly, it has amassed an international real estate listings dataset from 5 OECD countries, broadening the scope and reach of real estate data available for analysis.

Secondly, a spatial pricing model has been developed and evaluated, providing valuable insights into the crucial role spatial information plays in the practice of housing valuation. This model underscores the impact of location and surrounding features on property values, adding a new dimension to the field of real estate pricing.

The structure of the study is laid out in three chapters for easy comprehension and systematic presentation of the findings. The first chapter provides a comprehensive overview of property pricing models and explores recent advancements in alternative signals to these models, setting the context for the research.

The second chapter delves into the specifics of the dataset gathered for the study, detailing its empirical characteristics. It outlines the scope, diversity, and unique elements of the dataset, setting the foundation for the analysis presented in the next chapter.

The final chapter of the study presents the spatial housing pricing model and demonstrates its application to international markets, specifically in Colombia, Chile, Mexico, the Netherlands, and Poland. The chapter evaluates the following hypotheses: 1) Housing pricing models may achieve high prediction accuracy without employing only property size and location information 2) inclusion of information about counts of schools, restaurants, parks, universities and transport hubs in the property neighborhood leads to increased housing model performance. It consolidates the study's findings, offering a coherent summary and demonstrating the model's practical implications and efficacy in real-world scenarios.

Chapter 1

Background

1.1. AVM and mass appraisal methods

1.1.1. Property pricing

The economy of property pricing aims to explain the price-generating process which leads real estate transactions to happen at an agreed price (Mooya, 2017). In open markets property prices are influenced by the interaction of supply and demand factors. On the demand side economic theories assert that marginal utility derived from property ownership is a driving force behind prices. On the supply side the costs are meant to be a dominant driving force for price level setting process.

The literature traditionally assumes that homebuyers are driven by hedonic motivations. The assumption implies that houses with larger quantities of desirable features (and lower quantities of undesirable features respectively) can be sold for higher prices. That is because homebuyers are willing to spend additional money for extra utility (and smaller disutility) they are receiving with *hedonic property features* present in greater quantities. Hedonic features are usually tied to the home itself or its location. Most common home related features include property size, number of rooms and bathrooms, property condition, additional improvements e.g. swimming pool, big terrace, garden, garage. Location related features might include close access to transport hubs, schools and commercial centers, beautiful views, green and liveable neighborhood and other.

Supply side of the housing market plays a crucial role in long term price levels in the presence of robust demand. Supply is influenced by a range of factors:

- **Land scarcity** - scarce land may limit the number of houses that can be build. This effect is profoundly pronounced in big cities where more people are competing for land plots. Scarce land may lead to structural changes in the of houses market in the area - scarce land leads to smaller gardens, higher buildings, less detached housing and in some cases smaller houses surface.

- **Labor and building materials costs and availability** - key price components of a new housing are labor and material costs. Materials and labour availability crunches may disrupt construction sector developments and in effect constrain housing supply.

- **Legal constraints** - housing supply may be artificially constrained with restrictive legislation, lengthy building permit processes and local zoning laws. Local restrictive bills may be popular among members of local communities thanks to their effects of elevated houses price levels.

- **Risk and developers profit** - large scale housing construction projects may span multiple years over which massive pools of capital are frozen. Large projects face a plethora of risks including market risk (housing market, labor and building materials market), credit risk and regulatory risk. To attract investments in such projects and cushion potential headwinds the prices have to be elevated above the sum of land, labor and material costs.

In addition to these factors, market sentiment and expectations also play a role in determining the price of residential housing. In the heated market with quick property turnover buyers might be tempted to transact faster and agree on higher prices. Expectations around market conditions including changes to related markets e.g. land market, credit market, houses rental market might alter the views of market participants.

In the real estate analytics space three approaches towards property valuation are most popular (Dornfest et al., 2002):

- **Cost-based pricing or replacement cost valuation**: the approach relies on the fact that in an open market an individual or company may build the house/property on their own. The price of property is then a sum of land cost, labor and building materials cost, legal and organizational costs related the construction process (Eilers and Kunert, 2017).
- **(Discounted) Income approach**: Popular especially for commercial/office/residential for rent property pricing, the approach assess the economic value of property to be a sum of income over the certain period of time (Glumac and Rosiers, 2018)). Alternatively in the discounted income approach the value is equal to an infinite sum of discounted income stream. The approach depends on the existence of an active rental market operating in parallel with property sales market.
- **Comparative sales approach**: The approach relies on market transactions to derive the assessed property value. The assumption is that properties with comparable use, in similar condition, located nearby and with other features relatively similar can be sold for similar prices in the same time frame (Ciuna et al., 2017).

Hedonic models are constructed with a use of *hedonic features*. A derived formula can be used to predict transaction prices for properties based on the quantity and quality of its amenities.

Owing the high availability of property data, AVMs and mass appraisal models are regression tasks which attracted broad interest of researchers from fields of machine learning, statistics and computer science. Methods which have been applied with success to these problems range from neural network to evolutionary algorithms (Angrick et al., 2022).

1.1.2. Real estate listings

Real estate analytics tasks and in particular construction of property valuation models require reliable data sources. These include transactions databases, banks credit databases, land registers and other. Completeness, data delays and included information varies in these sources across countries. When market conditions are changing dynamically and valuation needs to be up-to-date, then application of alternative data sources can be undertaken.

Real estate online listings provide additional information about property markets. One of the key features of real estate listings data is the availability of a large number of variables that can be used as inputs for pricing models. This includes information on property characteristics such as size, age, and number of rooms, as well as location-specific variables such as neighborhood amenities and proximity to public transportation. Online listings have become extremely popular in real estate technology space as it provides timely source of relevant market information (Conway, 2018).

In addition, real estate listings data can provide up-to-date information on recent asking prices, allowing pricing models to adjust their estimates based on current market conditions. This can be particularly important in rapidly changing markets, where traditional methods of appraisal may not capture the most current trends.

However, there are also potential challenges associated with using real estate listings data for pricing models construction. One issue is the potential for bias in the data, as certain types of properties may be overrepresented or underrepresented in listings data. For example, luxury properties may be more likely to be listed than lower-priced properties, which could skew AVM estimates.

Another challenge is the quality of the data itself. Listings data may be incomplete or inaccurate, leading to errors in price estimates. It is important to carefully clean and preprocess the data to ensure that it is suitable for use in pricing models construction.

Despite these challenges, real estate listings data can be a valuable input to property pricing models and has been practically applied (Moosavi, 2017). By leveraging the wealth of information contained in listings data, pricing models can provide accurate and up-to-date estimates of property values, helping buyers and sellers make more informed decisions in the real estate market. A web scraping technique has been applied to gather a real estate online listings dataset for the purpose of this study. Details of the dataset are described in the next chapter.

1.2. Alternative signals in property valuation

The rising availability of data sources about property in the recent years has driven growth in incorporation of alternative signals into pricing models.

1.2.1. Floor plans

Floor plans are documents which precisely present property rooms dimensions and their relative position, traditionally on a 2-dimensional map. Along with the layout floor plans convey additional important information about property: windows location and property spatial orientation. These features might heavily impact the valuation of a property. Properties with uncommon and impractical layouts, low sunlight exposition or small number of windows may be hard to sell and therefore more likely to be valued at lower price.

Integration of floor plans into pricing models has been successfully applied leading to decrease pricing prediction errors (Solovev and Pröllochs, 2021). The method used deep convolutional neural networks (CNNs) to extract pricing sentiment from floor plans images and used the sentiment variables as inputs to pricing model.

1.2.2. Mobility data

Levels of human activity in certain areas might be an important factor in property valuation, especially for commercial properties. Recent developments (Coleman et al., 2022) in the big data space has enabled collection of large scale mobility data with use of Android and iOS smartphones. Dataset containing anonymized location data measured in 5 minute interval allow to distinguish between locations used for work and residential purposes and to extract measures of levels of activity in different areas. Extracted features have been used as inputs to property pricing model leading to decreased prediction errors.

1.2.3. Computer vision

Among different sources of property information, visual data sources distinguish themselves as sources capable of conveying most nuanced insights. Homebuyers, real estate agents and appraisers form their opinions about property after inspecting property visually. The emergence of online real estate listing websites has created an abundance of information in a form of property images. The quality of this data varies from photos taken with low-end equipment to professional photographic sessions. Virtual property walks have been growing in popularity with the emergence of required hardware technologies[1]. Analytics methods used to derive signals from property visual data have grown in popularity in the recent years helped with new developments in computer vision field. Main factors driving that growth were developments in convolutional neural networks methods and architectures, and greater availability of required compute resources.

Computer vision applications to property visual data processing are focused mainly on extracting property-level attributes (building amenities inclusion, architectural styles, property conditions) and extracting overall property sentiment. In the former case the extracted component is fed into further pricing models, therefore the component itself does not have to be interpretable. Owing to that computer vision methods may be employed to model more nuanced features of the property which otherwise would be difficult to incorporate into analytical process.

An attempt has been made to precisely predict property class and main use based on interior and exterior photos. Multimodal neural network architecture has been employed in that attempt (Stumpe et al., 2022).

Another promising application of computer vision in real estate is use of vision transformers, a self-supervised machine learning technique to extract embeddings from images and use them as inputs to pricing models. An attempt with use of DINO[2] model has been successfully undertaken (Yazdani and Raissi, 2023).

Main challenges which computer vision aided real estate analytics faces are related to data quality. Property imagery comes with all sorts of resolutions and quality. Required input normalization may remove or skew results by eliminating pieces of information from the data.

1.2.4. Spatial information

In the context of pricing models the local nature of real estate markets necessitates the inclusion of information from geographical neighborhood of appraised property. Effects of spatial autocorrelation are deeply embedded into the price generating process - home sellers set their prices primarily by observing the market in their immediate surroundings. A range of analytical processes have been employed to ensure proper inclusion of local-level information in the pricing models:

[1] Creating professional virtual walks require specialized equipment - multiple high resolution cameras paired with hardware responsible for collecting position information. In the process of virtual walk creation, a complete visual scan of all rooms is gathered and results form a digital twin of the property with precise floor plans and 3D model of interiors. One of the commercially employed scanning technologies is Matterport.

[2] Self-DIstillation with NO labels (DINO)

- **Neighboring amenities information:** Inclusion of variables describing counts of neighboring points of interest (restaurants, schools, etc.) may provide additional signals leading to higher model performance. This approach has been successfully applied to Turin property market prices prediction (Bergadano et al., 2019).

- **Spatial interpolation:** Spatial interpolation methods are employed to estimate property market conditions for areas lacking relevant observations and construct dense (non-sparse) price maps. Most common approaches are kriging and inverse distance weighting. Novel research applied geoadditive model based on penalized spline functions to Naples residential property rental market (Giudice and Paola, 2017).

- **Market microzones:** Local zoning set by laws may have profound influence on property market structure. City-level division into market microzones containing small areas ($1km^2 - 5km^2$) was applied to Turin housing market (Curto et al., 2017). Empirical results indicated that >50% of price variability has been explained with market microzones.

- **LVRS:** Location Value Response Function (LVRS) is an method which uses a continuous value surface as an input to the property appraisal process. In the study of Bari market (d'Amato, 2017b) the method application has resulted in better price prediction performance.

- **GWR:** Geographically Weighted Regression (GWR) is a model architecture which predicts values with a use of neighboring observations with weights assigned using a chosen kernel. In the investigation of Bloomington, Illinois single family housing market it has been shown that exponential kernel yielded most desirable results characteristics (Bidanset et al., 2017).

- **Spatial lag model:** The approach employs a spatial variable into the traditional hedonic modeling by creating a new explanatory variable, which conveys information about spatial effects. It has been shown that spatial lag models outperform traditional hedonic modeling in the Minsk housing market prices prediction (d'Amato et al., 2017).

1.2.5. Entity knowledge graphs

The real estate analytics field has long been associated with a great variety of data sources which are constantly being generated and collected. Although single source analytics already yields insights and value, it is the knowledge fusion which can unveil additional gains. Traditionally this has been achieved by joining data sources on property level keys into tabular form. The method however has some notable limitations when applied to datasets with multiple entities which span over large timeframes. Traditional models can not easily capture relations shifted in time and space. One tool which is being applied to overcome those limitations is entity knowledge graph (EKG).

An EKG is a type of knowledge graph that represents information about entities and the relationships between them. In the context of real estate, an EKG is used to model the relationships between properties, neighborhoods, cities, regions and past transactions. By representing these relationships in a graph structure, an EKG can provide a powerful foundation for further modeling.

EKG can be used to create a comparative sales pricing model. Linking only comparable properties with currently appraised property using EKG can yield more accurate price predictions. EKG-based approach[3] has been successfully applied to the task of appraising Taiwanese properties transactions (Li et al., 2022).

Property markets are prone to significant price shifts over time, which makes old pricing models obsolete over time. Periodical recalibration of pricing models is advisable to retain high prediction effectiveness (CoreLogic, 2011). For small markets with very few transactions information available this need can be difficult to achieve - the number observation in a selected time frame might be to small to achieve robust model performance. A EKG-based approach[4] has been applied to tackle a task of lifelong property valuation with use of sparse historical transaction data (Peng et al., 2020).

[3] Neighbor Relation Graph Learning Framework (ReGram).

[4] Lifelong house price prediction (LUCE)

Chapter 2

Dataset

The real estate sector is a pivotal component of the global economy, long recognized as a catalyst for economic growth, a substantial employer, and a generator of added value. As a constituent of the traditional sectors, it has not escaped the pervasive influence of digitization and the advent of the Internet. The emergence of online real estate listings portals has instigated a significant transformation within the industry by providing sellers and buyers with pertinent property and market information. Current listing websites are inundated with data pertaining to prices, property characteristics, and location. The ubiquity of Internet portals as a primary platform for real estate sales and inquiries has facilitated the introduction of a multitude of novel approaches to market insights and analysis. Consequently, trackers and indices reflecting market trends and dynamics have evolved to be more immediate and accurate.

The terms 'listing prices' and 'transaction prices' embody the seller's publicized intention to sell and the negotiated rate settled upon by all participating parties in the transaction, respectively. The relationship between these 'ask prices' and 'transaction prices' is characterized by several attributes: a significant degree of correlation, dynamism, and a dependency on prevailing market conditions. These features warrant deeper exploration in order to fully comprehend the complex dynamics that drive price setting and negotiation in markets.

In an idealized marketplace, the publicly declared intent to sell would precisely mirror the intrinsic value of the property market. However, it is commonplace in both developed and emerging property markets to witness a divergence between these two metrics. This variance typically manifests as a positive difference, as buyers infrequently find financial advantage in paying more than the asking price set by sellers. Numerous dynamic factors may underpin the degree of the observed discrepancy.

A pertinent factor influencing this discrepancy is the robustness of market demand for a specific category of properties. Flourishing economies, fueled by affordable credit, may engender a climate in which the growth of demand surpasses the expansion of supply. Under these market conditions, the noted discrepancy might invert to a negative value due to competitive bidding wars that result in transaction prices exceeding the initially listed prices. Conversely, when confronted with anemic demand, there may be an elongation in the duration properties remain listed in the market. Sellers exhibiting minimal patience for procuring a buyer may be enticed to finalize transactions at prices lower than initially listed.

Additionally, the degree of market dynamism significantly impacts the relationship between listing prices and transaction prices. In a market characterized by stability, the observed discrepancy may be comparatively minor, attributable to a greater convergence between sellers' and buyers' expectations. In contrast, within a volatile market, an escalating disparity might be evident, arising from swift alterations in market conditions and the consequent shifts in the expectations of market participants. High market dynamism may harm the applicability of real estate pricing models due to they decaying accuracy (Robson and Downie, 2008).

Other elements influencing the relationship between ask prices and transaction prices encompass the precision of the listing price, the negotiation prowess and leverage of both the buyer and seller, as well as the timeline designated for the transaction. These facets may be subject to fluctuations driven by demographic shifts, the accessibility and veracity of market information available to the public, and alterations in overall market sophistication.

The aims of this chapter are twofold: first the incremental data collection process is described; second the resulting dataset characteristics are documented and explored.

2.1. Data collection process

2.1.1. Introduction

Web scraping constitutes an expanding domain, purposed to harness publicly accessible Internet data sources for enhancing analytics and digital offerings. This segment delves into the intricacies of an incremental data acquisition methodology for online real estate listings across numerous OECD nations. Essential technologies and techniques are examined to afford a comprehensive comprehension of the entailed difficulties and constraints of the dataset.

2.1.2. Web Scraping Process

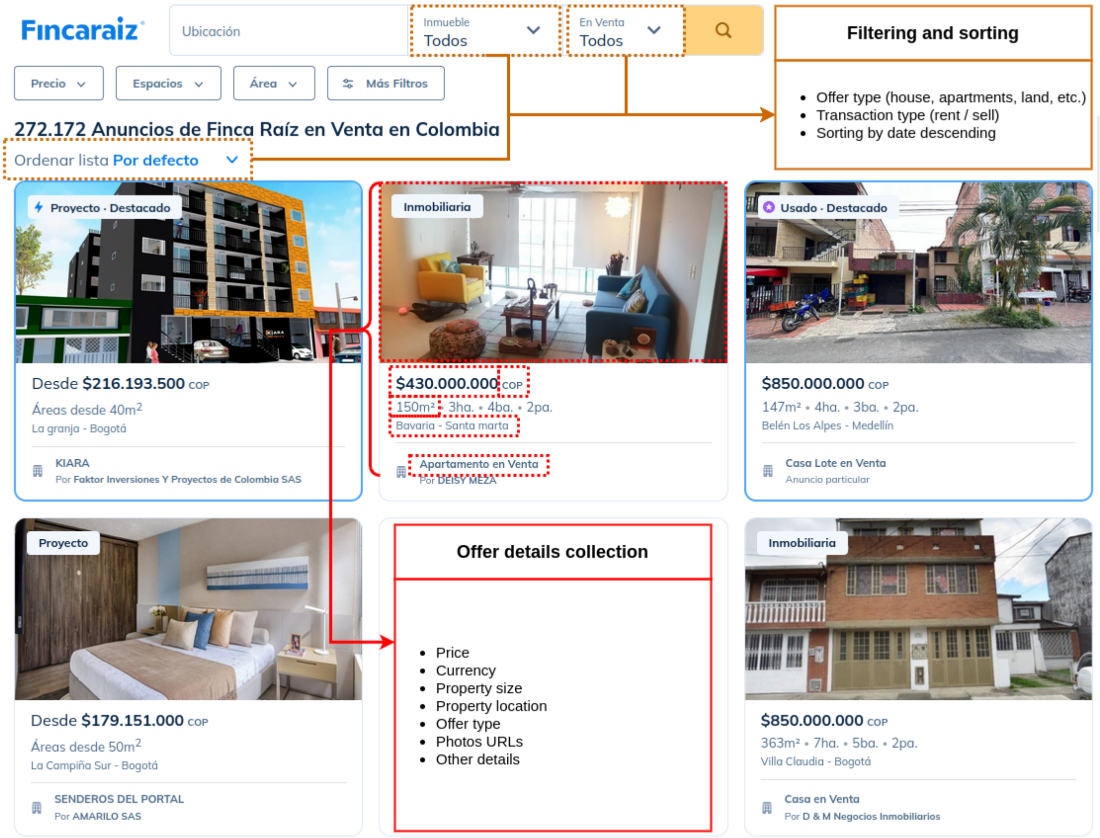
Web scraping, interchangeably referred to as web crawling or scraping, represents a data engineering methodology aimed at integrating internet data sources. The incorporation of webpages as a data source predominantly transpires without necessitating alterations to the original service. The primary function of web scraping processes lies in the acquisition, transformation, and preservation of data content originating from internet data sources. A substantial portion of webpages accessible on the internet are delivered by the backend software in HTML format.

In executing analytical studies employing web-scraped data, one must judiciously select the websites of interest. For the assembly of this dataset, a compilation of links to 1693 prominent real estate portals was curated, employing the link aggregator allyoucanread.com. This collection incorporated the most frequented property portals spanning nearly every country globally. Ultimately, the list was refined to exclusively encompass countries of interest and websites adhering to the subsequent criteria:

1. No aggressive bot protection present on the website
2. Available information about prices and property sizes
3. Location geographical coordinates available

Countries included in the study have been selected to provide a sample from diverse geographical and economic areas:

- **Colombia:** medium income country with vast areas covered with rainforest.
- **Chile:** mountainous and coastal country with unique geography.
- **Mexico:** huge country with vast rural areas and dense urban centers.



Filtering and sorting

- Offer type (house, apartments, land, etc.)
- Transaction type (rent / sell)
- Sorting by date descending

Offer details collection

- Price
- Currency
- Property size
- Property location
- Offer type
- Photos URLs
- Other details

Figure 2.1: Web scraping - key page components identification with Fincaraiz.com.co example.

- **Netherlands:** high income country with large population density.
- **Poland:** large european country with relatively low urbanization rate.

For each preselected data source, a specialized web scraping process necessitates development. This process should reliably procure HTML content, an outcome achieved through the employment of software retry schemes and rotating proxy pools. The subsequent phase involves transforming the procured HTML into one of the standard analytical formats. Amidst a plethora of tools capable of accomplishing this task, BeautifulSoup, a Python library, distinguishes itself as a remarkably convenient choice. It facilitates HTML querying utilizing CSS selectors. Owing to its user-friendly nature and high performance, it was elected as the tool of choice for this study.

Procured outcomes are archived to the cloud storage layer (AWS S3), ensuring reliability and fault tolerance. The preferred data storage file format for this investigation is Apache Parquet.

Once the web scraping procedure is operational, it necessitates systematic execution to maintain dataset currency. In this investigation, the Dagster ETL platform was elected to orchestrate web scraping workloads and schedules.

2.1.3. Data Cleaning and Validation

To undertake an analysis predicated on the collated dataset, the construction of a data cleaning and validation pipeline is imperative (Dornfest et al., 2018). This step is vital to ascertain that data points chosen for examination accurately represent the selected markets. Data consistency, crucial for international analysis, is maintained and imposed at the level of national markets.

The dataset underwent deduplication and filtration to incorporate solely non-extreme observations, thus circumventing the distortion of statistical analysis results. Properties with prices and sizes beneath the 5th percentile and exceeding the 95th percentile within each nation were excluded to accommodate for data entry errors, which are commonplace in datasets accrued via web scraping methods. As the study significantly hinges on the spatial characteristics of properties, observations devoid of geocoordinates were eliminated.

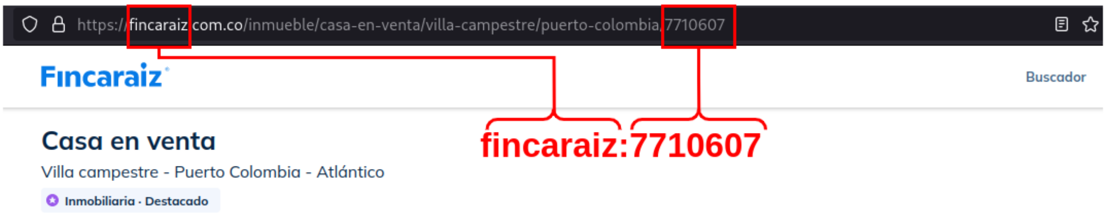
Unidad de Fomento, the Chilean non-circulating currency in which property prices are often presented, has been converted to the *Chilean Peso* using daily exchange rates from Chilean Internal Revenue Service (*Servicio de Impuestos Internos*).

2.1.4. Cloud Technologies

The emergence of cloud technologies has enabled a plethora of novel approaches towards data platforms architecture. For the purpose of this study cloud technologies have been utilized extensively as a compute and data storage layers of the architecture stack. Amazon Web Services Simple Storage Service (AWS S3) has been incorporated as a data storage layer. OVH Cloud VPS provided required compute infrastructure.

2.1.5. Unique Offer IDs

The dataset used for real estate market analysis should be free from observation duplication which can lead to overrepresentation of certain property groups and mislead conclusions. Since the study has limited the data sources selection to non-overlapping, single national websites, it is sufficient to deduplicate datasets within each country. This is achieved with use of unique portal offer IDs.



https://fincaraiz.com.co/inmueble/casa-en-venta/villa-campestre/puerto-colombia/7710607

Figure 2.2: Unique offer ID with Fincaraiz.com.co example.

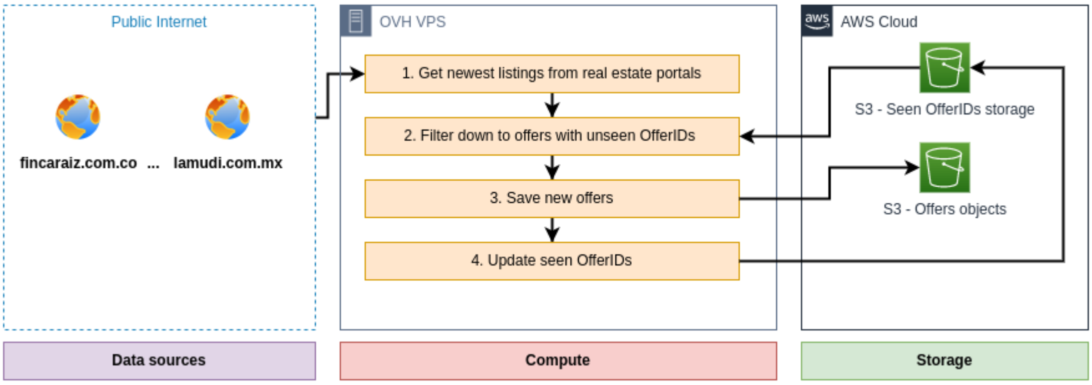


Figure 2.3: Key components of incremental data collection process.

Portal tag (usually derived from the website domain name) followed by colon and alphanumeric string of characters form together an unique property offer identifier which can be used for deduplication purposes. The alphanumeric string is extracted from property offers hyperlinks or respective id holding HTML elements.

2.1.6. Incremental Data Collection Process

The task of real estate market tracking based on Internet portals data requires careful consideration from the temporal modeling perspective. The dataset described in this chapter implements *unitemporal* data model. The single temporal column in the dataset is *date_scraped*, which holds information about the exact date and time when the listing was downloaded by a web scraping process. It has been chosen as an approximation of *bitemporal* data model with columns *date_scraped* and *date_posted* (when the listing was published on the source website). The ability to collect *date_posted* column from the source websites is constrained by limited availability of such information. The approximation has been achieved through incremental data collection process.

Incremental data collection process is organized as an iterative effort to keep the dataset up-to-date with the state of the source website.

1. Visit a source website and get first page of newest real estate listings.

2. Compare IDs of offers listed on the first page with *already seen* IDs.

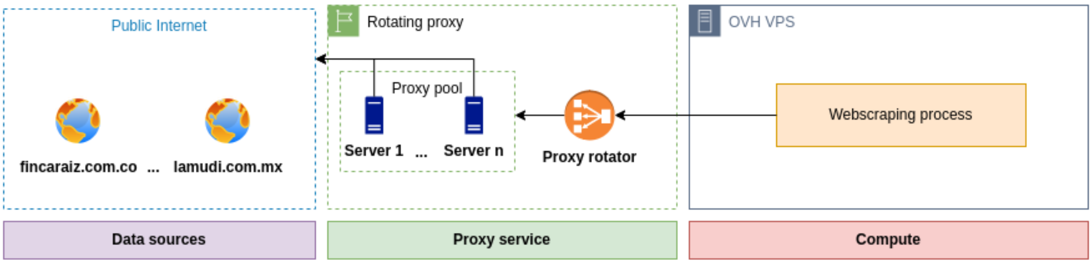


Figure 2.4: Rotating proxy as a component of web scraping pipeline.

3. Save new offers.
4. Update a data structure which holds *already seen* IDs.

If the pace of new offers additions on the source website is lower than collection rate, than the dataset will always reflect the state of the data source. In practice data collection schedule every 15-30 minutes provides sufficient data collection rate with a very conservative safety margin.

2.1.7. Proxies

Proxies play an important role in distributed data collection by providing stability and reliability to the data collection process. Distributed data collection involves collecting data from multiple sources simultaneously, which can result in network congestion, bandwidth limitations, and IP blocking. Proxies can help mitigate these issues by acting as intermediaries between the data collection tool and the target website.

Stability

Proxies can help improve stability by reducing network congestion and limiting the number of requests made to a website. By using multiple proxies, the data collection tool can distribute requests across different IP addresses, reducing the likelihood of getting blocked or blacklisted by the website. Proxies can also be configured to limit the number of requests made to a website within a specific time frame, reducing the risk of triggering rate limits or other security measures.

Reliability

Proxies can also help improve reliability by ensuring that the data collection process continues uninterrupted. If the IP address of the data collection tool gets blocked or blacklisted by the target website, the data collection process can be interrupted or terminated. With the use of proxies, the data collection tool can switch to a different IP address if one becomes blocked or blacklisted, ensuring that the data collection process continues without interruption.

Proxy Rotation

To further improve stability and reliability, proxies can be rotated on a regular basis. Proxy rotation involves switching between different proxies at regular intervals, such as every few minutes or every few hours. This can help prevent the target website from identifying and blocking the data collection tool, as the IP address will change frequently.

Proxy IP pool from which IP addresses are picked to relay request is exposed to the public Internet via proxy router. Proxy router is a component responsible for handling and authenticating incoming network traffic and distributing it across the pool. The traffic distribution may be performed by static iteration over each IP address in the pool or dynamically based on runtime IP addresses performance characteristics such as response times and request success rates.

Distributing the network traffic over a large IP pool does not guarantee a success for data acquisition process. Proxy failures do occur at rates comparable to connection errors to average webpage. Robust retry mechanisms are introduced as mitigation technique for failure prone network operations. Network request retries may be performed after static or dynamically increased time intervals called backoff. Exponential backoff is one of frequently used techniques.

Network requests failures may be caused by website blocks, network errors, server overload, or other. Web scraping processes may include logic to distinguish between these in order to apply appropriate reaction. One notable example is HTTP 429 request status code (*Too Many Requests*). It is returned by server when too much traffic is being sent via network. Clients receiving HTTP 429 should reattempt requests after time specified in the response header field *Retry-After*.

2.1.8. Apache Parquet

As a basis for data engineering operations required to conduct this study Apache Parquet has been chosen as data storage file format. Apache Parquet is an open-source columnar storage format that is designed to optimize data processing for big data analytics. It is highly optimized for performance and is commonly used in big data analytics platforms such as Hadoop, Spark, and Amazon EMR.

Notable benefits of use of Apache Parquet in real estate analytics application that lead to the choice involve:

1. **Columnar Storage:** Apache Parquet stores data in columns instead of rows, which allows for faster and more efficient analytical queries. In real estate analytics, this means that data can be quickly analyzed across a large number of listings or properties.

2. **Compression:** Apache Parquet uses advanced compression techniques to reduce the amount of storage space required for data. This makes it an ideal format for storing large and evolving datasets such as real estate listings, as it can help reduce cloud storage costs. Compression is achieved even in small files.

3. **Schema Evolution:** Apache Parquet supports schema evolution, which means that the schema can be updated over time without breaking existing queries. This is important for web scraping based real estate analytics, as new data may be added over time or the structure of the data may change.

4. **Performance:** Apache Parquet is highly optimized for performance and can handle large datasets with billions of rows and terabytes of data. This means that the format may support the dataset quickly and efficiently, even on at big data scale.

5. **Interoperability:** Apache Parquet is an open-source format and is supported by many big data platforms such as Hadoop, Spark, and Amazon EMR. This means that real

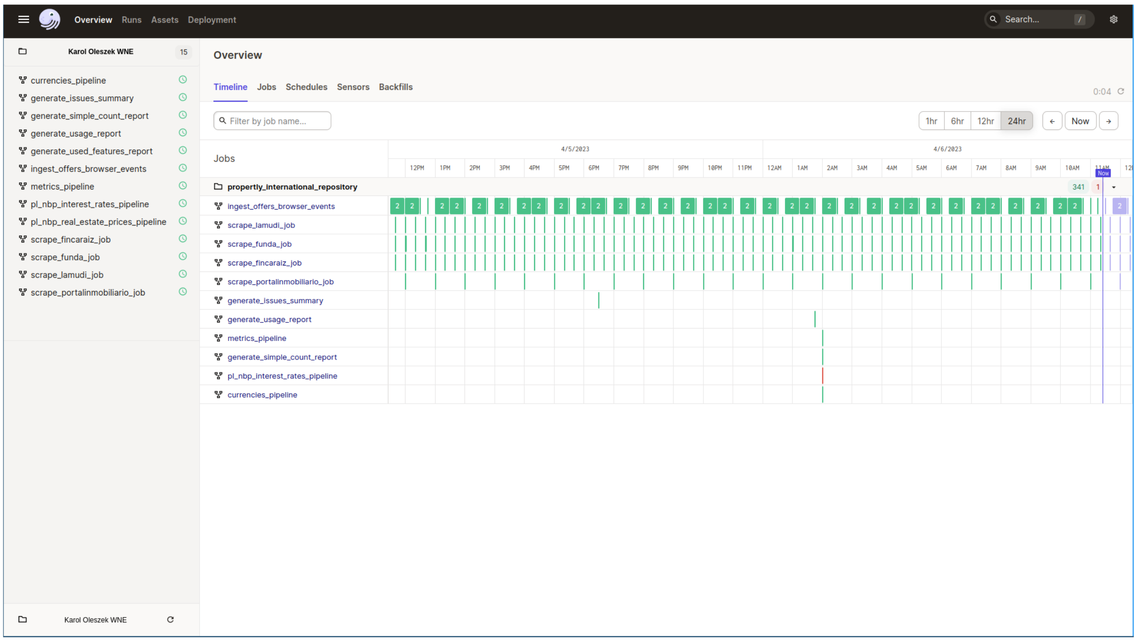


Figure 2.5: Dagster ETL platform - Dagit UI dashboard with dataset scrape processes orchestration.

estate analytics data stored in Parquet format can be easily integrated with other big data tools and platforms.

Apache Parquet is a highly efficient and optimized file format for analytics. Its columnar storage, compression techniques, schema evolution support, performance, and interoperability make it an ideal format for storing and analyzing large datasets in the real estate industry. By using Apache Parquet, analytics can be performed quickly and efficiently, while reducing storage costs and supporting future data growth and changes.

2.1.9. Dagster

Orchestrating web scraping processes which collect information from multiple countries with different schedules and operational setup poses a significant data engineering challenge. Web scraping scripts created for this study are managed with the use of Dagster.

Dagster is an open-source data orchestrator tool that can be used to manage and execute complex workflows for data processing, including web scraping workloads. It provides a simple and modular way to define the dependencies between the different components of a data pipeline, making it easy to test and debug the entire workflow.

Key characteristics which drove the Dagster adoption for web scraping workloads are:

1. **Declarative data pipeline definitions:** Dagster allows to declaratively define a data pipeline as a collection of independent components, each of which performs a specific task in the data processing workflow. For web scraping workloads, these components include tasks such as fetching web pages, extracting data, and storing results in a data storage layer.

2. **Dependency management:** With Dagster, dependencies between the different components of the pipeline are clearly defined, ensuring that each task is executed in the

correct order. The task of extracting data from a web page can only be performed after the task of fetching the web page is complete.

3. **Running the pipeline:** Once the pipeline is defined, Dagster is used to run the pipeline on a distributed computing infrastructure - Kubernetes cluster in the cloud. This allows pipelines to handle large volumes of data and to distribute the workload across multiple machines.

4. **Monitoring and debugging:** Dagster provides built-in tools for monitoring the progress of the data pipeline and identifying errors or issues. For web scraping workloads, this is particularly useful for identifying and addressing issues related to network connectivity, data quality, or performance.

5. **Testing and validation:** With Dagster, the task of testing and validation of each component of the data pipeline is performed both in isolation, as well as the entire workflow as a whole. This helps to ensure that the pipeline is functioning correctly and that it is producing accurate results.

2.1.10. AWS S3

This study is an attempt to deliver a dataset which can be used for analytics in the future. All data acquisition pipeline components were chosen to ensure project scalability beyond the scope of this study. A critical choice in any data project has to be made about a way in which data is being stored. Data storage layer should ensure linear scaling with dataset size, flexibility and cost effectiveness. After careful consideration of available databases and systems AWS S3 has been chosen for this dataset.

AWS S3 (Amazon Web Services - Simple Storage Service) is a cloud-based object storage service that provides a highly scalable and durable platform for storing and retrieving data. In context of storing incrementally collected real estate listings datasets in Parquet format it manifests high performance, scalability, and cost-effectiveness. Apache Parquet compression helps to further reduce data storage costs.

AWS S3 - technical considerations

There are several technical consideration which make AWS S3 a good choice for storing incrementally collected real estate listings datasets in Parquet format. These include:

1. **Scalability:** S3 provides virtually unlimited storage capacity, which makes it well-suited for storing growing datasets such as real estate listings.

2. **Durability:** S3 provides 99.999999999% durability[1], which means that the data is highly resistant to loss or corruption.

3. **Availability:** S3 provides high availability, which means that the data can be accessed quickly and reliably from anywhere in the world.

4. **Cost-effectiveness:** S3 is a cost-effective storage solution, as it provides usage based pricing model.

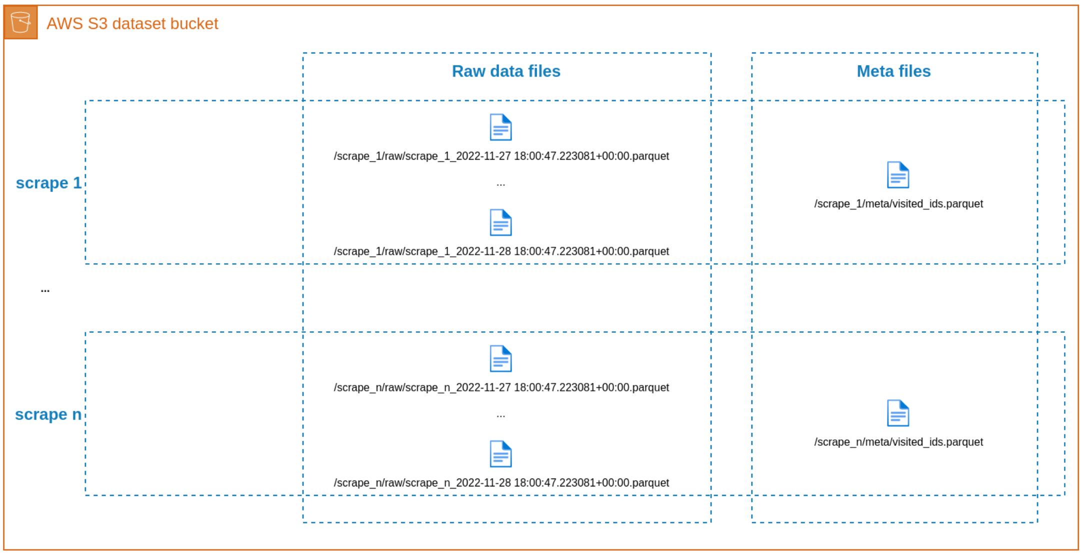

AWS S3 dataset bucket
Raw data files
Meta files
scrape 1
/scrape_1/raw/scrape_1_2022-11-27 18:00:47.223081+00:00.parquet
...
/scrape_1/raw/scrape_1_2022-11-28 18:00:47.223081+00:00.parquet
/scrape_1/meta/visited_ids.parquet
...
scrape n
/scrape_n/raw/scrape_n_2022-11-27 18:00:47.223081+00:00.parquet
...
/scrape_n/raw/scrape_n_2022-11-28 18:00:47.223081+00:00.parquet
/scrape_n/meta/visited_ids.parquet

Figure 2.6: Dataset persistence storage layer - AWS S3 key structure matrix.

Data storage structure

Although AWS S3 is a key-value object store, object keys can include forward slashes characters '/' and key prefixes can form logical groups *similar* to directories known from desktop file systems. This convention has adopted in the study. Each web scraping process writes data to a separate root level prefix. Data chunks and IDs metadata Apache Parquet files are stored in separate prefixes under each web scraping prefix. This convention forms a logical scrape <-> data/metadata matrix which helps to maintain uniformity and consistency across multiple scraping processes.

Retrieving the data from S3

Dataset snapshot has to be generated from data chunks stored on AWS S3 in order to perform analytics tasks. The process involves reading all data chunks contents into analytics compute and concatenating them into one data file. Dataset rows may be filtered to only include certain time period. After filtering step, the dataset is persisted to local disk for later use.

2.1.11. Compute infrastructure

The remaining infrastructure component required for dataset acquisition pipeline is compute layer and containers orchestrator. Kubernetes has been chosen to orchestrate Docker containers, OVH Cloud VPS was used as a compute layer.

Kubernetes

Dagster task orchestrator deployment has been paired with Kubernetes cluster. Dagster workers are deployed in Docker containers on Kubernetes and so are Dagster UI and main

[1] AWS S3 Data Durability.

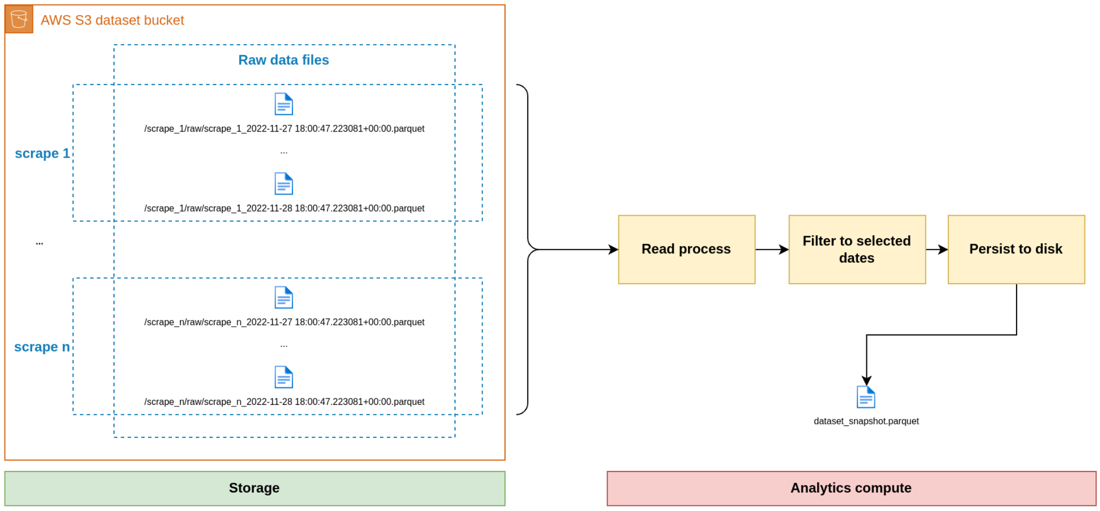

AWS S3 dataset bucket
Raw data files
scrape 1
/scrape_1/raw/scrape_1_2022-11-27 18:00:47.223081+00:00.parquet
...
/scrape_1/raw/scrape_1_2022-11-28 18:00:47.223081+00:00.parquet
...
scrape n
/scrape_n/raw/scrape_n_2022-11-27 18:00:47.223081+00:00.parquet
...
/scrape_n/raw/scrape_n_2022-11-28 18:00:47.223081+00:00.parquet
Read process
Filter to selected dates
Persist to disk
dataset_snapshot.parquet
Storage
Analytics compute

Figure 2.7: Dataset snapshot generation process.

engine components. Kubernetes is an open-source platform designed to automate the deployment, scaling, and management of containerized applications. It is a highly versatile choice for organizations and individuals that require a robust and efficient system for managing complex, multi-container architectures.

Key Kubernetes traits that drove this technical choice were:

1. **Scalability:** Kubernetes supports horizontal scaling, allowing the platform to handle an increase in demand by adjusting the number of running containers. This enables Dagster to easily scale its workers up or down based on the needs, ensuring that resources are used efficiently.

2. **Reliability:** Kubernetes ensures high availability of applications by redistributing workloads in case of a failure and providing self-healing capabilities, such as auto-restarting, re-scheduling, and replicating containers. The platform is designed with a built-in control plane to maintain system stability.

3. **Security:** Kubernetes provides robust security features, including secret management to handle sensitive data, network policies to control access to applications, and built-in service accounts. Moreover, the Kubernetes ecosystem is rich with additional security tools and add-ons that can be used to further enhance the security posture of the applications.

4. **Cost-effectiveness:** Kubernetes, being an open-source platform, significantly reduces software costs. Additionally, efficient resource utilization, facilitated by Kubernetes' automatic scaling and load balancing features, can result in considerable savings in infrastructure costs.

OVH Cloud VPS

Compute layer is a foundational part of any data pipeline. OVH Cloud VPS is a cloud-based virtual private server solution that provides a scalable and flexible hosting environment for a wide range of applications. OVH Cloud VPS has been chosen due to its cost-effectiveness and sufficient reliability.

2.2. Dataset variables

2.2.1. Price

Table 2.1: Dataset variables - price

Attribute	**Description**
Name	Price
Identifier	price
Units	Denoted by Currency Code
Type	Continuous
Range	5th - 95th percentile of prices in each country
Transformations	For Chile *Unidad de Fomento* transformation
Source	Real estate listings

The 'Price' attribute signifies the listing price of a property and serves as a continuous variable in our dataset. The range for this attribute is between the 5th and 95th percentile of all property prices within each respective country. For Chilean properties, the pricing is adjusted with the 'Unidad de Fomento' transformation, which is a unit of account used in Chile that is regularly adjusted for inflation. The data for this attribute is directly derived from real estate listings.

2.2.2. Currency Code

Table 2.2: Dataset variables - Currency Code

Attribute	**Description**
Name	Currency Code
Identifier	currency_code
Type	Categorical (ISO 4217)
Transformations	None
Source	Real estate listings

In the presented dataset, the 'Currency Code' represents a categorical variable adhering to the ISO 4217 standard. This attribute, derived directly from real estate listings, requires no transformation for analysis and provides essential context for interpreting the corresponding price values.

2.2.3. Property Size

The 'Property Size' attribute in the dataset represents a continuous variable detailing the size of a property in square meters. Sourced directly from real estate listings and encompassing a range of 10 to 500 square meters, this attribute does not necessitate any transformation for further analytical operations.

2.2.4. Date Scraped

The attribute 'Date Scraped' is a date-type variable in the dataset, formatted as 'YYYY-MM-DD'. This attribute, generated by a web scraping script, signifies the date at which the

Table 2.3: Dataset variables - Property Size

Attribute	**Description**
Name	Property Size
Identifier	property_size
Units	Square Meter
Type	Continuous
Range	10 - 500
Transformations	None
Source	Real estate listings

Table 2.4: Dataset variables - Date Scraped

Attribute	**Description**
Name	Date Scraped
Identifier	date_scraped
Units	YYYY-MM-DD
Type	Date
Transformations	None
Source	Web scraping script

corresponding property information was retrieved, requiring no transformations for downstream analysis.

2.2.5. Offer Type

Table 2.5: Dataset variables - Offer Type

Attribute	**Description**
Name	Offer Type
Identifier	offer_type
Type	Categorical
Categories	Rent, sale for houses, apartments, commercial buildings, land and other
Transformations	Unification across data sources
Source	Real estate listings

The 'Offer Type' attribute within the dataset denotes a categorical variable that distinguishes the nature of the property listing, such as rent or sale for a variety of property types including houses, apartments, commercial buildings, land, among others. This information is sourced directly from real estate listings. Given its categorical nature and potential inconsistencies across data sources, it required unification and standardization procedures to ensure uniformity and ease of analysis.

2.2.6. Portal Offer ID

The 'Portal Offer ID' is a text-type attribute in the dataset, signifying the unique identifier for a property listing as provided by the listing portal. This identifier is procured directly

Table 2.6: Dataset variables - Portal Offer ID

Attribute	**Description**
Name	Portal Offer ID
Identifier	portal_offer_id
Type	Text
Transformations	Website tag prepended
Source	Real estate listings

from real estate listings. To enhance the traceability and uniqueness of each listing across various sources, a transformation is implemented to prepend the website tag to the original offer ID.

2.2.7. Offer Original URL

Table 2.7: Dataset variables - Offer Original URL

Attribute	**Description**
Name	Offer Original URL
Identifier	offer_original_url
Type	Text
Transformations	None
Source	Real estate listings

The 'Offer Original URL' attribute within the dataset is a text-type variable, which represents the original web address from which a property listing was obtained. Sourced directly from the real estate listings, this attribute requires no transformations and serves as a vital reference for accessing the original online property listing.

2.2.8. Title

Table 2.8: Dataset variables - Title

Attribute	**Description**
Name	Title
Identifier	title
Type	Text
Transformations	HTML tags cleanup
Source	Real estate listings

The 'Title' attribute in the dataset constitutes a text-type variable, encapsulating the headline or primary descriptor of the property listing. Extracted directly from real estate listings, this attribute required a transformation process - HTML tags cleanup, to ensure readability and optimal utility in subsequent analyses.

Table 2.9: Dataset variables - Description

Attribute	**Description**
Name	Description
Identifier	description
Type	Text
Transformations	HTML tags cleanup
Source	Real estate listings

2.2.9. Description

The 'Description' attribute in the dataset is a text-type variable providing a detailed account of the property listing. This attribute, sourced directly from real estate listings, may potentially contain extraneous HTML tags. Hence, a transformation process, specifically HTML tags cleanup, has been used to maintain clean, comprehensible text data for further analyses.

2.2.10. Photos URLs

Table 2.10: Dataset variables - Photos URLs

Attribute	**Description**
Name	Photos URLs
Identifier	photos_urls
Type	List
Transformations	None
Source	Real estate listings

The 'Photos URLs' attribute within the dataset represents a list-type variable, containing a collection of URLs corresponding to images associated with the property listing. Derived directly from real estate listings, this attribute does not necessitate any transformations and provides an option to source visual context to supplement the textual property information.

2.2.11. Address Country

Table 2.11: Dataset variables - Address Country

Attribute	**Description**
Name	Address Country
Identifier	address_country
Type	Categorical
Transformations	None
Source	Real estate listings

The 'Address Country' attribute in the dataset is a categorical variable that specifies the country location of the property listing. Sourced directly from real estate listings, this attribute does not require any transformations and provides a crucial geographical context for each property listing.

2.2.12. Address Raw

Table 2.12: Dataset variables - Address Country

Attribute	**Description**
Name	Raw Address
Identifier	address_raw
Type	Text
Transformations	Text standardization
Source	Real estate listings

The 'Raw Address' attribute represents a text-type variable in the dataset, supplying the unprocessed, original address of the property listing. This information, extracted directly from real estate listings, may necessitate transformations such as text standardization to ensure uniformity and ease of data analysis.

2.2.13. Issuer Type

Table 2.13: Dataset variables - Issuer Type

Attribute	**Description**
Name	Issuer Type
Identifier	issuer_type
Type	Categorical
Categories	Individual, Agent, Developer
Transformations	None
Source	Real estate listings

The 'Issuer Type' attribute within the dataset constitutes a categorical variable, distinguishing who has listed the property: an individual, agent, or developer. This information is sourced directly from real estate listings and does not require any transformations, providing valuable insight into the listing party's category for each property.

2.2.14. Tag

Table 2.14: Dataset variables - Tag

Attribute	**Description**
Name	Tag
Identifier	tag
Type	Categorical
Transformations	None
Source	Web scraping script

The ’Tag’ attribute marks each property listing with the data source (website).

2.2.15. Location Longitude

Table 2.15: Dataset variables - Location Longitude

Attribute	**Description**
Name	Location Longitude
Identifier	location_longitude
Units	Decimal Degrees
Type	Continuous
Range	-180 to 180
Transformations	Country level filtering
Source	Real estate listings

The ’Location Longitude’ attribute constitutes a continuous variable within the dataset, delineating the longitudinal geographical coordinate of the property listing. This attribute, extracted from real estate listings, is represented in decimal degrees with a range from -180 to 180. For focused analysis, country-level filtering has been employed as a transformation method.

2.2.16. Location Latitude

Table 2.16: Dataset variables - Location Latitude

Attribute	**Description**
Name	Location Latitude
Identifier	location_latitude
Units	Decimal Degrees
Type	Continuous
Range	-90 to 90
Transformations	Country level filtering
Source	Real estate listings

The ’Location Latitude’ attribute represents a continuous variable in the dataset, defining the latitudinal geographical coordinate of the property listing. Derived from real estate listings, this attribute is presented in decimal degrees and spans a range from -90 to 90. To facilitate specific geographic analysis, a country-level filtering has been executed.

2.2.17. Restaurants

The ’Restaurants’ attribute is a non-negative integer variable that corresponds to the quantity of dining establishments in proximity to a given property listing. This data, obtained through augmentation from OpenStreetMap, is grouped into three distinct distance levels. This approach enhances the understanding of the property’s location relative to local amenities and lifestyle conveniences.

Table 2.17: Dataset variables - Restaurants

Attribute	**Description**
Name	Restaurants
Identifier	restaurants
Type	Non-negative integer
Transformations	Grouped into 3 distance levels
Source	Data augmentation - OpenStreetMap

2.2.18. Schools

Table 2.18: Dataset variables - Schools

Attribute	**Description**
Name	Schools
Identifier	schools
Type	Non-negative integer
Transformations	Grouped into 3 distance levels
Source	Data augmentation - OpenStreetMap

The 'Schools' attribute signifies the number of educational institutions within certain distance levels around a property, quantified as a non-negative integer. This data is procured through the augmentation from OpenStreetMap and subsequently categorized into three distinct distance levels. The categorization enriches the understanding of the property's environment, particularly in terms of its proximity to educational facilities, an important consideration for many prospective buyers or tenants.

2.2.19. Transport

Table 2.19: Dataset variables - Transport

Attribute	**Description**
Name	Transport
Identifier	transport
Type	Non-negative integer
Transformations	Grouped into 3 distance levels
Source	Data augmentation - OpenStreetMap

The 'Transport' attribute quantifies the presence of tram, subway, and bus stops within predefined radius levels around a property, denoted as a non-negative integer. This data is sourced through augmentation from OpenStreetMap and subsequently stratified into three distance levels. This spatially referenced data informs potential buyers or tenants about the accessibility of public transportation from the property, an influential factor in real estate decision-making.

Table 2.20: Dataset variables - Parks

Attribute	**Description**
Name	Parks
Identifier	parks
Type	Non-negative integer
Transformations	Grouped into 3 distance levels
Source	Data augmentation - OpenStreetMap

2.2.20. Parks

The 'Parks' attribute represents the number of public parks located within predefined distance categories from the property, enumerated as a non-negative integer. Derived from the OpenStreetMap database, this attribute is grouped into three distance levels post-acquisition. This variable provides an indication of recreational and green spaces surrounding a property, elements often associated with the livability of a neighborhood and thus the attractiveness of a property.

2.2.21. Universities

Table 2.21: Dataset variables - Universities

Attribute	**Description**
Name	Universities
Identifier	universities
Type	Non-negative integer
Transformations	Grouped into 3 distance levels
Source	Data augmentation - OpenStreetMap

The 'Universities' attribute represents the count of universities located within pre-established distance bands from the property, quantified as a non-negative integer. This data is gathered from the OpenStreetMap database, and subsequently grouped into three distance levels. The proximity to higher education institutions is a variable of interest as it may impact the value and desirability of a property, particularly for students, faculty, or those valuing educational accessibility.

2.3. Dataset empirical characteristics

Presented are latest statistics on population, surface, population density and urbanization rates reported at World Bank Open Data. Presented maps are rectangular projections of calculated price grids (See Chapter 3). Countries proportion may not match popular map projections e.g. Mercator projection.

2.3.1. Colombia

Table 2.22: Colombia - basic information

Detail	Value
Population	51,516,562
Area	1,109,500 km^2
Population density	$46/km^2$
Urbanization rate	82%

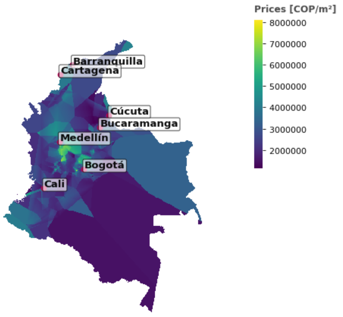


Figure 2.8: Colombian housing market - prices spatial distribution

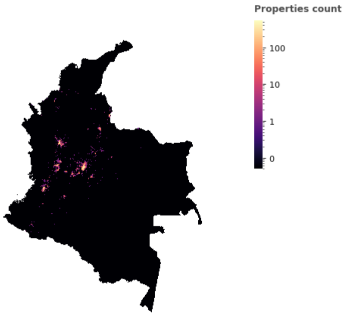


Figure 2.9: Colombian housing market - offers density

The scrutinized Colombian housing market exhibits pronounced concentration within the principal metropolitan areas. The urban triangle encompassing Bogotá, Medellín, and Cali epitomizes the costliest and most dynamic segment of the Colombian housing market. The extensive terrain covered by the Amazonian forest constitutes a significant proportion of Colombian territory, where market activity is discernibly minimal. Generally, house prices in these remote and inaccessible regions are markedly lower than those in proximity to urban centers.

Table 2.23: Colombian housing market - descriptive statistics

	Price [COP]	Price per m^2 [COP]	Size [m^2]
μ	763,384,296	3,558,622	209
σ	614,141,206	1,633,832	111
Minimum	50,000,000	1,042,945	20
Q1	330,000,000	2,321,429	120
Q2	580,000,000	3,250,000	187
Q3	980,000,000	4,454,545	280
Maximum	4,500,000,000	9,392,265	500

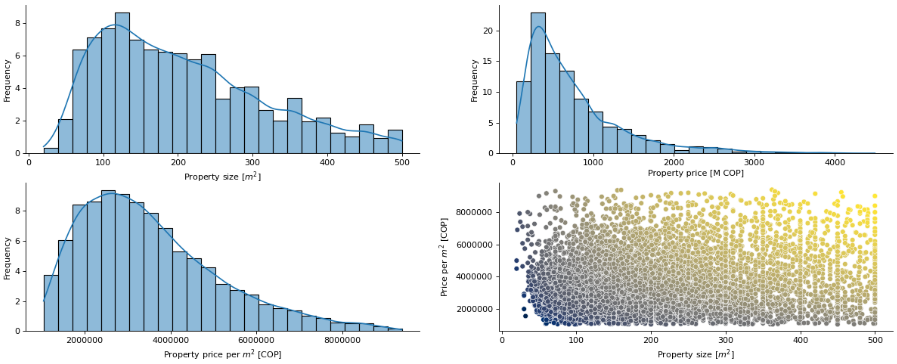


Figure 2.10: Colombian housing market - prices and sizes distributions

Reported statistics indicate that houses in Colombia on average tend to be larger than in other countries included in the study.

Table 2.24: Colombian housing market - dataset details

Detail	Value
Data time range	From 2023-03-27 to 2023-05-01
Observations	17165

2.3.2. Chile

Table 2.25: Chile - basic information

Detail	Value
Population	19,493,184
Area	743,532 km^2
Population density	$26/km^2$
Urbanization rate	82%

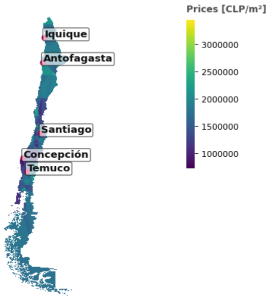


Figure 2.11: Chilean housing market - prices spatial distribution

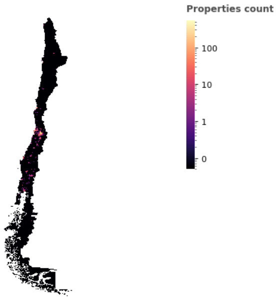


Figure 2.12: Chilean housing market - offers density

The Chilean housing market under examination is typified by a pronounced concentration within the Santiago metropolitan region. Other significant regions, encompassing Concepción, Temuco, and Antofagasta, account for the majority of housing market activity. The remote regions within the Andes are distinguished by minimal market activity and diminished price levels. Generally, when disregarding the inaccessible sections of the country, house prices exhibit a relatively uniform distribution across both rural and urban sectors.

Table 2.26: Chilean housing market - descriptive statistics

	Price [CLP]	Price per m^2 [CLP]	Size [m^2]
μ	435,457,600	2,157,531	194
σ	332,079,784	934,230	108
Minimum	15,900,000	650,379	13
Q1	178,899,584	1,409,911	110
Q2	342,382,971	2,001,118	160
Q3	600,000,000	2,782,841	260
Maximum	2,330,355,300	4,893,746	500

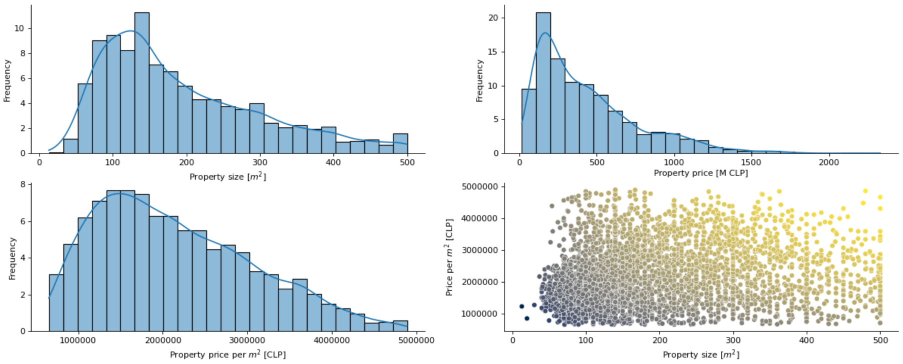


Figure 2.13: Chilean housing market - prices and sizes distributions

Reported statistics indicate that houses in Chile on average tend to be smaller than in other South American countries included in the study.

Table 2.27: Chilean housing market - dataset details

Detail	Value
Data time range	From 2022-11-24 to 2023-05-01
Observations	4690

2.3.3. Mexico

Table 2.28: Mexico - basic information

Detail	Value
Population	126,705,138
Area	1,943,950 km^2
Population density	$65/km^2$
Urbanization rate	81%

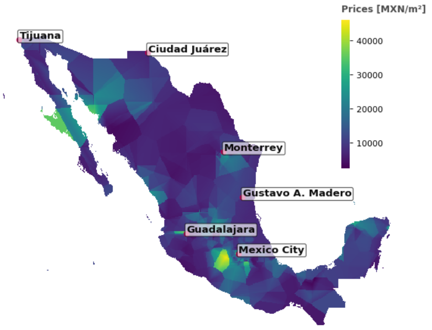


Figure 2.14: Mexican housing market - prices spatial distribution

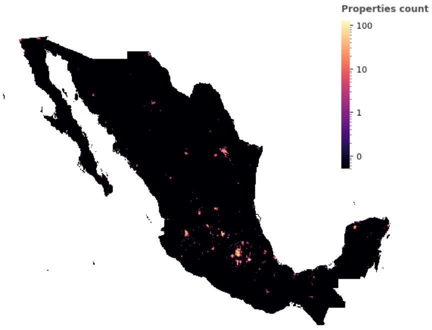


Figure 2.15: Mexican housing market - offers density

The scrutinized Mexican housing market displays significant activity primarily in the central region of the country, predominantly on the peripheries of Mexico City and other urban areas. The dataset offers minimal coverage of the extensive rural areas in Mexico. This limited rural coverage could be attributable to the restrained digitization of the market in the less developed Mexican regions. On average, price levels are elevated in central Mexico, regions in proximity to the United States, and the Yucatan Peninsula.

Table 2.29: Mexican housing market - descriptive statistics

	Price [MXN]	Price per m^2 [MXN]	Size [m^2]
μ	2,817,915	13,681	189
σ	3,203,673	9,689	100
Minimum	269,337	2,232	20
Q1	750,000	5,430	110
Q2	1,460,000	11,883	167
Q3	3,694,500	19,410	250
Maximum	27,000,000	59,375	500

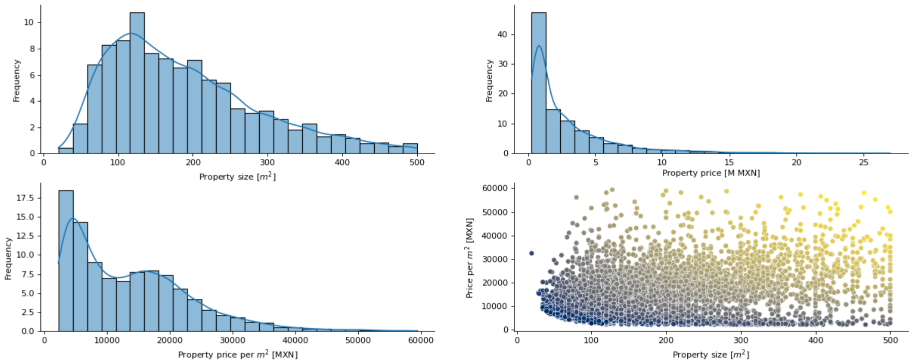


Figure 2.16: Mexican housing market - prices and sizes distributions

Reported statistics indicate that housing market in Mexico has an unusually high number of smaller (up to 120 square meters), cheap homes (up to 10k MXN per square meter). They make up for almost half of the market.

Table 2.30: Mexican housing market - dataset details

Detail	Value
Data time range	From 2023-03-25 to 2023-05-01
Observations	7275

2.3.4. Netherlands

Table 2.31: Netherlands - basic information

Detail	Value
Population	17,533,044
Area	33,670 km^2
Population density	$518/km^2$
Urbanization rate	93%

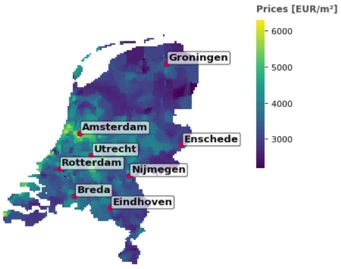


Figure 2.17: Dutch housing market - prices spatial distribution

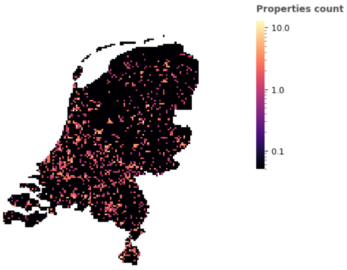


Figure 2.18: Dutch housing market - offers density

The Dutch housing market under scrutiny exhibits consistent activity across both urban and rural regions. The highest price levels are registered in the vicinity of Amsterdam and Rotterdam, with relatively elevated prices persisting in the central and southern sectors of the country. Contrastingly, the northern segment of the country displays the lowest price levels.

Table 2.32: Dutch housing market - descriptive statistics

	Price [EUR]	Price per m^2 [EUR]	Size [m^2]
μ	483,074	3,594	134
σ	235,983	923	50
Minimum	145,000	2,110	40
Q1	325,000	2,895	103
Q2	425,000	3,422	123
Q3	569,000	4,114	150
Maximum	2,495,000	6,860	476

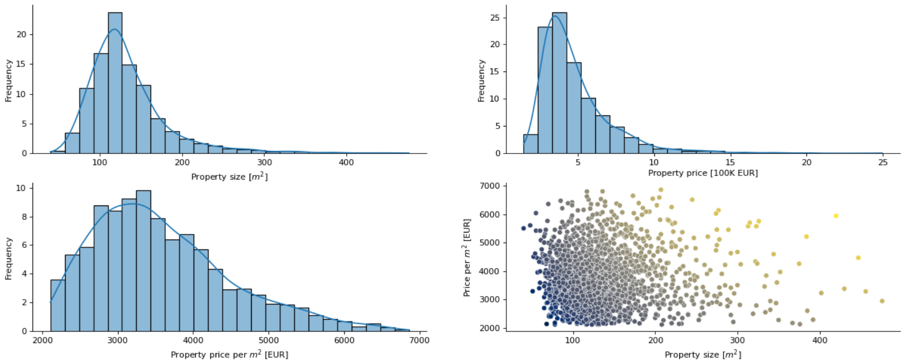


Figure 2.19: Dutch housing market - prices and sizes distributions

Reported statistics indicate that variability in house prices and sizes in Netherlands is limited. The majority of the offers on market fall into narrow range of 100-200 square meters in size and with prices in range of 2500-4500 EUR per square meter.

Table 2.33: Dutch housing market - dataset details

Detail	Value
Data time range	From 2023-04-10 to 2023-05-01
Observations	2268

2.3.5. Poland

Table 2.34: Poland - basic information

Detail	Value
Population	37,747,124
Area	306,130 km^2
Population density	$124/km^2$
Urbanization rate	60%

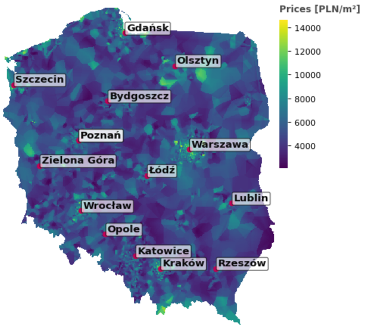


Figure 2.20: Polish housing market - prices spatial distribution

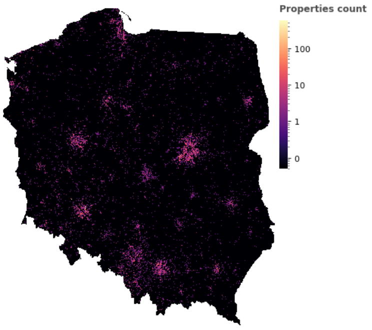


Figure 2.21: Polish housing market - offers density

The scrutinized Polish housing market displays vigorous activity within the primary metropolitan regions, with substantial clusters of activity proximate to Warsaw, Cracow, Wrocław, Gdańsk, and Poznań. Silesia distinguishes itself from other regions with its dispersed housing market devoid of a central point. Price levels peak within urban areas, along the Baltic Sea coast, and in the mountainous regions near the southern border.

Table 2.35: Polish housing market - descriptive statistics

	Price [PLN]	Price per m^2 [PLN]	Size [m^2]
μ	944,818	6,152	155
σ	624,515	2,398	74
Minimum	59,000	2,072	17
Q1	569,000	4,445	103
Q2	770,000	5,810	138
Q3	1,149,900	7,413	184
Maximum	6,500,000	14,650	500

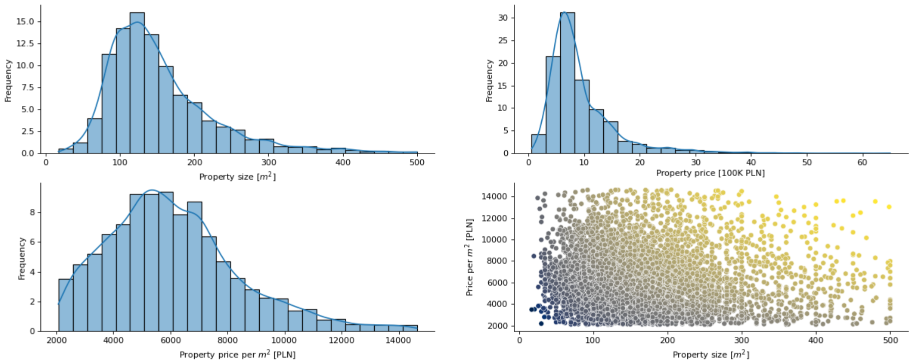


Figure 2.22: Polish housing market - prices and sizes distributions

Reported statistics indicate that housing market in Poland manifests greater variability in house sizes than in price levels.

Table 2.36: Polish housing market - dataset details

Detail	Value
Data time range	From 2023-01-01 to 2023-05-01
Observations	33510

2.4. Conclusions

The web scraping data collection process is an effective method for collecting data from websites on a large scale. The process involves identifying the websites that contain the data of interest, developing a web scrape process to collect the data, and cleaning and validating the data. For this study, web scraping was used to collect data on house sales offers from a number of countries. To ensure data collection process robustness and scalability a number of technical decision had to be made. Successful application of web scraping techniques requires careful balancing between aggressive data acquisition and integration of reliability and sustainability considerations. The next chapter will describe the data analytical process used to asses performance of alternative location based house pricing models withe the use of collected dataset.

Chapter 3

Model

One of the goals of this study is to research alternative property pricing methods without the use of any hedonic property features apart from property size while achieving robust ex-ante prediction performance. Traditionally used variables describing property condition and amenities are absent in the proposed approach. The study does not disregard the link between hedonic motivations of home buyers and property prices. It is observed though that different price levels may be explained with limited errors using only localization, neighborhood amenities and property size information. To achieve aforementioned research goal, a pricing model with a spatially autoregressive component has been constructed. With the use of a housing pricing model and its diagnostics the following hypotheses are evaluated:

1. Pricing houses solely based on their locations and sizes yields accurate results.

2. Inclusion of information about property neighborhood amenities improves pricing models performance.

    (a) Information about quantity of nearby parks improves pricing accuracy.

    (b) Information about quantity of nearby restaurants improves pricing accuracy.

    (c) Information about quantity of nearby schools improves pricing accuracy.

    (d) Information about quantity of nearby bus, tram and subway stops improves pricing accuracy.

    (e) Information about quantity of nearby universities improves pricing accuracy.

Proving hypothesis 1. may provide evidence towards claims that housing markets demonstrate structural homogeneity in the local spatial groupings. That could mean that local property groupings are either in relatively similar condition and/or the market does not perceive their condition to be relevant from the pricing perspective. The homogeneity hypothesis falls out of the scope of this study.

Proving hypothesis 2. may provide useful clues for choices made by AVM/CAMA/other pricing models developers and researchers in the practical setting.

The pricing model with a spatially autoregressive component used in the study has been trained on international dataset described in the previous chapter.

The chapter starts with model formulation, training and inference setting. Then model performance is evaluated in regressor, urban/rural, market segment and feature set breakdowns.

3.1. Model formulation

3.1.1. Price grid

The spatial autoregressive pricing model (SAR) is a popular modeling choice used to incorporate spatial relationships into analytics process. The proposed modeling approach shares a spatial component with SAR models.

Traditionally SAR models are formulated as (d'Amato, 2017a):

$$Y = X\beta + \rho W_y Y + \epsilon \tag{3.1}$$

where Y is a vector of housing prices, X is a matrix of hedonic variables, β is a vector of coefficients, ρ is a spatial autoregressive coefficient, W_y is a spatial weight matrix, and ϵ is the error term.

Setting spatial matrix weights can be performed in a way that for each property only neighboring properties are contributing to the pricing prediction. Specifically, they can be set to include only observations from within the same cell from a rectangular spatial grid.

In this study all properties were projected to a rectangular geocoordinate grid. In each grid cell an average house price per square meter has been calculated. All properties in the dataset have been assigned a new variable *local market housing price per square meter* denoted as p_{local_market}. Vector of local market prices is an equivalent of the spatial autoregressive term.

$$p_{local\_market} = \rho W_y Y \tag{3.2}$$

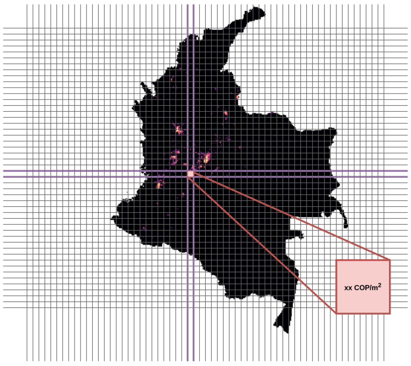


Figure 3.1: National price grids

Grid cell sizes have been set individually for each country in a way that divides large cities (1M+ inhabitants) to at least tens of cells to ensure difference between city districts are properly captured. Due to sparse nature of rural markets, placing dataset on a regular grid leaves a large number of cells without any price level information. Leaving empty grid cells without any price level approximation would greatly constrain model applicability to property pricing at the national level. The approximation for empty cells has been calculated as an average of linear and nearest neighbor interpolation using Numpy numeric computation library.

3.1.2. Formula

The model is defined with a two-stage formula (spatial component, neighborhood component). Multi-level models have been applied to separate value influence from earthquake risks factor in a case of Stambul property market (Dunning et al., 2017).

The model pricing formula is:

$$Y = s_{property} p_{property} + \epsilon_Y \tag{3.3}$$

where $s_{property}$ is house surface in square meters and $p_{property}$ is its predicted price per square meter. ϵ_Y denotes the global error term.

$s_{property}$ is an observed variable, it is property price per meter which needs to predicted. It has been modeled as a sum of local market price per meter and respective deviation from that price and an error term, denoted as $\delta p_{property}$ and $\epsilon_{p_{property}}$ respectively.

$$p_{property} = p_{local\_market} + \delta p_{property} + \epsilon_{p_{property}} \tag{3.4}$$

For each property p_{local_market} can be observed in the national pricing grid. $\delta p_{property}$ is further modeled with $F(X)$, a function approximated using machine learning regression methods. In this study $F(X)$ has been trained with ExtraTreesRegressor and RandomForestRegressor models from Python SciKit-Learn machine learning library.

$$\delta p_{property} = F(X) + \epsilon_{\delta p_{property}} \tag{3.5}$$

Local price per square meter does not introduce further variability into error term, therefore:

$$\epsilon_{p_{property}} = \epsilon_{\delta p_{property}} \tag{3.6}$$

Consequently, the global error term can be calculated as:

$$\epsilon_Y = s_{property} \epsilon_{\delta p_{property}} \tag{3.7}$$

Unwinding the equation yields the pricing formula:

$$Y = s_{property}(p_{local\_market} + F(X) + \epsilon_{\delta p_{property}}) \tag{3.8}$$

Formula yielded from setting $F(X) = 0$ (using only national price grid) is referred to and further evaluated as BaselineModel.

3.1.3. Model training

Pricing model training pipeline requires a series of data transformations before a chosen regressor can be fit. Starting with raw property prices, the pipeline calculates $p_{property}$, a property price per square meter. By subtracting $p_[local\_market]$ a $\delta p_{property}$ is derived and it becomes an explained variable. Dataset localization features and property size act as explanatory variables in the regression problem. For each studied country a machine learning model of choice is trained with the input and output features using 90% randomly chosen national dataset records.

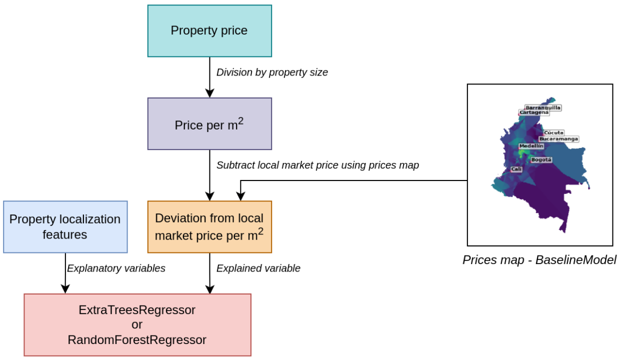

Property price
Division by property size
Price per m^2
Subtract local market price using prices map
Barranquilla
Cartagena
Cucuta
Bucaramanga
Medellin
Bogota
Cali
Prices map - BaselineModel
Property localization features
Deviation from local market price per m^2
Explanatory variables
Explained variable
ExtraTreesRegressor
or
RandomForestRegressor

Figure 3.2: Model training diagram

3.1.4. Inference

Housing pricing prediction requires an inversion of the model training pipeline. localization features and property size are used to infer house price per square meter deviation from the local market. After summing it with local market property prices per square mater $p_{property}$ is derived. Multiplying it with property size yields the price prediction.

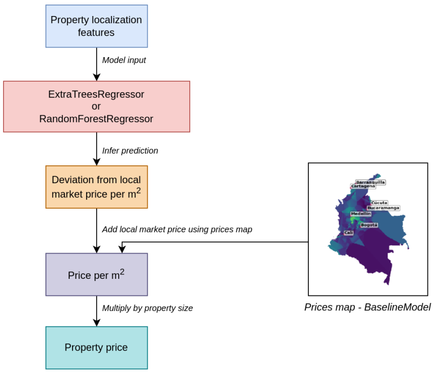

Property localization features
Model input
ExtraTreesRegressor
or
RandomForestRegressor
Infer prediction
Deviation from local market price per m^2
Add local market price using prices map
Barranquilla
Cartagena
Cucuta
Bucaramanga
Medellin
Bogota
Cali
Prices map - BaselineModel
Price per m^2
Multiply by property size
Property price

Figure 3.3: Model inference diagram

3.1.5. RandomForestRegressor

For the purpose of this study two ensemble machine learning techniques have been selected. One of them is Random Forrest algorithm, a method aimed at constructing an ensemble of base regression/classification trees each trained on the training dataset subsample. The randomness is introduced also at the base learner level - each tree split is performed on a random subset of dataset features. Reference SciKit-Learn RandomForestRegressor implementation have been adopted in the study (Breiman, 2001). In the context of the selected features and house prices regression problem, the Random Forrest algorithm has been selected due to:

- **Presence of non-linear features characteristics** - localization features in the form of amenities counts in different distance ranges from the property may be linked to the property price in multiple, nuanced and non-linear ways. Random Forests and their base learners enables the pricing formula to account for that.

- **Importance of pricing model performance** - Random Forests are recognized and established method used to achieve high *ex-post* prediction performance in various regression problems, including property appraisal and pricing. The study aims to maximize prediction performance achieved without the use of traditional hedonic model features.

3.1.6. ExtraTreesRegressor

The Extremely Random Trees algorithm, often referred to as Extra Trees, is an advanced machine learning methodology which stems from the core principles of Random Forests. However, two key distinguishing factors exist between these methods:

- Firstly, every base learner within the Extra Trees algorithm is trained employing the entirety of the available training dataset.

- Secondly, the split points for these base learners are chosen at random, selected from the empirical feature ranges identified within the training dataset.

In this research, the implementation of the ExtraTreesRegressor from the SciKit-Learn library is leveraged (Geurts et al., 2006). This algorithmic selection was made with the objective of deeply investigating and assessing the performance capabilities that this method potentially offers.

3.2. Model diagnostics

The following section will scrutinize the empirical performance results obtained with the proposed model, utilizing the dataset gathered prior. The analysis commences with a broad evaluation of the model's performance in the task of price prediction. This is succeeded by a presentation of the model's performance on a country-specific level, coupled with a detailed analysis of performance of different $F(X)$ functions.

Subsequently, a comprehensive comparison is undertaken, focusing on regional variations and market segments. To assess the contribution of individual features to the overall performance of the model, an explanatory variable removal procedure is undertaken.

The section culminates with an analysis of the residuals' distribution. This enables a comprehension of the model's strengths and weaknesses, thereby providing insights to potentially refine the model for future implementations.

All diagnostics have been conducted with the use of a holdout sample derived for each country from 10% randomly selected dataset observations. Different models have been compared with use of accuracy metrics (MAPE) as recommended in he CRC Guide to Automated Valuation Model (AVM) Performance Testing (Consortium, 2003).

3.2.1. Model performance

MAPE and R2 metrics have been selected to evaluate pricing model performance. These metrics are universally employed in AVM, CAMA and property pricing models evaluation. National and model-level breakdown is presented below:

Table 3.1: MAPE - breakdown by countries

Model	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
Country			
Chile	32.69	26.24	27.88
Colombia	39.22	29.58	29.96
Mexico	92.13	67.51	71.06
Netherlands	17.80	17.68	17.53
Poland	30.44	26.10	26.64

Reported accuracy scores are lower than ones achieved by high-performance, commercial AVMs in the United States, but they are in line with international levels of models accuracy (Matysiak, 2017). BaselineModel metrics provide insights into performance of a pricing model solely based on prices of properties in the vicinity of the priced house. The MAPE difference between BaselineModel and ExtraTreesRegressor/RandomForestRegressor (ET/RF) allows for an insightful peak into level of additional localization features overall performance contributions. The decrease is largest when BaselineModel MAPE is largest. The performance contribution is visible consistently across all countries under study evaluation:

- Colombia: **-9.64** ET MAPE / **-9.26** RF MAPE
- Chile: **-6.45** ET MAPE / **-4.81** RF MAPE
- Mexico: **-24.62** ET MAPE / **-21.07** RF MAPE
- Netherlands: **-0.12** ET MAPE / **-0.27** RF MAPE
- Poland: **-4.34** ET MAPE / **-3.8** RF MAPE

Table 3.2: R2 - breakdown by countries

Model	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
Country			
Chile	0.165040	0.535352	0.522542
Colombia	0.139418	0.371194	0.414225
Mexico	0.013072	0.157298	0.171261
Netherlands	0.284025	0.307087	0.322227
Poland	0.175984	0.284821	0.306110

Observed R2 metrics for models are consistently larger than BaselineModel R2 metrics in all countries:

- Colombia: **+0.23** ET R2 / **+0.27** RF R2
- Chile: **+0.37** ET R2 / **+0.36** RF R2
- Mexico: **+0.14** ET R2 / **+0.16** RF R2
- Netherlands: **+0.02** ET R2 / **+0.04** RF R2
- Poland: **+0.11** ET R2 / **+0.13** RF R2

3.2.2. Performance in major metropolitan areas

An evaluation has been conducted to investigate model performance in the proximity of major metropolitan areas and the remaining parts of countries included in the study. The metropolitan areas division approach has been originally employed in the hotels valuation study (O'Neill, 2004). Areas in the vicinity[1] of the city centers has been included in the *cities_ test* sample. Observations not included in the *cities_ test* form *non_ cities_ test* sample.

MAPE metrics observed are smaller in the *cities_ test* than in *non_ cities_ test* in **4/5 countries** except for Mexico.

Table 3.3: Model performance in Colombia major metropolitan areas [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
cities_test	37.84	27.91	28.27
non_cities_test	40.79	31.47	31.88

- Colombia: **-3.56** ET MAPE in *cities_ test*[2]
- Colombia: **-3.61** RF MAPE in *cities_ test*

[1] Vicinity is defined as a longitudinal and latitudinal decimal degree distance smaller than 0.135 which corresponds to the rectangle area with sides of size 26-30 km centered at city center coordinates. The exact size depends on the city position on the globe.

[2] Bogotá, Medellín, Cali, Barranquilla, Cartagena, Cúcuta, Bucaramanga

Table 3.4: Model performance in Chile major metropolitan areas [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
cities_test	30.26	21.56	22.80
non_cities_test	35.14	30.93	32.99

- Chile: **-9.37** ET MAPE in *cities_test*[3]
- Chile: **-10.19** RF MAPE in *cities_test*

Table 3.5: Model performance in Mexico major metropolitan areas [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
cities_test	130.21	89.58	94.22
non_cities_test	77.84	59.23	62.37

- Mexico: **+30.35** ET MAPE in *cities_test*[4]
- Mexico: **+31.85** RF MAPE in *cities_test*

Table 3.6: Model performance in Netherlands major metropolitan areas [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
cities_test	16.92	15.95	16.44
non_cities_test	18.04	18.15	17.82

- Netherlands: **-2.2** ET MAPE in *cities_test*[5]
- Netherlands: **-1.38** RF MAPE in *cities_test*

[3] Santiago, Antofagasta, Temuco, Iquique, Concepción

[4] Mexico City, Guadalajara, Ciudad Juárez, Tijuana, Gustavo A. Madero

[5] Amsterdam, Rotterdam, Utrecht, Eindhoven, Groningen, Breda, Nijmegen, Enschede

Table 3.7: Model performance in Poland major metropolitan areas [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
cities_test	27.15	19.22	20.13
non_cities_test	31.83	29.00	29.38

- Poland: **-3.56** ET MAPE in *cities_test*[6]
- Poland: **-3.61** RF MAPE in *cities_test*

3.2.3. Performance in market segments

Model diagnostics has been expanded to cover market segments partitioned by house sizes. The following segments have been selected:

- $0m^2 - 100m^2$
- $100m^2 - 200m^2$
- $200m^2 - 300m^2$
- $300m^2 - 500m^2$

The metrics evaluation did not reveal any pattern in model performance evolution along with property size increases/decreases across countries. Difference between biggest and smallest MAPE across sectors can act as a consistency measure for model performance in different prediction settings.

Table 3.8: Model performance across Colombian housing market segments [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
market_segment_0_100	29.90	25.41	25.84
market_segment_100_200	40.88	28.84	29.50
market_segment_200_300	40.25	29.09	29.57
market_segment_300_500	43.01	35.35	35.04

- Colombia: **+9.94** ET MAPE min-max range
- Colombia: **+9.2** RF MAPE min-max range

[6] Poznań, Zielona Góra, Szczecin, Rzeszów, Cracow, Olsztyn, Bydgoszcz, Poznań, Wrocław, Opole, Łódź, Lublin, Warsaw, Katowice, Gdańsk

Table 3.9: Model performance across Chilean housing market segments [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
market_segment_0_100	32.29	26.43	27.92
market_segment_100_200	36.15	27.31	29.28
market_segment_200_300	32.50	23.53	26.37
market_segment_300_500	25.50	26.46	26.29

- Chile: **+3.78** ET MAPE min-max range
- Chile: **+2.99** RF MAPE min-max range

Table 3.10: Model performance across Mexican housing market segments [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
market_segment_0_100	65.43	50.07	52.30
market_segment_100_200	83.43	55.92	60.34
market_segment_200_300	125.71	92.50	96.53
market_segment_300_500	108.92	90.46	92.91

- Mexico: **+42.43** ET MAPE min-max range
- Mexico: **+44.23** RF MAPE min-max range

Table 3.11: Model performance across Netherlands housing market segments [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
market_segment_0_100	17.04	15.43	15.23
market_segment_100_200	17.87	18.14	17.98
market_segment_200_300	18.12	21.20	20.38
market_segment_300_500	28.97	11.27	18.42

- Netherlands: **+9.93** ET MAPE min-max range
- Netherlands: **+5.15** RF MAPE min-max range

Table 3.12: Model performance across Polish housing market segments [MAPE]

Model Test scheme	BaselineModel	ExtraTreesRegressor	RandomForestRegressor
market_segment_0_100	22.73	21.36	21.64
market_segment_100_200	29.13	25.50	25.97
market_segment_200_300	41.00	31.94	33.06
market_segment_300_500	51.46	38.59	39.50

- Poland: **+17.23** ET MAPE min-max range
- Poland: **+17.86** RF MAPE min-max range

In all housing markets under examination except for Chile the performance consistency measure is large relative to the overall MAPE metrics. This evidence indicates that for a practical application, e.g. mass property appraisal, more specialized models would be required. The findings are consistent with recommendations presented in the IAAO Standard on Mass Appraisal of Real Property which signifies the importance of homogeneity of the appraised market segments.

3.2.4. Features performance contribution

Individual localization features used in the model as explanatory variables may contribute to the model performance in the varying degree. This contribution has been examined in a series of feature teardown tests. The procedure compares results achieved with model with all variables to a model with individual feature removed. The procedure is repeated for each localization variable. In each procedure run only single variable is removed from the original model. Performance decrease after variable removal indicates that a features is contributing positively to the model performance. Near-zero changes or performance increases indicate that variable inclusion in the model does not contribute positively to the model performance and robustness.

Table 3.13: Model feature exclusion performance teardown - Colombia [MAPE]

Model Ablation scheme	ExtraTreesRegressor	RandomForestRegressor
all_columns	29.58	29.96
parks	31.34	31.24
restaurants	31.40	31.68
schools	31.56	31.35
transport	30.20	30.39
universities	30.12	30.61

- Colombia: **+1.76** ET MAPE change without *parks* variable
- Colombia: **+1.28** RF MAPE change without *parks* variable
- Colombia: **+1.82** ET MAPE change without *restaurants* variable
- Colombia: **+1.72** RF MAPE change without *restaurants* variable
- Colombia: **+1.98** ET MAPE change without *schools* variable
- Colombia: **+1.39** RF MAPE change without *schools* variable
- Colombia: **+0.62** ET MAPE change without *transport* variable
- Colombia: **+0.43** RF MAPE change without *transport* variable
- Colombia: **+0.54** ET MAPE change without *universities* variable
- Colombia: **+0.65** RF MAPE change without *universities* variable

In the Colombian housing market case all explanatory variables are contributing positively to the overall model performance.

Table 3.14: Model feature exclusion performance teardown - Chile [MAPE]

Model Ablation scheme	ExtraTreesRegressor	RandomForestRegressor
all_columns	26.24	27.88
parks	28.23	29.69
restaurants	27.41	28.34
schools	28.14	29.08
transport	27.36	28.99
universities	26.40	28.01

- Chile: **+1.99** ET MAPE change without *parks* variable
- Chile: **+1.81** RF MAPE change without *parks* variable
- Chile: **+1.17** ET MAPE change without *restaurants* variable
- Chile: **+0.46** RF MAPE change without *restaurants* variable
- Chile: **+1.9** ET MAPE change without *schools* variable
- Chile: **+1.2** RF MAPE change without *schools* variable
- Chile: **+1.12** ET MAPE change without *transport* variable
- Chile: **+1.11** RF MAPE change without *transport* variable
- Chile: **+0.16** ET MAPE change without *universities* variable
- Chile: **+0.13** RF MAPE change without *universities* variable

In the Chilean housing market case all explanatory variables are contributing positively to the overall model performance.

Table 3.15: Model feature exclusion performance teardown - Mexico [MAPE]

Model Ablation scheme	ExtraTreesRegressor	RandomForestRegressor
all_columns	67.51	71.06
parks	73.75	75.75
restaurants	69.16	73.99
schools	72.52	76.11
transport	70.90	73.10
universities	67.71	71.84

- Mexico: **+6.24** ET MAPE change without *parks* variable
- Mexico: **+4.69** RF MAPE change without *parks* variable
- Mexico: **+1.65** ET MAPE change without *restaurants* variable
- Mexico: **+2.93** RF MAPE change without *restaurants* variable
- Mexico: **+5.01** ET MAPE change without *schools* variable
- Mexico: **+5.05** RF MAPE change without *schools* variable
- Mexico: **+3.39** ET MAPE change without *transport* variable
- Mexico: **+2.04** RF MAPE change without *transport* variable
- Mexico: **+0.2** ET MAPE change without *universities* variable
- Mexico: **+0.78** RF MAPE change without *universities* variable

In the Mexican housing market case all explanatory variables are contributing positively to the overall model performance.

Table 3.16: Model feature exclusion performance teardown - Netherlands [MAPE]

Model Ablation scheme	ExtraTreesRegressor	RandomForestRegressor
all_columns	17.68	17.53
parks	17.59	17.32
restaurants	17.77	17.65
schools	18.85	17.95
transport	18.08	17.95
universities	17.68	17.70

- Netherlands: **-0.09** ET MAPE change without *parks* variable
- Netherlands: **-0.21** RF MAPE change without *parks* variable
- Netherlands: **+0.09** ET MAPE change without *restaurants* variable
- Netherlands: **+0.12** RF MAPE change without *restaurants* variable
- Netherlands: **+1.17** ET MAPE change without *schools* variable
- Netherlands: **+0.42** RF MAPE change without *schools* variable
- Netherlands: **+0.4** ET MAPE change without *transport* variable
- Netherlands: **+0.42** RF MAPE change without *transport* variable
- Netherlands: **-0.0** ET MAPE change without *universities* variable
- Netherlands: **+0.17** RF MAPE change without *universities* variable

Features performance contribution is smaller in Netherlands compared to other countries included in the study. Inclusion of *parks* and *universities* variables yields performance decreases or only a minor increase.

Table 3.17: Model feature exclusion performance teardown - Poland [MAPE]

Model Ablation scheme	ExtraTreesRegressor	RandomForestRegressor
all_columns	26.10	26.64
parks	26.67	27.17
restaurants	26.69	27.02
schools	26.48	26.80
transport	27.66	27.83
universities	26.33	26.68

- Poland: **+0.57** ET MAPE change without *parks* variable
- Poland: **+0.53** RF MAPE change without *parks* variable
- Poland: **+0.59** ET MAPE change without *restaurants* variable
- Poland: **+0.38** RF MAPE change without *restaurants* variable
- Poland: **+0.38** ET MAPE change without *schools* variable
- Poland: **+0.16** RF MAPE change without *schools* variable
- Poland: **+1.56** ET MAPE change without *transport* variable
- Poland: **+1.19** RF MAPE change without *transport* variable
- Poland: **+0.23** ET MAPE change without *universities* variable
- Poland: **+0.04** RF MAPE change without *universities* variable

In the Polish housing market case all explanatory variables are contributing positively to the overall model performance.

3.2.5. Detailed performance

Detailed performance evaluation reveals further insights into overall model performance in the out-of-sample predictive tasks. By examining distribution of APE scores and prediction residuals a more complete view on the model performance can be formed.

Table 3.18: ExtraTreesRegressor detailed performance metrics

Country	Mean residual	1Q residuals	3Q residuals	1Q APE	2Q APE	3Q APE
Colombia	22114.27	-548625.58	755770.20	8.45	20.80	39.63
Mexico	-149.88	-2919.01	3448.23	8.25	30.72	66.14
Chile	24542.68	-231733.98	306474.61	3.97	12.31	31.56
Netherlands	37.49	-455.91	548.69	7.02	15.09	23.95
Poland	18.29	-755.56	940.81	2.94	14.92	36.27

Table 3.19: RandomForestRegressor detailed performance metrics

Country	Mean residual	1Q residuals	3Q residuals	1Q APE	2Q APE	3Q APE
Colombia	19541.68	-576428.09	744842.77	10.15	21.17	38.59
Mexico	3.22	-3187.19	4006.18	11.56	32.54	71.27
Chile	35556.41	-242062.18	358573.02	5.39	14.35	32.53
Netherlands	36.98	-470.92	558.88	8.42	13.51	23.73
Poland	37.83	-815.00	1020.17	4.75	16.07	35.60

Close examination of the above tables reveals that:

- Median APE is smaller than 25 in **4/5** countries: It means that 50% of houses were priced with lower or equal prediction error in 4/5 countries. Such results can be viewed as a high model performance.
- MAPE > Median APE: In all countries long tail of larger prediction errors contribute to the inequality. In a general case of an average property performance is higher than MAPE indicate.
- Mean residuals are relatively close to zero: examined models don not present tendencies to undervalue or overvalue houses.

3.2.6. Residuals distribution - Chile

Presented and discussed below are histograms and scatterplots for residuals of Chilean house pricing models. Results for other countries are in line with Chilean results. They are reported in the Appendix A.

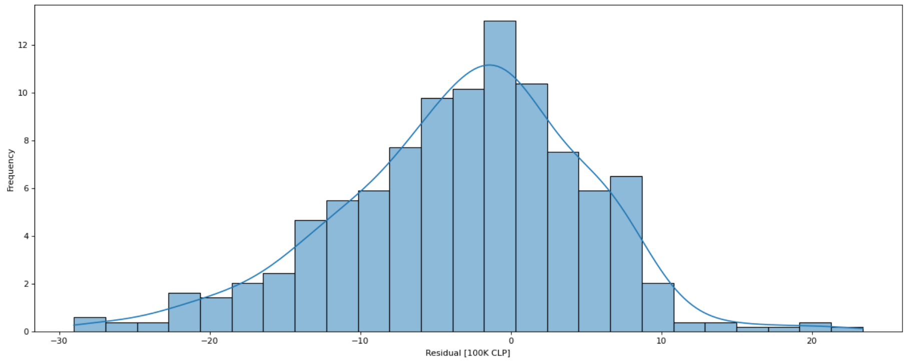


Figure 3.4: Chile - BaselineModel residuals distribution

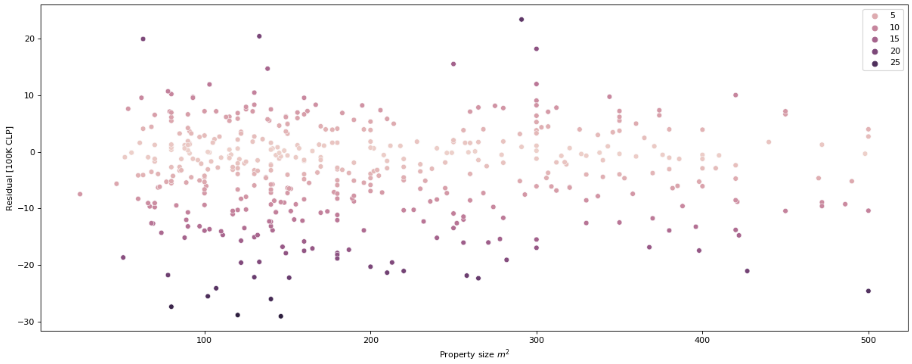


Figure 3.5: Chile - BaselineModel residuals scatterplot

Application of price grid based model yields near-zero-centered distribution without very long tails.

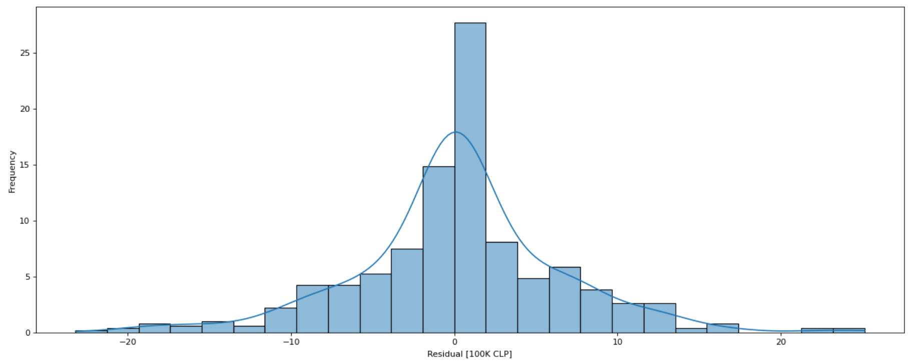


Figure 3.6: Chile - ExtraTreesRegressor residuals distribution

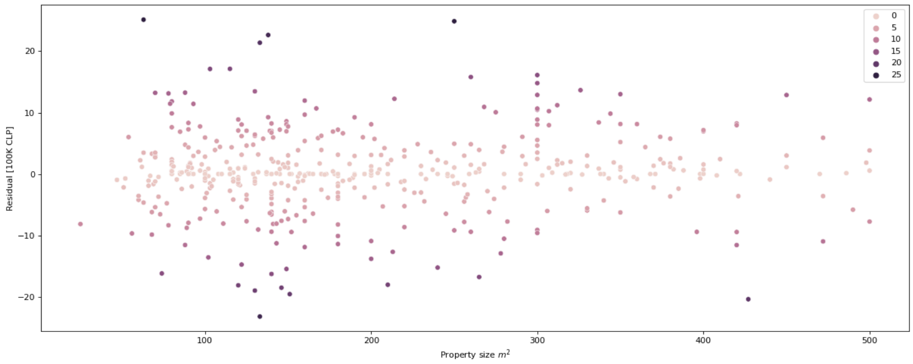


Figure 3.7: Chile - ExtraTreesRegressor residuals scatterplot

Application of the ExtraTreesRegressor based model yields distribution which is more symmetrical and which clearly manifests higher model performance compared to the BaselineModel.

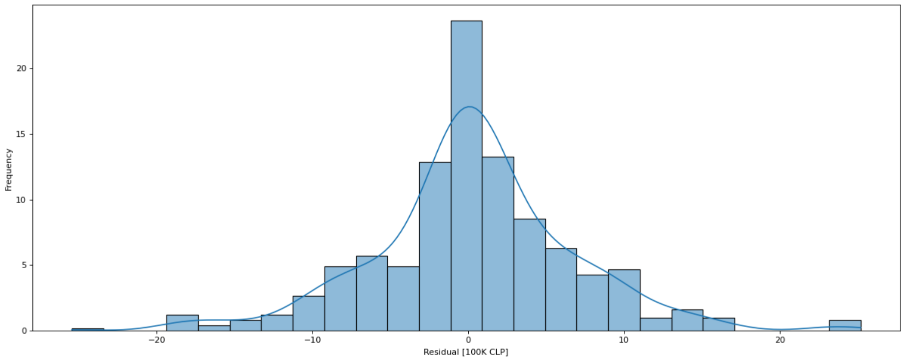


Figure 3.8: Chile - RandomForestRegressor residuals distribution

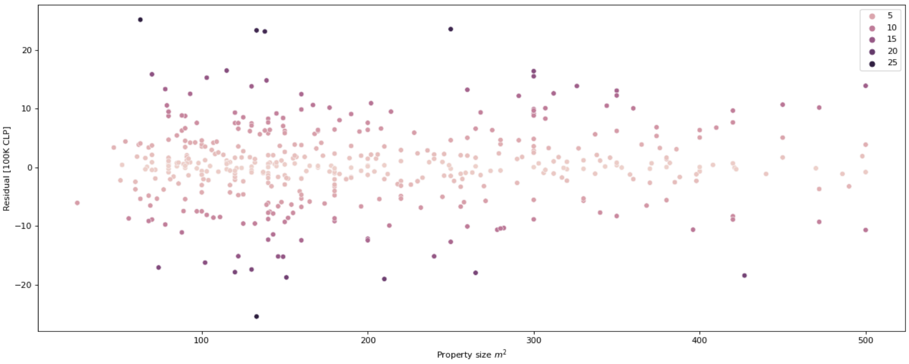


Figure 3.9: Chile - RandomForestRegressor residuals scatterplot

Application of the RandomForestRegressor based model yields distribution which is more symmetrical and which clearly manifests higher model performance compared to the BaselineModel. The results are consistent with ExtraTreesRegressor results.

3.3. Conclusions

Proposed pricing model has been successfully applied to analysis in multiple countries and achieved high prediction performance in all markets except for Mexican housing market. As evidenced by model diagnostics to improve model performance in Mexican market a greater degree of segmentation would be required.

High prediction accuracy contributes valuable evidence supporting hypothesis 1. Absence of all hedonic features but property size did not impede the attempt to develop robust and accurate property pricing model.

The study contributes significant evidence supporting hypotheses 2. (a-d). In almost all countries and for all examined variables results conclusively support the claim that inclusion of information about nearby amenities improves predictive performance of pricing models. The only exceptions are *parks* and *universities* variables in Dutch housing market where achieved performance is very high. In that case variables inclusion decreased performance only slightly or had no effect.

The dataset and pricing model can be expanded towards inclusion of hedonic features. These might contribute further towards dataset usefulness and model performance.

Chapter 4

Summary

4.1. Dataset

The dataset of real estate listings from 5 OECD countries, compiled in the study, has proven to be an instrumental resource, offering a wealth of information on the dynamics of housing markets. This expansive data source has not only helped analyze broad trends but has also delved into finer details, providing a holistic view of the real estate scenario in these countries.

The analysis of the dataset has provided critical insights into the spatial distribution of housing markets, highlighting both regional and local differences in terms of market activity and price levels. This has shed light on the patterns of distribution and varying scales of market activity, helping to understand the disparity in real estate prices across different locations.

Additionally, the study also provides a deep dive into the distributions of house prices and sizes, giving a clear perspective on housing affordability and spatial constraints. It aids in drawing comparisons between various house sizes and their corresponding prices, painting a comprehensive picture of the housing landscape across the countries.

This dataset could serve as a valuable resource for researchers from a broad array of fields including economics, geography, urban planning, and data science. It could provide them with the necessary tools to derive valuable insights, allowing them to conduct detailed studies, extrapolate trends, and make meaningful contributions to their respective fields.

Efforts by the author are in progress to further enhance the dataset by incorporating hedonic features of properties. This will involve looking at factors beyond size and location, such as the physical condition of the property, the presence of amenities, and unique characteristics that might contribute to its value.

The geographical coverage of the dataset is also set to expand, broadening the scope and applicability of the research. This would increase the depth and breadth of the dataset, making it more versatile and comprehensive, thereby strengthening its relevance and applicability.

Furthermore, the author plans to share this dataset for scientific purposes with researchers from all aforementioned domains. This decision reflects a commitment to promoting open access to data, fostering collaborative research, and supporting the advancement of knowledge in these critical fields.

4.2. Model

The study's modeling results provide substantial evidence about the critical role of spatial information in real estate asset valuation. It emphasizes that the geographical context of properties, including their location and surrounding infrastructure, significantly influences their value.

The findings of this analysis align seamlessly with other reported research conducted in the field of real estate. This consistency with existing literature further bolsters the credibility of the study and the validity of its conclusions.

Interestingly, the model has demonstrated high performance even in the absence of traditionally included hedonic features. This shows the robustness of the model, and suggests that it is capable of accurately assessing property value even without these typically considered characteristics.

The study has confirmed that certain features that play a pivotal role in improving the accuracy of real estate pricing models. These include factors that contribute to the quality of life and accessibility in a given area:

- number of restaurants in the neighborhood
- number of schools in the neighborhood
- number of parks in the neighborhood
- number of tram, bus and subway stops in the neighborhood
- number of universities in the neighborhood

Bibliography

Angrick, S., Bals, B., Hastrich, N., Kleissl, M., Schmidt, J., Doskoč, V., Molitor, L., Friedrich, T., & Katzmann, M. (2022). Towards explainable real estate valuation via evolutionary algorithms. *arXiv.* https://arxiv.org/pdf/2110.05116.pdf

Bergadano, F., Bertilone, R., Paolotti, D., & Ruffo, G. (2019). Learning real estate automated valuation models from heterogeneous data sources. *arXiv.* https://arxiv.org/pdf/1909.00704.pdf

Bidanset, P. E., Lombard, J. R., Davis, P., McCord, M., & McCluskey, W. J. (2017). Further evaluating the impact of kernel and bandwidth specifications of geographically weighted regression on the equity and uniformity of mass appraisal models. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Breiman, L. (2001). Random forests. *Machine Learning 45.* https://doi.org/10.1023/A:1010933404324

Ciuna, M., Salvo, F., & Simonotti, M. (2017). An estimative model of automated valuation method in italy. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Coleman, W., Johann, B., Pasternak, N., Vellayan, J., Foutz, N., & Shakeri, H. (2022). Using machine learning to evaluate real estate prices using location big data. *arXiv.* https://arxiv.org/pdf/2205.01180.pdf

Consortium, C. R. M. (2003). The crc guide to automated valuation model (avm) performance testing. https://professional.sauder.ubc.ca/re_creditprogram/course_resources/courses/content/344/CRC_AVM.pdf

Conway, J. (2018). Artificial intelligence and machine learning: Current applications in real estate. *MIT.*

CoreLogic. (2011). Automated valuation model testing. https://www.corelogic.com/wp-content/uploads/sites/4/downloadable-docs/automated-valuation-model-testing.pdf

Curto, R., Fregonara, E., & Semeraro, P. (2017). A spatial analysis for the real estate market applications. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

d'Amato, M. (2017a). A brief outline of avm models and standards evolutions. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

d'Amato, M. (2017b). Location value response surface model as automated valuation methodology a case in bari. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

d'Amato, M., Siniak, N., & Amoruso, P. (2017). Dealing with spatial modelling in minsk. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Dornfest, A., Marchand, B., Warr, D., Dettbarn, A., Forde, W., Neihardt, J. M. C., & O'Connor, P. M. (2018). Standard on automated valuation models (avms). https://www.iaao.org/media/standards/Standard_on_Automated_Valuation_Models.pdf

Dornfest, A., Warr, D., Gloudemans, R., Prestridge, M., Reavey, M., Deegear, D., & Bennett, C. (2002). Standard on mass appraisal of real property. *International Association of Assessing Officers*. https://www.livingstoncounty-il.org/wordpress/wp-content/uploads/2015/02/PR-Ex.-184-IAAO-Standard-On-Mass-Appraisal.pdf

Dunning, R., Keskin, B., & Watkins, C. (2017). Using multi level modeling techniques as an avm tool: Isolating the effects of earthquake risk from other price determinants. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Eilers, F., & Kunert, A. (2017). Automated valuation models for the granting of mortgage loans in germany. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning 63*. https://doi.org/10.1007/s10994-006-6226-1

Giudice, V. D., & Paola, P. D. (2017). Spatial analysis of residential real estate rental market with geoadditive models. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

Glumac, B., & Rosiers, F. D. (2018). Real estate and land property automated valuation systems: A taxonomy and conceptual model. *SSRN Electronic Journal*.

Li, C.-C., Wang, W.-Y., Du, W.-W., & Peng, W.-C. (2022). Look around! a neighbor relation graph learning framework for real estate appraisal. *arXiv*. https://arxiv.org/pdf/2212.12190.pdf

Matysiak, G. A. (2017). Automated valuation models (avms): A brave new world? *Wroclaw Conference in Finance 2017*. https://www.researchgate.net/publication/319355261_Automated_Valuation_Models_AVMs_A_brave_new_world

Moosavi, V. (2017). Urban data streams and machine learning: A case of swiss real estate market. *arXiv*. https://arxiv.org/pdf/1704.04979.pdf

Mooya, M. M. (2017). Automated valuation models and economic theory. *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis, Springer.*

O'Neill, J. W. (2004). An automatedvaluation model for hotels. *Cornell Hotel and Restaurant Administration Quarterly*. https://www.academia.edu/85154776/An_Automated_Valuation_Model_for_Hotels

Peng, H., Li, J., Wang, Z., Yang, R., Liu, M., Zhang, M., Yu, P. S., & He, L. (2020). Lifelong property price prediction: A case study for the toronto real estate market. *arXiv*. https://arxiv.org/pdf/2008.05880.pdf

Robson, G., & Downie, M. L. (2008). Automated valuation models: An international perspective. https://core.ac.uk/download/4147458.pdf

Solovev, K., & Pröllochs, N. (2021). Integrating floor plans into hedonic models for rent price appraisal. *arXiv*. https://arxiv.org/pdf/2102.08162.pdf

Stumpe, E., Despotovic, M., Zhang, Z., & Zeppelzauer, M. (2022). Real estate attribute prediction from multiple visual modalities with missing data. *arXiv*. https://arxiv.org/pdf/2211.09018.pdf

Yazdani, M., & Raissi, M. (2023). Real estate property valuation using self-supervised vision transformers. *arXiv*. https://arxiv.org/pdf/2302.00117.pdf

List of Figures

List of Tables

Appendix A

Residuals distribution - all countries

A.1. Chile

A.1.1. BaselineModel

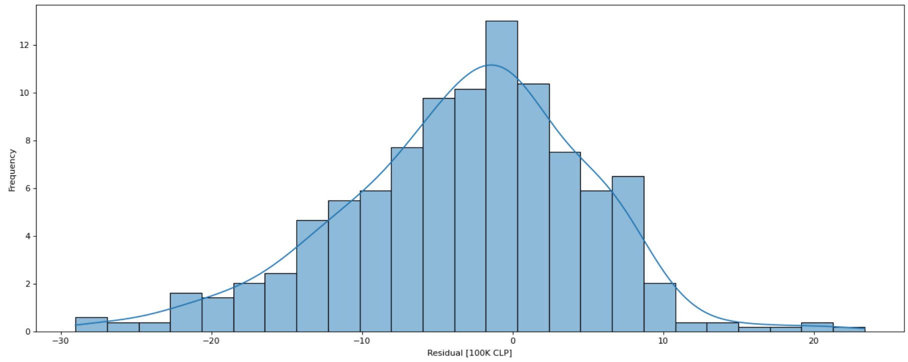


Figure A.1: Chile - BaselineModel residuals distribution

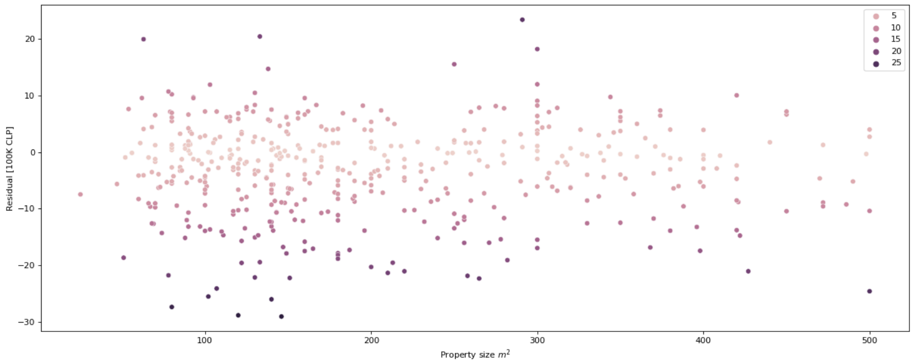


Figure A.2: Chile - BaselineModel residuals scatterplot

A.1.2. ExtraTreesRegressor

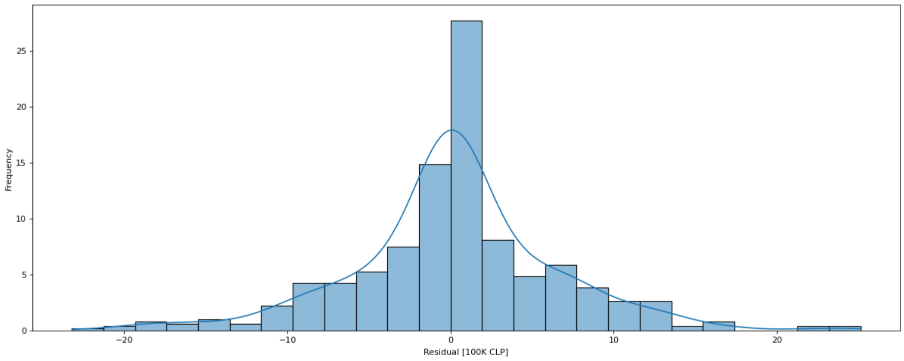


Figure A.3: Chile - ExtraTreesRegressor residuals distribution

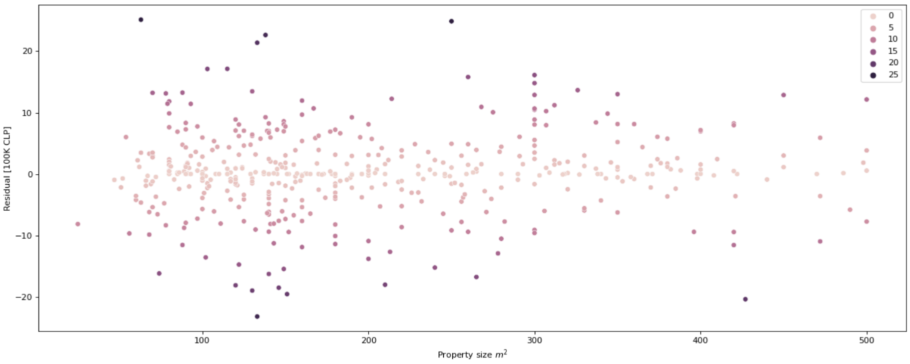


Figure A.4: Chile - ExtraTreesRegressor residuals scatterplot

A.1.3. RandomForestRegressor

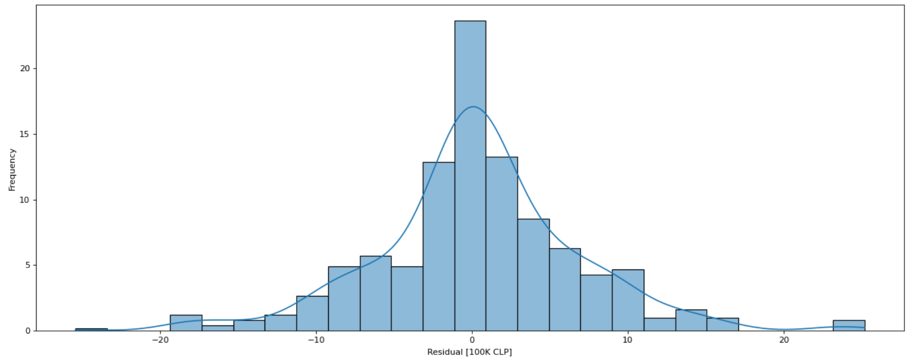


Figure A.5: Chile - RandomForestRegressor residuals distribution

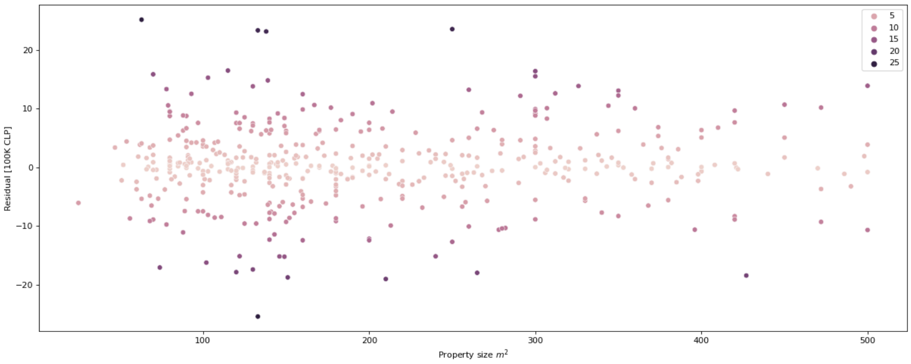


Figure A.6: Chile - RandomForestRegressor residuals scatterplot

A.2. Colombia

A.2.1. BaselineModel

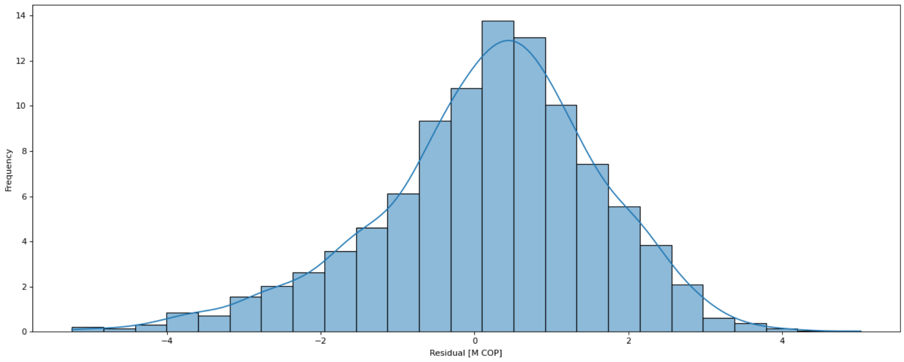


Figure A.7: Colombia - BaselineModel residuals distribution

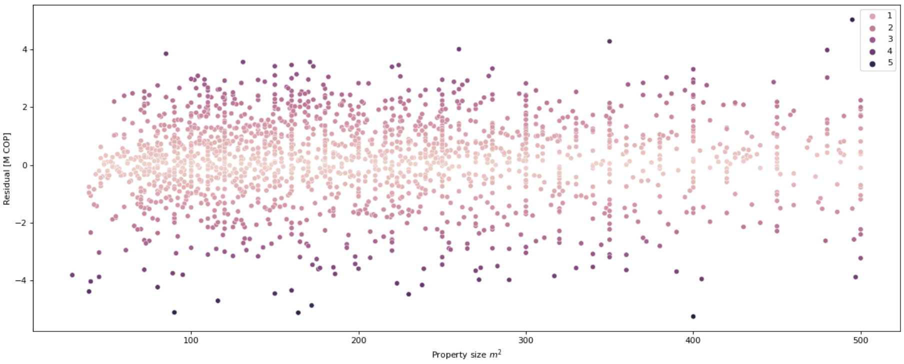


Figure A.8: Colombia - BaselineModel residuals scatterplot

A.2.2. ExtraTreesRegressor

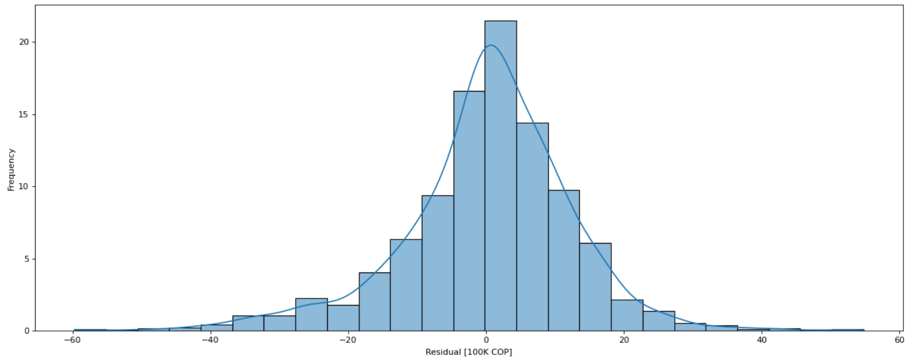


Figure A.9: Colombia - ExtraTreesRegressor residuals distribution

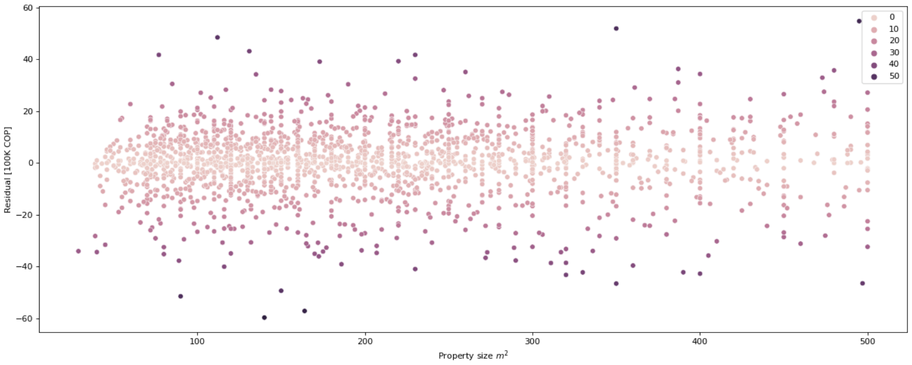


Figure A.10: Colombia - ExtraTreesRegressor residuals scatterplot

A.2.3. RandomForestRegressor

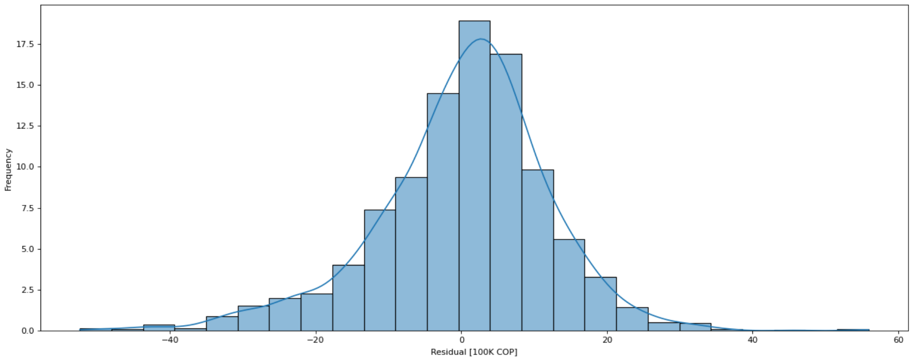


Figure A.11: Colombia - RandomForestRegressor residuals distribution

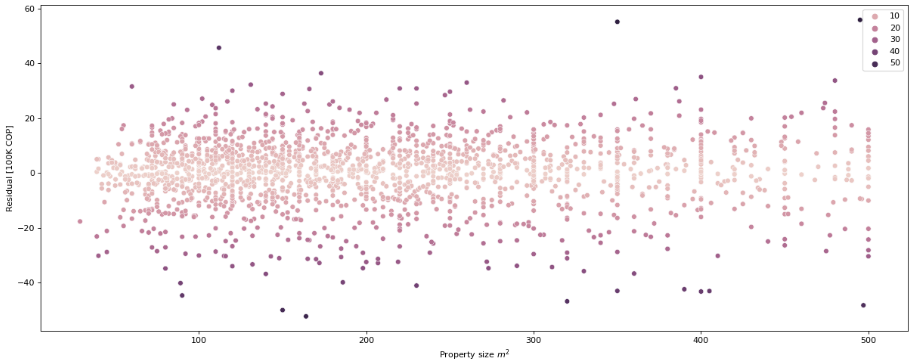


Figure A.12: Colombia - RandomForestRegressor residuals scatterplot

A.3. Mexico

A.3.1. BaselineModel

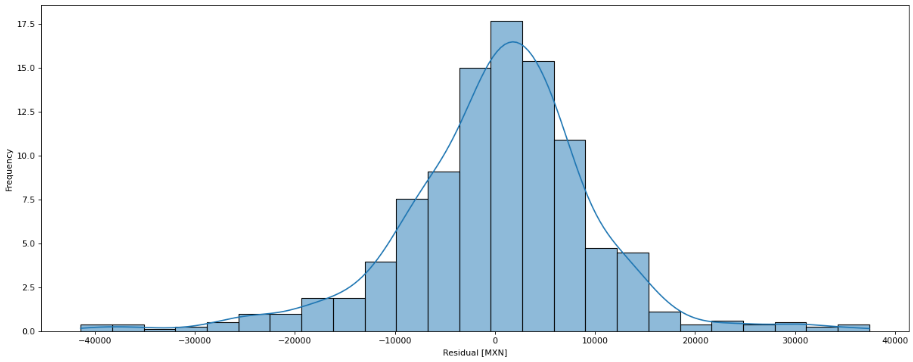


Figure A.13: Mexico - BaselineModel residuals distribution

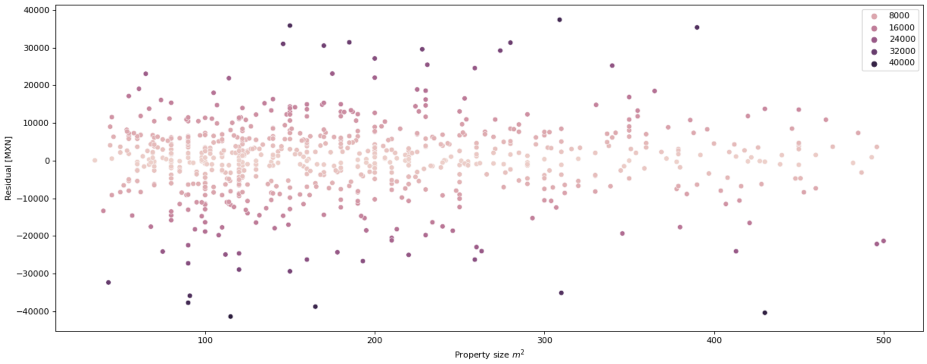


Figure A.14: Mexico - BaselineModel residuals scatterplot

A.3.2. ExtraTreesRegressor

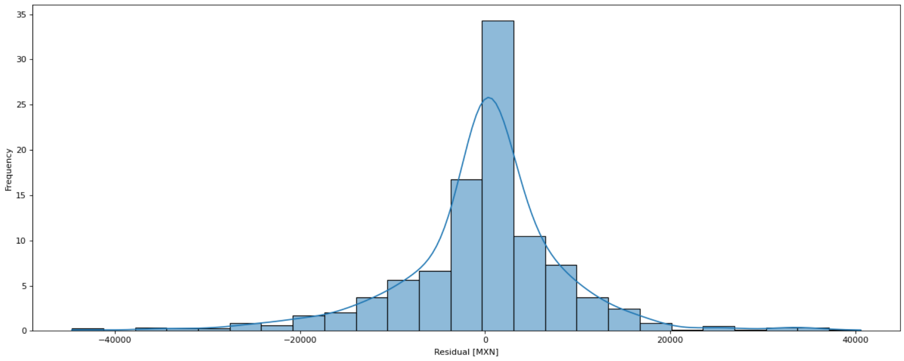


Figure A.15: Mexico - ExtraTreesRegressor residuals distribution

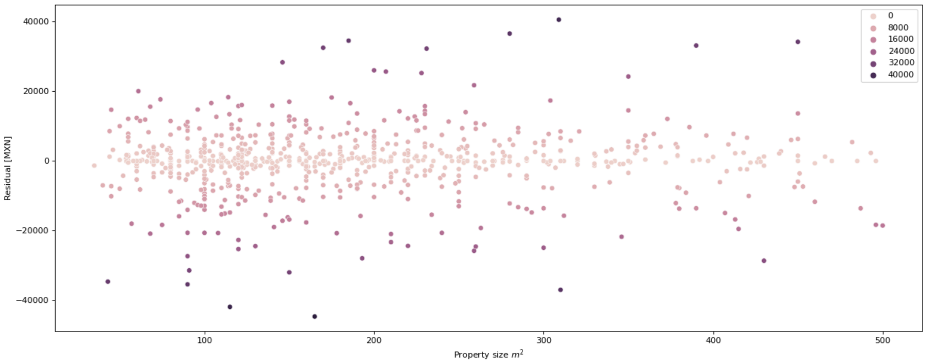


Figure A.16: Mexico - ExtraTreesRegressor residuals scatterplot

A.3.3. RandomForestRegressor

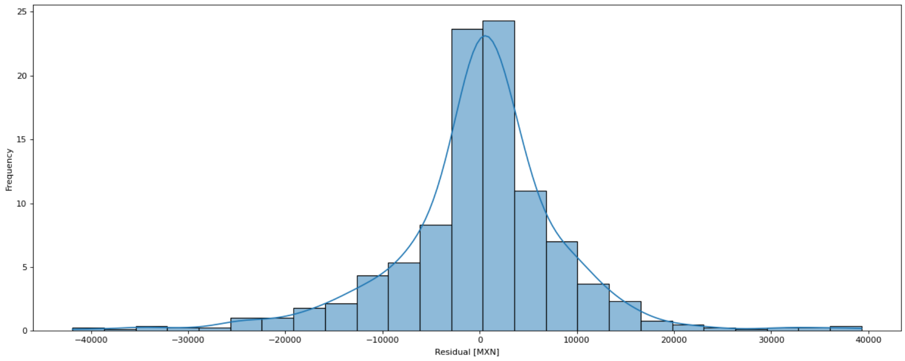


Figure A.17: Mexico - RandomForestRegressor residuals distribution

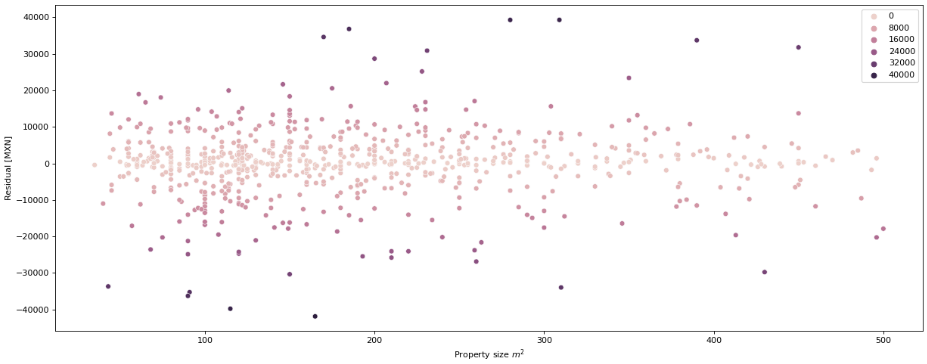


Figure A.18: Mexico - RandomForestRegressor residuals scatterplot

A.4. Netherlands

A.4.1. BaselineModel

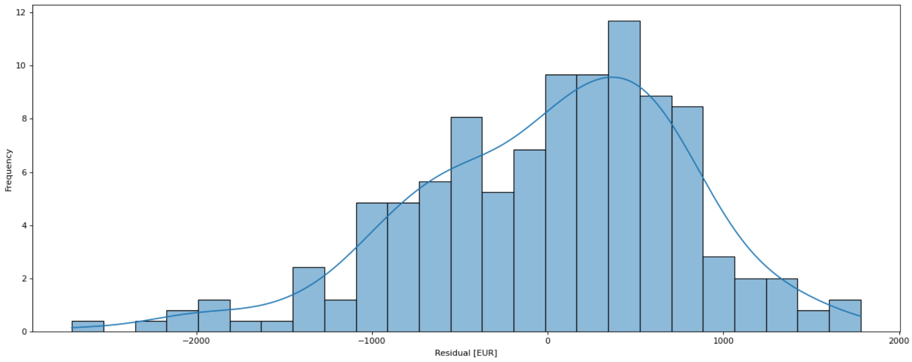


Figure A.19: Netherlands - BaselineModel residuals distribution

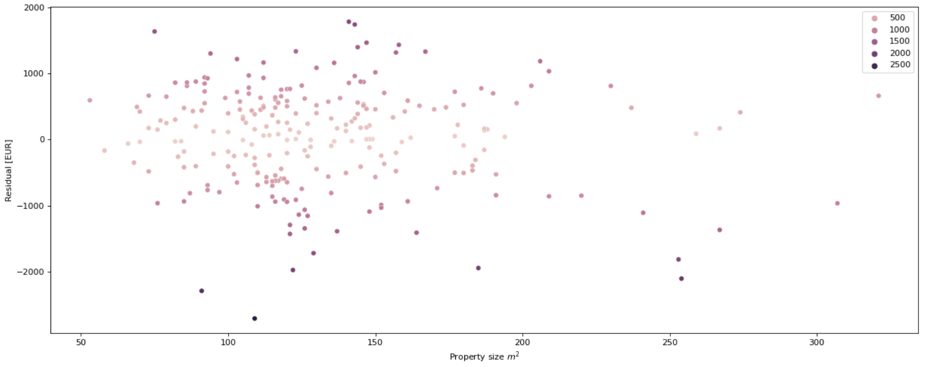


Figure A.20: Netherlands - BaselineModel residuals scatterplot

A.4.2. ExtraTreesRegressor

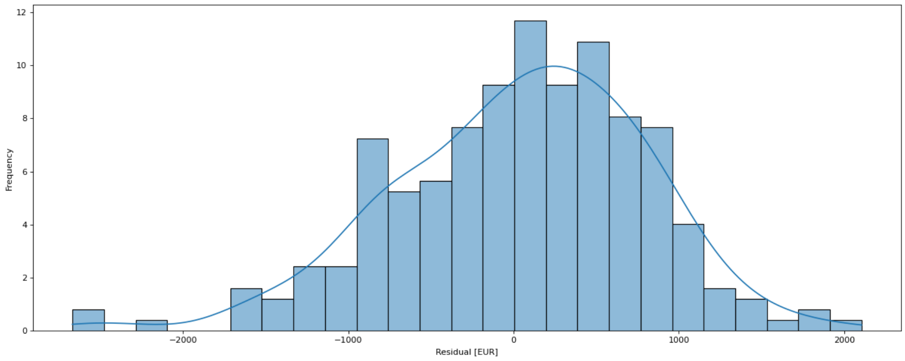


Figure A.21: Netherlands - ExtraTreesRegressor residuals distribution

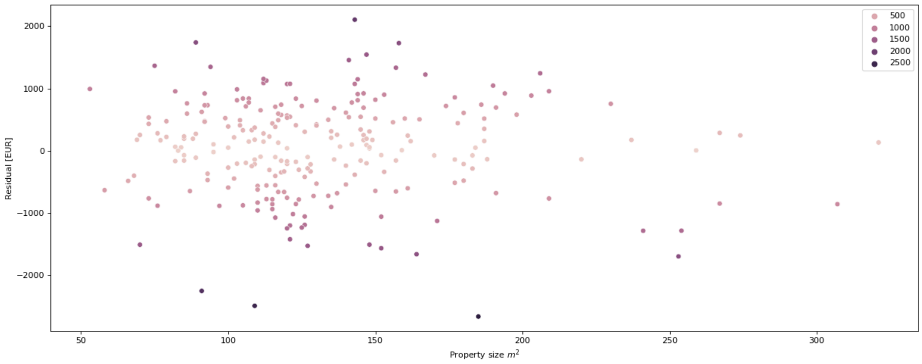


Figure A.22: Netherlands - ExtraTreesRegressor residuals scatterplot

A.4.3. RandomForestRegressor

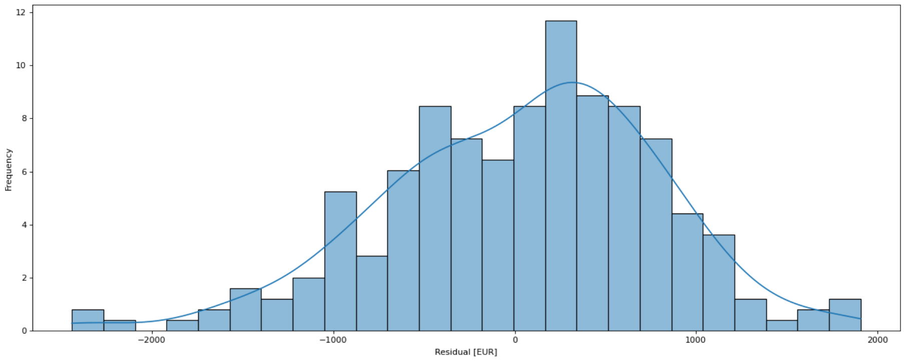


Figure A.23: Netherlands - RandomForestRegressor residuals distribution

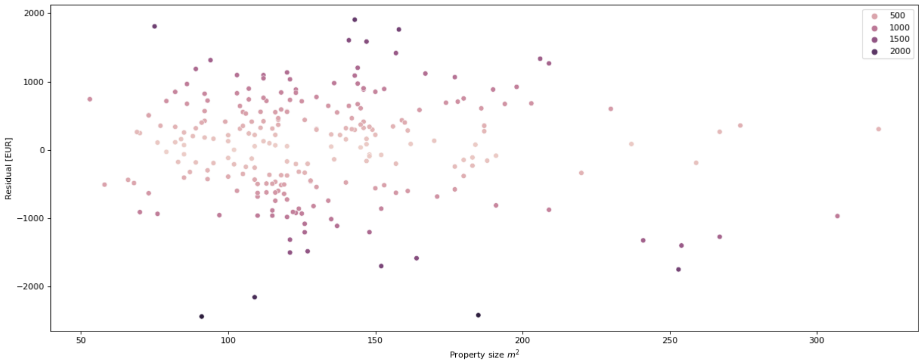


Figure A.24: Netherlands - RandomForestRegressor residuals scatterplot

A.5. Poland

A.5.1. BaselineModel

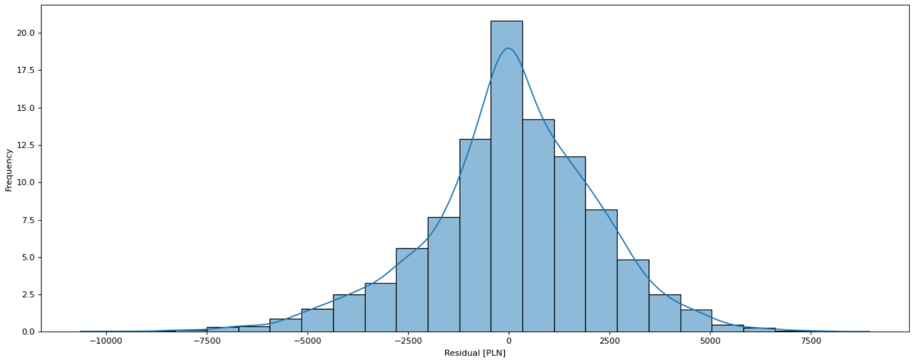


Figure A.25: Poland - BaselineModel residuals distribution

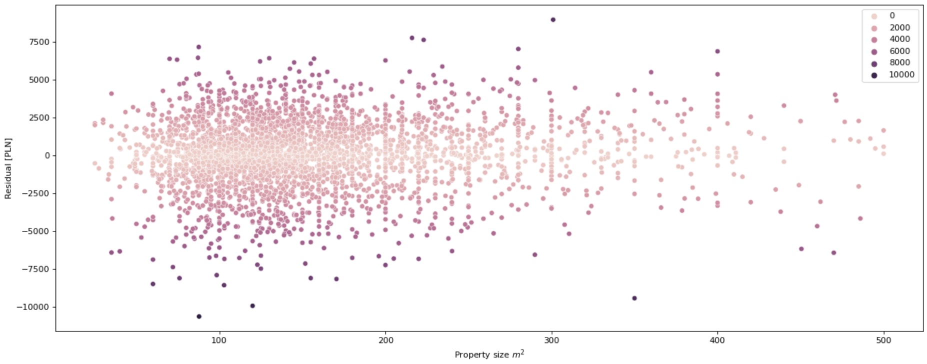


Figure A.26: Poland - BaselineModel residuals scatterplot

A.5.2. ExtraTreesRegressor

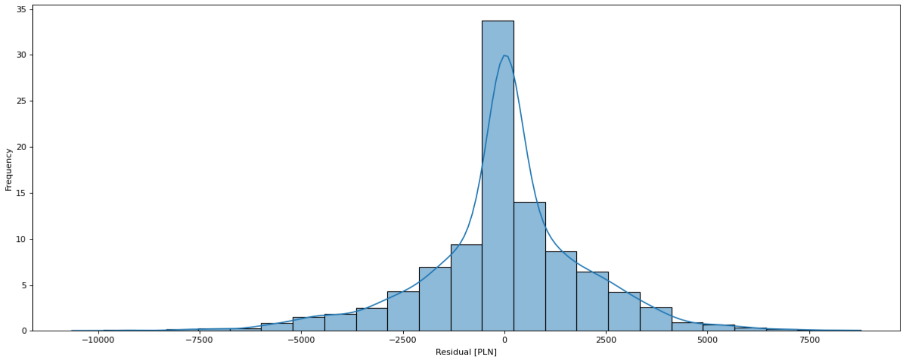


Figure A.27: Poland - ExtraTreesRegressor residuals distribution

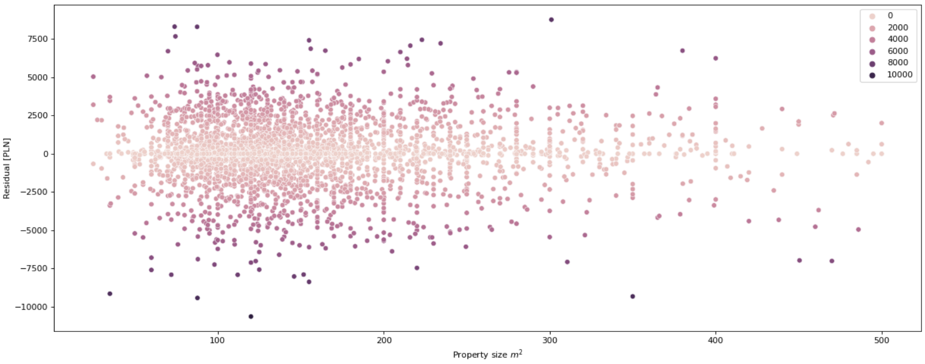


Figure A.28: Poland - ExtraTreesRegressor residuals scatterplot

A.5.3. RandomForestRegressor

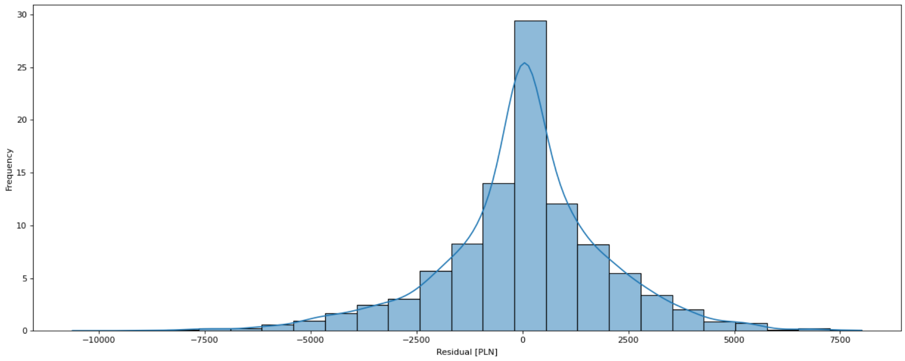


Figure A.29: Poland - RandomForestRegressor residuals distribution

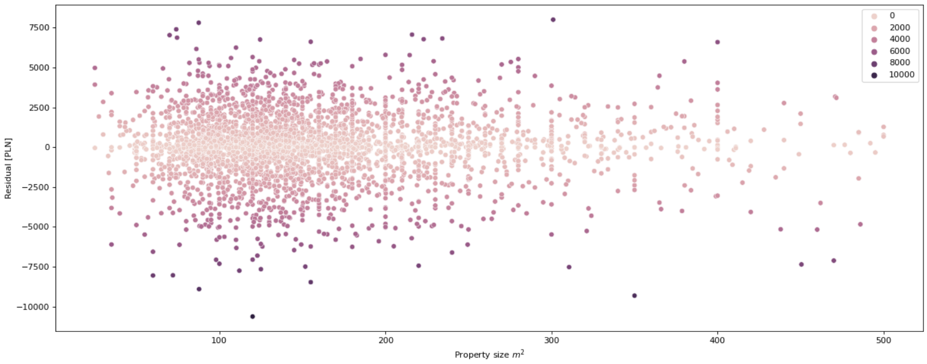


Figure A.30: Poland - RandomForestRegressor residuals scatterplot