Submitted to *Conference on Big Data, AI, & Fin Econ*

ML-Enhanced Arbitrage in Real Estate: Unveiling Intertemporal Investment Opportunities Under Asymmetric Information

LUCAS BALI
ESCP Business School

GONZALO FERNANDEZ DE CORDOBA
ESCP Business School

JAIME LUQUE
ESCP Business School

DAMIAN PIERRI
ESCP Business School

This paper investigates how big data and Machine Learning (ML) can optimize real estate investment strategies in online marketplaces by addressing information asymmetries. Using a large dataset of property listings in Madrid, we develop a model that assigns an opportunity index to each property, linking online traffic with investor information costs. The model ranks neighborhoods based on arbitrage opportunities arising from discrepancies between posted and demand prices. Employing machine learning, we find that unit size and neighborhood income explain two-thirds of price variability. Through dynamic pricing and active learning, the model guides institutional investors in maximizing returns while minimizing liquidity constraints. Our approach extends urban economics models by incorporating ML and data-driven decision-making, giving market value to information frictions in real estate finance.

KEYWORDS. machine learning, real estate investments opportunities, asymmetric information.

1. INTRODUCTION

In the evolving landscape of real estate investment, integrating economic theory with advanced big data and Machine Learning (ML) analytics offers new opportunities for identifying profitable arbitrage opportunities. As ML increasingly interacts with large, unstructured datasets and decision-making in financial markets, it is essential to explore how these technologies can enhance investment strategies in real estate. This study presents a novel approach to predicting the optimal selling price of residential

Lucas Bali: juan.lucas.bali@gmail.com
Gonzalo Fernandez de Cordoba: gfernandezdecordoba@escp.eu
Jaime Luque: jluque@escp.eu
Damian Pierri: dpierri@escp.eu

properties within a city, particularly where sellers face asymmetric information in online marketplaces. By leveraging 346,035 property listings from an intermediary platform in Madrid, Spain, we aim to uncover the "first-best" optimal investment strategy in the city's existing urban structure, thereby addressing key challenges at the intersection of big data, machine learning, and financial decision-making.

Our model assigns an opportunity index to each residential building and plot of land, reflecting the relationship between online traffic for a listing and the benefits for an investor to acquire market information. This approach directly tackles how ML interacts with data resources and human decision-makers, allowing investors to mitigate the effects of information frictions by purchasing and processing relevant data. This creates value through large-scale data analysis, which aligns with recent advances in ML technology that provide tools for analyzing unstructured financial data, such as real estate listings. Using this framework, we derive the optimal portfolio strategy for institutional investors seeking to invest in residential real estate. The model offers a comprehensive methodology for computing the spatial equilibrium of a city, ranking neighborhoods based on ML-enhanced data analysis, and generating arbitrage opportunities from unobserved demand and supply imbalances.

Our large dataset integrates observations from four sources, including real estate listings, demographic data, and geographical information. We derive the geographical distribution of real estate prices and per capita income across Madrid using state-of-the-art ML techniques, such as Permutational Feature Importance. This analysis aligns with the growing trend of employing ML and big data in investment decisions, particularly the rise of machines in asset management and high-dimensional data analysis. Our findings suggest that Madrid operates as a monocentric city, where unit size and neighborhood income explain two-thirds of the variability in real estate prices per square meter. By using ML tools to process large-scale datasets, we address both methodological challenges and real-world applications of ML in financial economics.

One of the key insights from our model is that the trade-off faced by real estate investors differs from the standard exploration-exploitation problem discussed in information theory. In traditional models, exploring the market might uncover better opportunities over time. However, our model, which applies dynamic pricing strategies enabled by ML, shows that the quicker an agent sells a property, the wider the gap between the posted price and the true demand price due to information asymmetries. This trade-off between price discrimination and liquidity emphasizes the role of ML in enabling investors to make data-driven decisions, a critical concern in both financial markets and corporate decision-making. The demand for information services, which can process and interpret vast amounts of market data using machine learning techniques, is growing as financial institutions seek to optimize investment portfolios and reduce risks associated with asymmetric information.

Our model is grounded in a dynamic pricing framework where supply and demand evolve. Sellers learn about demand through active experimentation, continuously adjusting posted prices based on observed outcomes. The recursive structure of the model allows us to estimate how investors can exploit arbitrage opportunities by using ML-enhanced data processing to identify neighborhoods where demand exceeds posted

prices. This contributes to the broader discourse on the role of ML in asset management, where trading and algorithmic strategies rely on big data to optimize financial decisions.

We segment the market using clusters created through unsupervised machine-learning techniques to further enhance the model's predictive accuracy. These clusters are ordered by price and liquidity preferences, forming a hierarchical "ladder" of real estate opportunities. The calibrated model is applied to analyze the investment decisions of institutional real estate investors, and the resulting strategy suggests a cascading approach: investing in the most attractive neighborhood first, selling all available units, and then moving on to the next best neighborhood. This systematic approach maximizes the investor's returns while minimizing liquidity constraints, a critical concern in both financial markets and corporate finance, where ML-driven decision-making is reshaping how investments are managed.

Our study builds on foundational works in urban economics and real estate finance while addressing emerging challenges related to ML and big data use in financial markets. We employ a modified monocentric model inspired by earlier research (e.g., Duranton and Puga (2014)), providing a comprehensive framework for understanding city structures. However, we extend the classical static model to account for dynamic pricing in online platforms by incorporating moving agents actively engaging in the market. Another difference from Duranton and Puga (2014) is that in our model, households that visit online platforms are assumed to live in inhabited units, and through an arbitrage equation, we derive the distribution of demand prices as a function of per capita income and the size of the inhabited unit—two of the main determinants of real estate costs, as reflected in our dataset.

To model dynamic learning in real estate markets, we draw on the Bayesian learning framework (e.g., Bruneel-Zupanc (2020)) and introduce a flexible numerical approach that adapts to the granularity of online marketplace data. This contrasts with traditional models focusing on brick-and-mortar real estate, allowing us better to capture the dynamics of online pricing and demand structures. Additionally, we build on active learning theories from Koulovatianos, et. al. (2011) and Aghion, et. al. (1991). Unlike passive learning models, where sellers' actions do not affect market beliefs, our model incorporates active learning, where sellers adjust their pricing strategies based on new market information.

We address asymmetric information in real estate transactions, expanding on work by Guerrieri, et. al. (2010). In our framework, informed agents (buyers) do not fully internalize the probability of matching their desired property, creating a knowledge gap between buyers and sellers. This gap emphasizes the importance of acquiring market data to uncover arbitrage opportunities, a core concern in financial economics as ML-driven data analysis becomes more integral to market efficiency. By reducing the cognitive frictions associated with real estate pricing, ML techniques can assist in closing this gap, thereby enhancing decision-making in real estate markets.

Finally, our approach explores the standard exploitation-exploration tradeoff commonly discussed in reinforcement learning literature, e.g., Charpentier et al. (2020).

While passive learning frameworks akin to bandit problems limit the seller's ability to influence payoffs, our model employs active learning, where sellers continuously update their understanding of demand prices, leading to a more nuanced exploration of pricing strategies. In this context, we argue that the persistence of information frictions gives lasting value to large-scale data processing techniques in both real estate markets. Our approach is model-based, contrary to the standard reinforcement learning approach. This allows us to estimate the distribution of equilibrium prices structurally and derive the subtle tradeoffs investors face due to the interaction of significant information frictions.

2. Model

This section proposes a theoretical framework to compute the spatial equilibrium of a city by integrating real estate markets, data analysis, and ML-enhanced decision-making. The model ranks neighborhoods based on arbitrage opportunities, considering the interplay between demand prices and online posted prices. By representing land use and investment potential through an opportunity index, we offer a comprehensive methodology for predicting the best investment opportunities within a city's urban structure. This framework addresses the trade-off between price discrimination and liquidity that real estate investors face, differing from the classical exploration-exploitation trade-off in information theory. The model assumes that investors can improve their decision-making by using machine learning to process and interpret large volumes of real estate market data.

2.1 *The Supply Side*

We describe an environment where an infinitely-lived monopolist wants to sell a house or apartment using an online platform. There is geometric discounting $\beta \in (0,1)$, and the owner is risk-neutral. The owner must post a price $x^p_{j,t}$, where p stands for "posted", without knowing the demand price $x^d_{i,t}$. The owner observes the posted prices of similar units $x^p_{-j,t}$. Each unit belongs to a neighborhood. Table A.1 gives the unit's set of characteristics, which determines the range of admissible posted prices for the unit. The seller's objective is to "learn" the distribution of demand prices and obtain the maximum possible price for the unit.

The seller faces a trade-off between selling the unit and exploring the market, which takes time and is conditioned by an unknown demand. This adds the benefits and costs of exploring the market in an unknown environment to the classic trade-off between selling and waiting for a better bid. The reinforcement learning literature (e.g., Charpentier et al. (2020)) refers to this as a trade-off between "exploitation and exploration". The learning literature (e.g., Bruneel-Zupanc (2020)) calls this a "learning externality". The critical difference from a standard dynamic pricing model is an unknown reservation value for the demand.

We depart from the Bayesian learning literature and assume the seller learns through experimentation, as in Aghion, et. al. (1991). Experimentation is the simplest form of

active learning and does not rely on a Bayesian updating procedure. In this framework, the seller updates the upper bound of the demand side of the market only if the unit is not sold. If the unit is sold, the seller stops learning.

For simplicity, let x be the posted price and x_m the demand price. We assume each posted unit belongs to a set of coordinates in a given neighborhood, omitting any reference to the unit's location. In a recursive environment, time subscripts can be eliminated. We assume the seller knows that reservation prices for the demand side of the market are distributed according to a power law distribution with parameters $\alpha, \kappa > 0$:

$$x_m \sim f(\alpha, \kappa, x_m^{UB}) = \alpha \left(\frac{x_m^{UB}}{x_m} \right)^{-\kappa},$$

where x_m^{UB} is the maximum possible price at a given game stage, smaller or equal to the maximum possible price in a given set of coordinates, denoted $\overline{x}$. The information structure implies the seller does not know the exact distribution of demand prices but knows the maximum possible price, which is updated at each game stage. If the unit is not sold, the seller updates the perceived demand price distribution by setting $x = x_m^{UB}$. The seller believes that demand prices are below the most recent posted price if the unit is not sold, leading to the setting of the maximum possible demand price in the next game stage equal to the latest posted price. As in Aghion, et. al. (1991), the seller "learns by experimentation". The seller also knows the minimum and maximum prices: $0 < \underline{x} < \overline{x}$ in their neighborhood, which can be obtained by browsing the online marketplace. If the unit is not sold, the seller updates the price x' such that: $0 < \underline{x} \leq x' < x \leq \overline{x}$. To post a new price, the seller must pay a maintenance cost $r > 0$, increasing with the minimum price $(\underline{x} - x')r$. In the calibration, we add curvature to the cost function: $(\underline{x} - x')^{\zeta} r/\zeta$ with $\zeta > 0$. These costs capture the expenditures associated with preparing the unit before being listed online. As the agent does not have liquid assets, he must borrow at an interest rate r to pay for the housing unit' rehabilitation cost. The seller pays this cost only if the unit is sold.

The seller knows the minimum and maximum possible demand prices in their neighborhood, $(\underline{x}, \overline{x})$ respectively, and assumes demand prices are distributed according to the power law distribution mentioned earlier. In the calibration section below, the set of possible demand prices will depend on the clusters presented in table A.1.

The sequencing of the game is as follows. If the unit is sold, the owner receives $\frac{\eta x}{1-\beta}$. If the unit is not sold, the owner updates the upper bound $x = x_m^{UB}$ and pays the maintenance cost $(\underline{x} - x')r$. Then, the owner could: a) Sell at the updated price with probability $\overline{F}(x'; \alpha, \kappa, x)$; b) Re-post with probability $F(x'; \alpha, \kappa, x)$.

Formally, the seller faces the following problem:

$$V(x; \alpha, \kappa) = \max_{\underline{x} \leq x' \leq x} (\underline{x} - x')r + \beta \left[\frac{\overline{F}(x'; \alpha, \kappa, x)\eta x'}{1-\beta} + V(x')F(x'; \alpha, \kappa, x) \right]$$

where

$$F(x';\alpha,\kappa,x) = \int_{\underline{x}}^{\eta x'} \alpha \left(\frac{\eta x}{x_m}\right)^{-\kappa} dx_m$$

$$\overline{F}(x';\alpha,\kappa,x) = \int_{\eta x'}^{\eta x} \alpha \left(\frac{\eta x}{x_m}\right)^{-\kappa} dx_m$$

To understand the problem, note that if the seller posts $x' = \underline{x}$, the unit is sold at the lowest possible price without any cost. This is the property owner's outside option. If the owner decides to explore the market by posting $x' > \underline{x}$, a cost is incurred, and the unit sells with probability $\overline{F}(x';\alpha,\kappa,x)$. There is a trade-off: setting the posted price at the highest value x results in a zero probability of selling, while reducing the price increases the probability of selling but lowers the discounted utility of the transaction $\eta x'/(1-\beta)$. If the unit is not sold, the owner learns that the reservation value lies between $\underline{x}$ and $\eta x'$, and the search value $V(x')$ is weighted using $F(x';\alpha,\kappa,x)$. The set of possible demand prices shrinks at every game stage, similar to the standard cake-eating problem where the cake size decreases with each stage of the recursive problem. Thus, V is increasing in x, and the policy function $x'(x)$ is increasing, with $x'(x) < x$ if $x \in (\underline{x}, \overline{x}]$ and $x'(x) = x$ if $x = \underline{x}$.

2.2 *The Demand Side*

We employ the standard monocentric model as outlined by Duranton and Puga (2014) to model the demand with a slight modification. In our model, there is no production. Hence, we cannot use the marginal cost of producing a unit to determine the demand price. Instead, we assume that a household lives in a unit priced at x_m and seeks to move to a new unit priced at x'_m. The household must be indifferent between locations, which provides a threshold utility level from which we can derive the entire demand curve.

Assuming the household has an income ω and incurs a commuting cost factor τ. Commuting cost τd varies with distance d to the household's center. We depart from the standard monocentric model by assuming the existence of a specific center for each household. This center has the same location as his job. Thus, he commutes to work and then shops in the same location. He consumes housing h and a numeraire good z. The relative housing price x_m is a function of distance d to the household's center, denoted as $x_m(d)$. The budget constraint is $\omega - \tau d = x_m(d)h + z$. The household's utility function $U(h,z)$ depends on both goods. From standard consumer theory, we derive the Marshallian demands $(h(x_m(d)), z(x_m(d)))$ and the indirect utility $\nu(x_m(d), \omega - \tau d)$.

Now, if the household decides to move, he must choose a new pair (x'_m, d') such that they remain indifferent: $\nu(x'_m(d'), \omega - \tau d') = \nu(x_m(d), \omega - \tau d)$. This defines a function $x'_m(d', x_m(d), d, \omega)$, omitting τ for simplicity. For simplicity, we denote this as $x'_m(d', x_m, \omega)$, dropping the explicit dependence on d. As in the standard monocentric model, x_m (x'_m) decreases with d (d'). The relationship between real estate prices and distances to the household's center represents a trade-off between commuting costs and

the size of the real estate unit. Proximity to the household's center results in higher real estate prices but lower commuting costs.

The arbitrage equation between two locations (d, d') yields a moving distance. We need the exact location from which each household moves to estimate the model. As this information is unavailable, we modify this version of the demand model in the next section.

2.3 *Tractable Demand Side*

Given the recursive structure of the supply side, solving it numerically is straightforward. For the demand side, we modify the previous model slightly. In particular, we adopt a Cobb-Douglas demand structure for (h, z) with a housing parameter γ that represents the housing expenditure share after paying for the commuting cost (see equation (1)). The household's demands $\hat{h}$ and $\hat{z}$ and value function $V(\omega, x_m)$ are given by:

$$\hat{h} = \frac{\gamma(\omega - \tau d)}{x_m(d)} \tag{1}$$

$$\hat{z} = (1-\gamma)(\omega - \tau d) \tag{2}$$

$$V(\omega, x_m) = \left[\frac{\gamma(\omega - \tau d)}{x_m}\right]^{\gamma} [(1-\gamma)(\omega - \tau d)]^{1-\gamma} \tag{3}$$

This system of equations allows solving for demand prices $x_m(d')$ through arbitrage between two locations:

$$x_m(d') = \left[(\omega - \tau d')\right]^{1/\gamma} \frac{x_m(d)^{1-(1/\gamma)} \gamma^{1/\gamma}}{\hat{h}^{1/\gamma}} \tag{4}$$

We get a relationship among $d, d', x_m(d')$ and $x_m(d)$ by eliminating the arbitrage opportunity between two locations d and d' using the value function (3). We eliminate d (not observable) by using equation (1). We can get rid of $x_m(d)$ from equation (4) if assuming a stable relationship between wages and house prices (in our Data section, we show that there is a strong linear correlation between these two variables. See figure A.3). We will derive a version of the demand price that depends on observable variables later. Of course, in the model, this information is unavailable to the seller or, if available, he does not know how to process it.

2.4 *Monocentric Partial Equilibrium*

From the supply side, we obtain $x'(x_s, \underline{x}, \overline{x}, L)$, where x_s denotes the state price for the supplier and L is the geographical location of the posted unit. In the learning game, an owner adjusts the price to x' after failing to sell at x. Given the information structure, the owner does not know demand prices and assumes $x \geq x_m$, where x_m is the demand

price. When failing to sell, the owner updates the price by setting $x' \leq x$. Thus, the previous price influences the current price, making it a state variable.

We denote the translation of the property's location into the distance between the property and the household's center as L^d. Of course, L^d, is demand-specific even though L is the same for all potential buyers. From the demand side, we derive $x'_m(d', x_m, \omega)$. Note that the distance associated with the demand price, denoted by d', may differ from L^d, even if $x' = x'_m$. This is because the agent's preferences *only* determine d'. [1] The relationship between posted prices and the effective location of the unit, (x', L), may be independent of the household's preferences. Thus, even assuming that demand and posted prices are equal, d' and L^d may differ as $x_m(d') \neq x_m(L^d)$.

Recall that L^d is *not observable* as the households' center is unknown because geographical information for the distribution of wages at the neighborhood level for Madrid is not publicly available. Thus, not only does L^d differ across households, but we also do not have sufficient granularity to identify it empirically. However, we can define an equilibrium notion based on observable variables. Given asymmetric information on the supply side, a unit is sold if the reservation price for demand is at least equal to the posted price, $x' \leq x'_m$, and the unit is closer to the household's center than the distance consistent with the demand price, $L^d \leq d'$. These two conditions are independent between each other and, thus, must be stated separately in the definition of equilibrium.

Let Θ be a vector of parameters of a given distribution, with Θ_m and Θ_s characterizing the distribution of x'_m and x', respectively. Later, we shall allow the parameters associated with posting price distribution to be location-specific and different across categories of real estate units. Also, the parameters associated with demand will be location-specific through the distribution of wages. Complete learning occurs if $\Theta_m = \Theta_s$. We can now define an equilibrium.

Definition: Spatial Equilibrium (SE). *A spatial equilibrium (SE) comprises price functions x' and x'_m, a distance d', and parameters $[x_s, x_m, L, \underline{x}, \overline{x}, \omega]$. The agent purchases the unit if: (i) $x'(x_s, \underline{x}, \overline{x}, L) \leq x'_m(d', x_m, \omega)$, and (ii) $L^d \leq d'$. Moreover, (iii) complete learning occurs if $\Theta_m = \Theta_s$. It is possible to have a SE that only satisfies (i) and (ii). We will call this equilibrium inefficient. If, in addition to these two conditions, the SE satisfies (iii), we will say the equilibrium is efficient.*

Definition: Threshold Distance *(d'_{TR}). A threshold distance (d'_{TR}) is given by $x'(x_s, \underline{x}, \overline{x}, L) = x'_m(d'_{TR}, x_m, \omega)$. Thus, we can write $d'_{TR}(x_m, x_s, L, \omega, \underline{x}, \overline{x})$.*

Condition (iii) of the definition of equilibrium specifies *complete learning*. We assume the seller observes only a subset of all possible parameters of posted prices' cross-sectional distribution for a given period in their neighborhood. Thus, even though the

[1] We are not modeling the effect of the spatial distribution of firms on wages and, thus, on the households' center. In a model with production, the location of firms may affect the distribution of wages, which in turn implies that the distance of a given agent d' can be affected by the preferences of other households.

agent can update the expected demand price distribution, convergence through learning to the actual distribution of demand, in the sense of (iii), is impossible as this would imply that the agent can adjust the expected distribution to match any moment of the empirical distribution, which, by hypothesis, is not allowed.[2] This lack of convergence is essential for our analysis of arbitrage as it generates a metric of *best arbitrage opportunities*, to be defined later. If condition (iii) is met, there are no arbitrage opportunities for the seller across regions, leaving no surplus to extract from the consumer. We will artificially generate an efficient equilibrium by imposing $x' = x'_m$. However, this is only to study a benchmark as this type of equilibrium cannot emerge through decentralized learning.

Because the demand price x_m increases with income ω, we can define a contour $x_m(\omega)$ to compute d'_{TR}. For every combination $(x_s, L, \omega, \underline{x}, \overline{x})$, we find a distance by setting $x' = x'_m$, which is a function of the inhabited unit $\hat{h}$. The next subsection deals with this fact. We call this distance d'_{TR}, and we can express it as $d'_{TR}(x_s, L, \omega, \underline{x}, \overline{x})$. This translates the geographical location of the unit into its distance to the household's city center.

2.5 *The Distribution of Distances in the Efficient Model*

We say that the price system is efficient if there are no informational frictions generated by the seller's inability to know the demand distribution. To understand the basics of the model, we first assume that the pricing system is *efficient*, i.e., $X_m = X_s$, so condition (iii) is satisfied. We must set $x_m(d') = x'$ to solve the above equilibrium notion. Unfortunately, we don't observe d. Then, as we did in section 2.3, we can use equation (1) to express the demand price in terms of observable variables. After some algebra, we obtain an expression for the distance as a function of observable variables:

$$d' = \left[\omega - \left(\frac{x'}{\psi(x_m, \hat{h})}\right)\right]\frac{1}{\tau}, \tag{5}$$

where:

$$\psi(x_m, \hat{h}) \equiv \frac{x_m^{1-(1/\gamma)}\gamma^{1/\gamma}}{\hat{h}^{1/\gamma}} \tag{6}$$

If we assume that all units are sold when the demand and the supply prices are equal, we can solve the model. We call this version *efficient*. Due to asymmetric information on the supply side, we may observe that i) the demand side of the market keeps all the surplus, and ii) the supply side forces some buyers to increase their reservation price. In the first case, we observe $X_m^i > X_s$ for some consumers i. This implies that, by accepting the deal, these consumers keep all the surplus associated with the bounded rationality assumption used to model digital platforms. In the second case, we observe $X_m^j < X_s$,

[2] For instance, the seller does not observe either the kurtosis or any higher moment of the empirical distribution, which prevents an exact match of demand prices for most empirical distributions.

for consumers j, in which case the buyer accepts a price above his reservation price, giving the surplus to the seller. By setting $X_m^i = X_s$ for all i, we eliminate the frictions associated with asymmetric information, or equivalently, we assume that the equilibrium is efficient.

We obtain the distribution for distances illustrated in Figure 1 by assuming $X_m = X_s$ and the following two conditions derived from our empirical analysis for Madrid: i) a lognormal distribution of wages with mean equal to zero and standard deviation equal to 0.5, ii) that the household living in an apartment of 58 square meters pays the mean price for housing x_m.

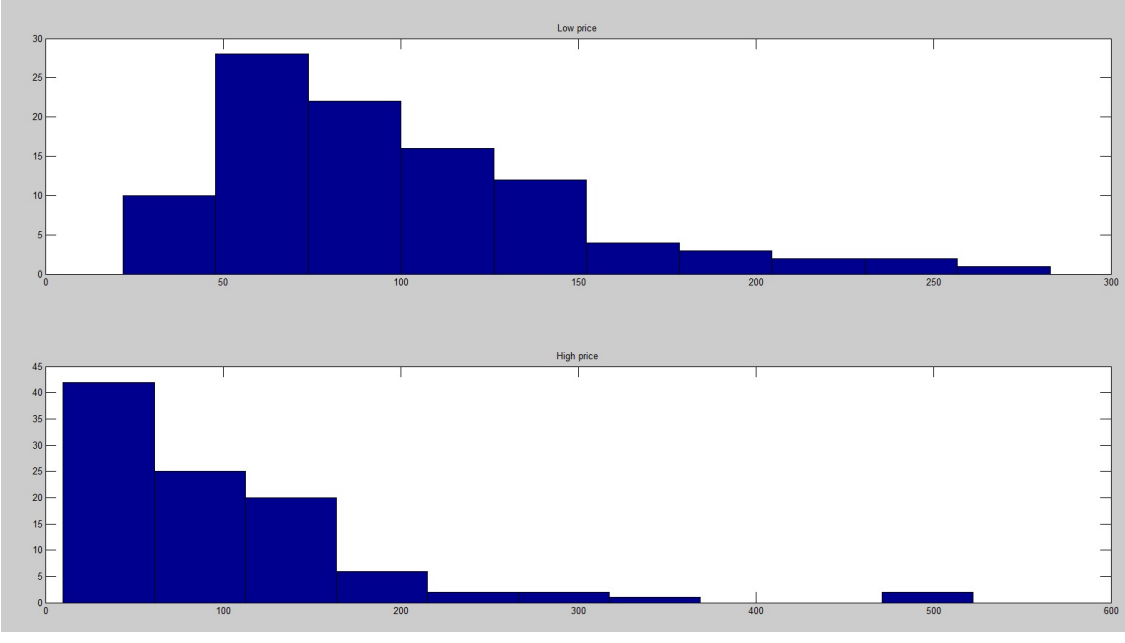


FIGURE 1. Distribution of distances

The top graph in Figure 1 contains the distribution of the household's distances of low-price apartments (i.e., at the 1st quartile of the price distribution), and the bottom figure depicts these distances for high-price apartments (i.e., at the 3rd quartile of the distribution). From the definition of Spatial Equilibrium, it is clear that, even after setting $X_m = X_s$, we have $L^d \leq d'$ for all the units sold. As we depict d' using equation (5), we observe an upper contour of the actual distribution of owner's distances for each unit L^d.

From the bottom part of Figure 1, we observe that the agents living in high-priced apartments faced smaller commuting distances to their centers than agents living in low-price apartments. In the model, the center is household-specific. Thus, we see larger distances for households residing in low-price apartments, typically far from downtown.

Equation (5) implies that for a given set of prices (x', x_m), distance is monotonically increasing in wages. This implies that wealthy households live far away from their center and thus commute a lot, and poor households live close to their center. This urban

configuration is observed in some cities in the United States, where rich households live in the suburbs. However, this does not exactly match the urban configuration of Madrid, where there are only a few rich neighborhoods far away from the city's center, while most of the rich households live in downtown. [3] To accommodate this pattern into our model, we assume that households paid a uniform fraction of their income for housing before moving to another unit. That is, $x_m = \theta w$. This assumption is borrowed from figure A.3. Under this assumption, equation (5) becomes:

$$d' = (1/\tau)\left[w - \Lambda(x', \hat{h})\,(\theta w)^{\frac{1}{\gamma}-1}\right] \equiv d'(w), \tag{7}$$

where $\Lambda(x', \hat{h}) \equiv x' \hat{h}^{1/\gamma} \gamma^{-1/\gamma}$. If households spend less than half of their income on housing, $\gamma < 0.5$, then equation (7) is concave and increasing (decreasing) if $w < w^*$ ($w > w^*$, respectively) with w^* given by:

$$w^* \equiv \theta^{\frac{1-\gamma}{2\gamma-1}} \left[(1-\gamma)\Lambda(x', \hat{h})\right]^{\frac{\gamma}{2\gamma-1}}. \tag{8}$$

Moreover, distance has a maximum at w^*, $d'_{max} = d'(w^*)$ and, as $\gamma < 0.5$, d'_{max} is decreasing in x'. That is, the maximal distance decreases in the price of the purchased unit. Finally, $d'(0) = 0$ and $d'(\bar{w}) = 0$, where:

$$\bar{w} \equiv \left[\theta^{\frac{1-\gamma}{\gamma}} \Lambda(x', \hat{h})\right]^{\frac{\gamma}{2\gamma-1}}. \tag{9}$$

Using the set of equations given by (7) - (9), the distribution of wages induces a distribution of distances. As d'_{max} decreases in the unit's price, the model implies that expensive units are closer to the center of households. Moreover, as equation (9) also increases in x', more affluent households live in expensive units and are closer to the center. We depict these patterns in figure 2 below.

Housing affordability is a frequent topic in the literature. In our model, γ represents the fraction of income devoted to housing. The paragraphs above imply a threshold value for γ, 0.5, that generates a reasonable commuting pattern based on income. Usually, for any $\gamma > 0.5$, we say housing is severely unaffordable. Thus, the model predicts that frequent commuting patterns, with wealthy households (between the 80th and 100th percentile in the horizontal axis in figure 2) purchasing high-price units (the green line in figure 2) and living close to their center (the values in the vertical axis for the green line are below their analogous for the blue line), are deeply related to housing affordability (as we need to assume that $\gamma < 0.5$ to obtain this figure).

[3] In the empirical section, we depict how prices peak near Salamanca and then decrease as we approach the limit of the first ring of the city, formed by the M-30 highway.

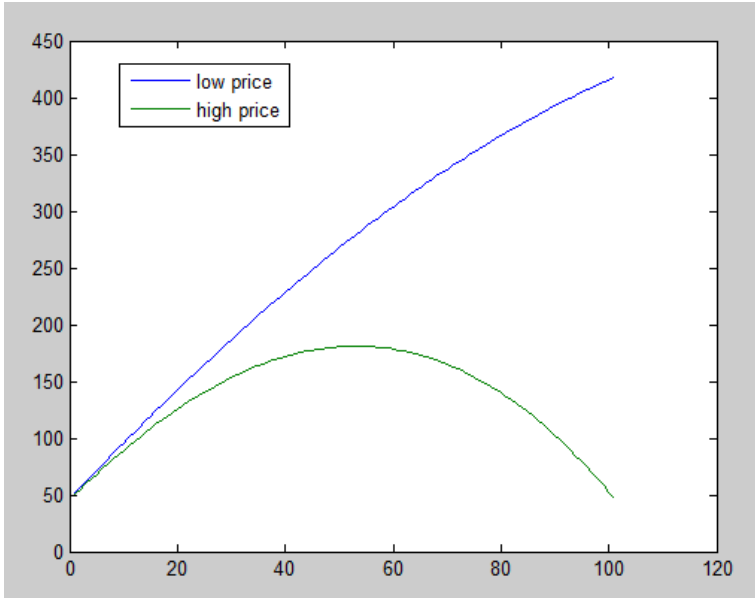


FIGURE 2. Distance as a function of wages. We depict distance according to equation (7) in the vertical axis and the grid position of wages in the horizontal axis. As d' is a function of the price of the house, we set it to the average value in the first (i.e., the low price) and third (i.e., the high price) quartile of the distribution.

2.6 *General Equilibrium*

The previous definition of equilibrium does not include a market-clearing condition and, thus, cannot explain the spatial distribution of properties and their prices from both the supply and demand sides. To address this issue, we introduce a Poisson arrival rate to the model. As is standard in the literature, a Poisson distribution has realizations in the positive integers, allowing us to model the number of visits to a given ad in the online marketplace. Note that equations (4)-(6) imply that the demand price is given by:

$$x_m(d') = \left[(\omega - \tau d')\right]^{1/\gamma} \frac{\gamma^{1/\gamma}\theta w^{1-(1/\gamma)}}{h^{1/\gamma}}. \tag{10}$$

A household can live and work in different municipalities, significantly affecting its commuting distance. We shall consider additional household characteristics to accommodate this possibility and compute distance d'. A household is of type 1 if he lives where he works and is of type 2 if he works elsewhere.

The Autonomous Region of Madrid ("Comunidad de Madrid") contains multiple municipalities, including the City of Madrid. Each municipality is divided into multiple non-overlapping neighborhoods. We assume a continuum of households of each type per municipality for tractability. For simplicity, we assume that households living in the City de Madrid also work there. There are J municipalities, and in each of them, $\phi^j \in [0,1]$ households are of type 1. For each municipality, we assume that the household

lives in the geographical center of the municipality, except for the City of Madrid, where we differentiate among neighborhoods and assume that households live in the center of the neighborhood.

Let $j = M$ be municipality de Madrid. We will consider the market clearing condition for this municipality. Note that for $j \neq M$, type 2 households that reallocate to M will experience a significant reduction in their commuting distance. We will assume they work in the center of M, the location of which will be estimated below in our empirical analysis. Thus, we will compute the distance from the advertised unit to the center of M. Similarly, for type 1 households in other municipalities, we will compute the distance from the center of municipality j to the advertised unit. Finally, let b denote a neighborhood. There are B neighborhoods inside M. We denote the number of units advertised in M by N_s^M. We only observe household socioeconomic data at the neighborhood level inside the City of Madrid. Thus, we only differentiate across neighborhoods for the City.

Given equation (10), the number of households of type l, who are characterized by a wage w, an apartment of size h, and location (j,b), that move to an apartment with price X_s at location c are given by:

$$N_m([D_l', H](j,b;c)) \sim P_p\left[\lambda P_{LN}\left(X_m^{j,b}(d_l', h) > X_s(c)\right)\right], \tag{11}$$

where P_p stands for a Poisson distribution with parameter λ, which will be used to match data; P_{LN} is the distribution of the demand price X_m that has an underlying uncertainty given by the log-normal distribution of wages; and $[D_l', H](j,b;c)$ is the distribution of distances and apartment sizes of agents of type l that move from (j,b) to c. Equipped with equation (11), we can compute the ex-post distance of households $D_l'(j,b;c)$ after moving to the advertised unit $X_s(c)$. Then, we have the following market clearing conditions:

$$\sum_c \sum_{j,b} \phi^j N_m(D_1'(j,b;c)) + (1-\phi^j) N_m(D_2'(j,b;c)) \leq \sum_{c \in M} N(X_s(c)) \equiv N_s^M, \tag{12}$$

where $N(X_s(c))$ is the number of units advertised at coordinate c. Note that expression (11) implies that markets are not segmented, which in turn means that equation (12) is compatible with different prices across neighborhoods. Later, based on table A.1, we will segment the market by introducing housing types and household preferences restrictions, which implies different market clearing conditions for each category. We feed the model with the observed $X_s(c)$ and, given the mean of wage distributions per (j,b), we calibrate the model to match the dispersion of X_s as a function of the dispersion of wages. The calibrated model gives an *endogenous* distribution of distances that can be used to evaluate the impact of transportation costs τ and the efficiency of online market platforms λ on the *distribution of demand prices* per destination (j,b,c) and origin $\{c_1, ..., \} \equiv M(b)$, where $M(b)$ is a neighborhood inside the City of Madrid and we classify origin and destination according to the left and right-hand side of equation (12) respectively.

We can now define a *General Equilibrium*.

Definition: General Equilibrium (GE). *A General Equilibrium (GE) for a region defined by a set of municipalities J, a set of neighborhoods B, and a set of coordinates $c \in M$ is a collection of policy functions $(x(c), x_m^{j,b})$ for parameters $\Theta_m^{j,b}, \Theta_s(c), \gamma$, and λ, such that equation (12) holds.*

Note that there is only one market-clearing equation and more than one parameter even if we set the mean of wages in $\Theta_m^{j,b}$ based on available data. The only equation that defines this equilibrium may generate an under-determined calibration system as, even if the cardinality of $\Theta_m^{j,b}$ and $\Theta_s(c)$ are equal, we can match the only equilibrium equation using only one parameter. Thus, as markets are not segmented (i.e., potentially, a consumer can purchase any unit), we must match either γ or λ. To improve the quality of the match, we segment the market using the clusters presented in Section A.1.

2.7 *Best Arbitrage Opportunities*

Using the definition of GE and the fact that realized transactions have variable consumer surplus $X_m^{j,b}(\omega) > X_s(c)$ depending on the structural parameters of the model and the distribution of wages in each location, we can rank different neighborhoods based on the distance between the demand and the supply price. The smaller this distance, the lower the information friction, and thus, we observe a better, more efficient housing market from the seller's perspective.

Let the surplus S of neighborhood b be:

$$S(b) \equiv \sum_{c \in b} \sum_{j,b} \phi^j N_m(D_1') P_{LN}\left(X_m^{j,b}(d_1') > X_s(c)\right) + (1-\phi^j) N_m(D_2') P_{LN}\left(X_m^{j,b}(d_2') > X_s(c)\right) \tag{13}$$

That is, a neighborhood b has a low surplus if the probability of having information frictions $X_m > X_s$ is low *and* the number of visits to its post is also low. Then, we can define the Best and Worst Arbitrage Opportunity as follows:

Definition: Best and Worst Arbitrage Opportunities (BAO, WAO). *The neighborhood $\underline{b} \in B$ with the Best Arbitrage Opportunity (BAO) is given by: $\underline{b} = argmin_{b \in B}[S(b)]$. Similarly, the Worst Arbitrage Opportunity (WAO) is given by $\overline{b} = argmax_{b \in B}[S(b)]$.*

Intuitively, the neighborhood with the BAO has the minimum possible consumer surplus and, thus, the highest possible offer price given the distribution of demand and information structure. That is, as we have $\underline{b} < \overline{b}$, *abstracting from waiting times* (i.e., if the real interest rate is lower or equal to zero), any investor would want to sell units in $\overline{b}$ to purchase units in $\underline{b}$ as it maximizes their expected return. Of course, as they do not observe X_m, they cannot take this arbitrage opportunity. Further, as we can compute $S(b)$ for any neighborhood b, we can construct a ranking S.

Notice that neighborhoods b with high posted prices $X(c)$, with $c \in b$, do not necessarily have a low surplus $S(b)$, thus better arbitrage opportunities. This is possible as we

have two opposite effects: i) a high offer price reduces the informational friction for a given distance, and ii) according to figure A.1, as these units are closer to the (empirical) center of the city, the number of visits to their post increases as it attracts many type 2 households. As these households do not work where they live, they have incentives to move closer to the city center, which concentrates on high-priced units in Madrid. Thus, high-traffic in the online marketplace increases informational friction, affecting the arbitrage opportunities from the seller's perspective. However, the first effect tends to dominate. Among the thirty-two municipios in the Madrid Autonomous Region, only five of them (Coslada, Majadahonda, Pozuelo de Alcorcon, Rivas-Vaciamadrid, and Las Rozas) have more than 50% of type 2 agents.[4] This fact implies that a neighborhood with a higher than average offer prices implies better arbitrage opportunities and a lower a) ranking, b) probability of selling the unit, c) visits to the units posted in that neighborhood, and d) information frictions. Critically, arbitrage opportunities generate a tradeoff for any investor. Building a portfolio composed entirely of neighborhoods with high arbitrage opportunities implies longer waiting times to sell the unit and, thus, higher opportunity costs given any positive interest rate. In this sense, a ranking S across neighborhoods indicates only efficiency in the match from the seller's perspective but not necessarily the best option from an investor's perspective.

3. Empirical Analysis

3.1 *Data Sources Description*

In this study, we collected data from multiple sources.

1. *Fotocasa Real Estate Listings*: The Fotocasa dataset includes data from real estate listings detailing various properties available for sale in Madrid. It contains key information such as listing price, property size, floor number, and available amenities (e.g., swimming pool), offering a comprehensive view of the market offerings in the city. Listing data from March 2020 up to December 2021 is available. We have 346,035 property listings distributed in time as shown in Figure 3.

2. *Google Places Data*: This dataset, obtained from the Google Places API, provides information on points of interest (POIs) within Madrid. These POIs include cultural sites such as museums and essential infrastructure like metro stations. The concentration and proximity of these amenities can significantly influence property values. In Figure 4, we show Madrid's 10 most prominent POIs and surroundings.

3. *Geographical Description of Madrid's Neighborhoods*: This dataset provides detailed geographical information for the 131 neighborhoods in Madrid. It includes boundaries, area size, and geographical characteristics essential for spatial analysis.

[4] See Informacion Urbanistica, Diagnostico de Ciudad, Vol. 1, page 72.

4. *Sociodemographic Information*: This dataset contains socioeconomic and demographic data for each neighborhood in Madrid. Key variables include average household income and per capita income.

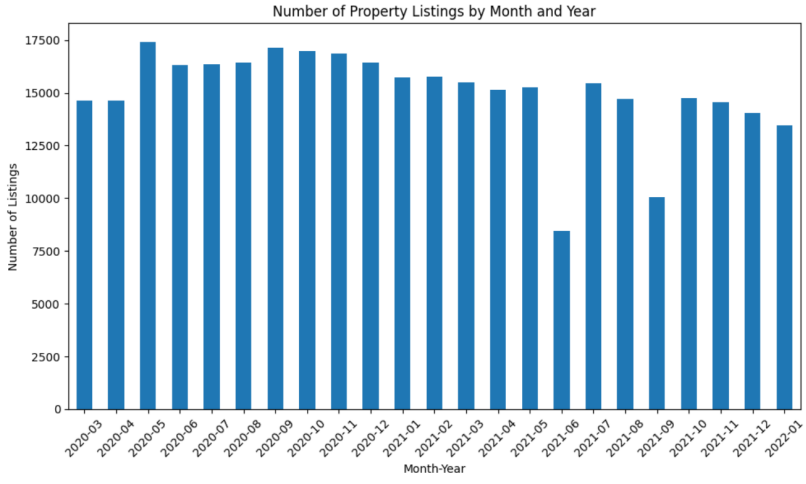


FIGURE 3. Number of property listings in Madrid

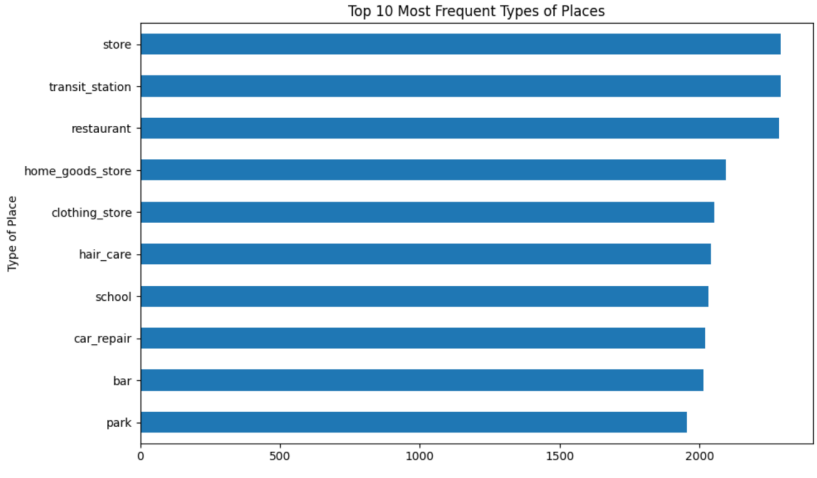


FIGURE 4. Number of POIs in Madrid and surroundings

1. *Fotocasa Real Estate Listings*: The Fotocasa dataset includes data from real estate listings, detailing various properties available for sale in Madrid. It contains key information such as listing price, property size, floor number, and available amenities (e.g., swimming pool), offering a comprehensive view of the market offerings in the city. Listing data from March 2020 up to December 2021 is available.

2. *Google Places Data*: This dataset, obtained from the Google Places API, provides information on points of interest (POIs) within Madrid. These POIs include cultural sites such as museums and essential infrastructure like metro stations. The concentration and proximity of these amenities can significantly influence property values.

3. *Geographical Description of Madrid's Neighborhoods*: This dataset provides detailed geographical information for the 131 neighborhoods in Madrid. It includes boundaries, area size, and geographical characteristics that are essential for spatial analysis.

4. *Sociodemographic Information*: This dataset contains socioeconomic and demographic data for each neighborhood in Madrid. Key variables include average household income and per capita income..

5. *Cadastre Information*: The cadastre dataset offers comprehensive details about each parcel in Madrid. This includes information on the type of land use (e.g., residential, commercial) for each unit within a parcel, which is necessary for understanding the distribution and utilization of real estate across the city.

3.1.1 *Data Analysis and Predictive Modeling of Property Values in Madrid* We have carefully examined the real estate market data, with a focus on property prices, income levels, and the development of a predictive model for estimating property values. For the sake of brevity, we refer the reader to the Appendix. Here we summarize our analysis and findings.

Our analysis begins by mapping average property prices across different neighborhoods, showcasing the wide variation in real estate costs throughout the city. Central neighborhoods exhibit the highest property prices, while more affordable areas are located toward the outskirts. A visual representation of income levels complements this analysis, highlighting a strong correlation between neighborhood wealth and property values, with higher-income areas displaying significantly higher prices per square meter.

A regression model is introduced to estimate property prices based on a comprehensive set of features, including property characteristics, neighborhood income levels, and proximity to amenities. The model employs advanced machine learning techniques, specifically CatBoost, optimized through hyperparameter tuning using the Optuna framework. The resulting model demonstrates strong predictive performance, achieving a high degree of accuracy in estimating property prices per square meter. A detailed feature importance analysis reveals that property size, average personal income, and proximity to key amenities, such as embassies, are the most influential factors in determining property values.

We also delve into a clustering analysis that segments properties into five distinct groups based on intrinsic characteristics like size, number of bedrooms and bathrooms, and the presence of amenities such as lifts, swimming pools, and garages. These clusters represent different market segments, ranging from more affordable housing to premium properties, offering a nuanced understanding of the property market in Madrid. Each

neighborhood is assigned to the cluster most representative of its property characteristics, and the price distributions within each cluster are analyzed to provide insights into the housing market's structure.

In addition to the data analysis, we explore the relationship between property size and price per square meter across neighborhoods, using a LOWESS curve to highlight the trend that smaller properties in denser, central areas tend to command higher prices per square meter. Larger properties, particularly those in less central neighborhoods, often see diminishing returns in terms of price per square meter.

Overall, we provide valuable insights into the dynamics of Madrid's real estate market, identifying key factors that influence property prices and demonstrating the effectiveness of machine learning techniques in predicting real estate values. For further details on the data, visualizations, and models discussed, see the Apendix.

3.1.2 *Why Have We Used a Gradient Boosting Procedure* Why have we used a Gradient Boosting procedure instead of a simpler or more standard approach like linear regression? The primary distinction between these modern methods and classical linear regression lies in their ability to capture complex relationships within the data. Gradient Boosting methods, and CatBoost in particular, excel due to their flexibility in modeling non-linearities, interactions, and their ability to handle categorical features effectively (Friedman, 2001). This makes them highly suitable for real-world datasets, where such complexities are common and can significantly impact the model's performance.

On the other hand, linear regression's simplicity is both its strength and its limitation. While it provides easily interpretable coefficients that describe the linear relationship between features and the target variable, it often fails to capture key patterns if the true relationship is non-linear or if there are significant interactions between features (Montgomery et al., 2021). Additionally, linear regression can be sensitive to multicollinearity and outliers, which can distort the results, making the model less reliable in more complex settings (Kutner et al., 2004).

In the context of property price prediction, as seen in our study, CatBoost has demonstrated its ability to leverage a variety of features, including both numerical and categorical variables, to build a robust model. The flexibility of CatBoost allows for the automatic encoding of categorical features, which avoids the need for manual preprocessing like one-hot encoding, a common step in linear regression or other econometric models (Prokhorenkova et al., 2018). Furthermore, the use of Permutational Feature Importance provided clear insights into the most significant factors affecting property prices, such as property size and average personal income, which might not have been fully captured by a linear regression model due to its limitations in modeling interactions and non-linear relationships. Once the main features are obtainer by the Feature Importance procedure, further analysis on those features may be pursued in a more classical way, as it was done with *size* and *Average Personal Income*.

It is important to acknowledge, however, that the interpretability of machine learning models like CatBoost can be more challenging compared to linear regression. While tools such as SHAP (SHapley Additive exPlanations) and Permutational Feature Importance help bridge this gap by attributing importance scores to individual features, they

often require more computational resources and are less straightforward to interpret than the coefficients of a linear model (Lundberg & Lee, 2017, Breiman, 2001). Despite this, the added complexity of Gradient Boosting models often results in superior predictive performance, which justifies their use in scenarios where accuracy is paramount.

3.2 *Calibration*

This section presents the procedure for taking the model to the data. To sharpen the calibration, we will define a vector of attributes of units $l = 1, ..., L$ that will allow us to segment the market, generating L different market clearing equations (12). Based on table A.1, $L = 5$, each segment l is associated with a different cluster and, according to figure A.8, the average price per segment is increasing.

Equation (10) gives the demand price as a function of the distance d', which is not observable. However, we can use equation (7) to get an observable expression as a function of the *previous* posted supply price at coordinate c, x, with $X_s(c,x)$. Note remarkably that if we set $x' = X_s(c)$ in equation (7), given the wage rate ω, equation (10) implies $X_m = X_s(c)$. That is, we are in an efficient equilibrium. As we know that the seller faces information frictions as he does not observe the distribution of demand, we *assume* that any potential buyer computes his distance d' as a function of the previously posted price x. We assume that the household *moves first* computing distance d' as a function of x. Then, the seller updates the posted price $X_s(c,x)$. As distance d' is decreasing in the price of the unit x through equation (7), $d'(x)$, as it is also the current posted price as a function of the past posted price $X_s(c,x)$, equation (10) implies $X_s(c,x) < X_m(d'(x))$ as desired. Note remarkably that *learning dynamics, through the sequence of posted prices and the game timing, affects arbitrage opportunities.*

To determine the demand price, besides distance d', we need to include the wage level w, the current unit inhabited by the agent h^l, the fraction of expenditure devoted to housing γ, transportation costs τ, and the wage-rent ration θ [5]:

$$X_m^{j,b,l}(d'(x), h(l), \omega(j,b); \gamma, \tau, \theta), \tag{14}$$

where (j,b) stands for the municipality and the neighborhood, respectively, and ω is a realization from a log-normal distribution centered at the observed mean of net per capita income at (j,b). Because we only observe data at the neighborhood level when $j = M$, we set $b = 1$ for all $j \neq M$.

The critical assumption for identifying demand prices is that we use equation (7) as a function of the posted price. This is possible as the demand side of the market has access to the online posted price. This equation coupled with (10) allows us to derive a (partial equilibrium) closed form for the demand price for a unit posted at coordinate $c \in b$:

[5] Note that θ is the wage-rent ratio per square meter and γ is the share of housing-related expenditure over after-commiting income. Thus, we have $\gamma = \frac{\theta w h^l}{w - \tau d}$, which in turn implies that distance to the agent's center of the inhabited house is determined.

$$X_m^{j,b,l;c} = \omega(j,b)^{\epsilon_1} h(l)^{\epsilon_2} x(c)^{\epsilon_3} \Gamma, \tag{15}$$

where $\Gamma \equiv \gamma^{(1/\gamma)(1-(1/\gamma))} \theta^{(1/\gamma)+1} > 0$, $\epsilon_1 \equiv (1/\gamma)(\gamma - 2 + (1/\gamma)) > 0$, $\epsilon_2 \equiv (1/\gamma)((1/\gamma) - 1) > 0$, and $\epsilon_3 \equiv 1/\gamma > 0$. From equation (15), we know that the demand price increases in per capita income and in unit size, as predicted by Figures A.5 and A.6 respectively. As transportation costs enter proportionally in equations (10) and (7), the demand price in (15) is independent of them. Of course, these costs affect the distribution of equilibrium distances through equation (7).

We now turn to specify the observable version of the supply. We assume that each municipality and neighborhood (j,b) and type of unit l form a cluster of posted prices compactified by $\left[\underline{X}^{j,b,l}, \overline{X}^{j,b,l}\right]$. To complete the description of the supply, we need the previously posted price x, the scale and curvature parameter of the distribution of expected demand prices α, κ, the interest rate r, and the curvature parameter associated with the cost function ζ:

$$X_s^{j,b,l}\left(c, x; \alpha, \kappa, r, \zeta, \underline{X}^{j,b,l}, \overline{X}^{j,b,l}\right). \tag{16}$$

Finally, we segment the real estate market through a *ladder assumption* (Rosenthal (2014)). We assume that if the agent lives in a unit of type l, h^l, he wants to move to a unit of at least that category $X_m^{l'}$, with $l' \geq l$. This assumption generates L market-clearing equations (12). Remarkably, the higher the agent is on the ladder, the fewer options to move, generating lower consumer surplus $S^l(b)$ and thus better arbitrage opportunities from the investor's point of view, as he discriminates prices more efficiently. As described in the previous subsection, the seller prefers locations with lower expected consumer surplus $X_m > X_s$. Due to asymmetric information, he cannot bargain this surplus and thus try to minimize it. From an investor's perspective, selling a location with a high expected consumer surplus and purchasing one with a lower surplus is an arbitrage opportunity as it increases the probability of efficient price discrimination. Typically, when we move up through the ladder, we increase posted prices on average; the ladder assumption implies that better arbitrage opportunities are in high-priced units. As mentioned, these units are harder to sell due to less website traffic, increasing investor's waiting times. The following section addresses these issues.

3.3 *Investors*

Equation (12) allows us to compute market clearing prices for every market segment l. As we showed in figure A.8, these segments are *ordered*. For instance, it could be the case that $\underline{X}^{j,b,l} < \underline{X}^{j,b,l'}$ or $\overline{X}^{j,b,l} < \overline{X}^{j,b,l'}$, with $l < l'$, or both, which in turn implies $[\underline{X}^{j,b,l}, \overline{X}^{j,b,l}] \subset [\underline{X}^{j,b,l'}, \overline{X}^{j,b,l'}]$. In any case, there is *ladder* in the market with an arbitrary number of steps L and a given distance between steps $X(l') - X(l)$ for

any pair of segments in the market. This structure generates investment opportunities. Any investor with wealth Ω can construct a portfolio $\{a_{j,b,l}\}$, with $a_{j,b,l} \in [0,1]$ and $\sum_{j,b,l} a_{j,b,l} = 1$, using the online platform for more than one unit. He has an opportunity cost r as selling a unit may take several periods $k = 0, 1, ...$ In this setting, the tradeoff for the investor is straightforward. He can sell cheaper units faster, reducing the opportunity cost of liquidy Ω, or he can *climb the ladder* searching for better average demand prices.

Based on equation (13) we can construct a ranking. A location b with a surplus value $S(b)$ has a better position in the ranking compared with location b' if $S(b) < S(b')$. An investor is willing to pay for the raking S to curve information frictions. A lower score in the ranking implies that he can discriminate demand with high probability as these units have a lower average consumer surplus $X_m > X_s$. Note that, according to equation (11), the investor has incentives to pay for the ranking as units with high information frictions also have high arrival rates. So, investing in the lower part of the ladder implies reducing the opportunity cost of liquid funds, as he can sell these units rapidly at the cost of reducing the probability of accurate price discrimination. Finally, any investor has access to the online platform and uses a pricing strategy $X_{s,k}^{j,b,l}(x_{k-1})$, with $k = 0, 1, ...$ and $x_0 = \overline{X}^{j,b,l}$, given by equation (16).

As with any seller, the investor faces information frictions. However, he knows that a unit with a low ranking also has a low arrival probability. Thus, he *estimates* the following arrival rates, which has the form of a Poisson distribution:

$$f(k, X_{s,k}^{j,b,l}) = \frac{S(X_{s,k}^{j,b,l})^k e^{-S(X_{s,k}^{j,b,l})}}{k!}, \tag{17}$$

where the dependence of S on $X_{s,k}^{j,b,l}$ follows from equation (13) and $k!$ is the factorial of k. We estimated the arrival rate in equation (17) using our data and found interesting properties: i) $(1/2.7) < f(0, X_{s,k}^{j,b,l}) < 1$ and it is decreasing in S, which implies that there is at least a probability of $1/2.7$ of not selling the unit at the highest possible price for the segment, and high-quality lower-ranked units have a higher probability of not being sold in period $k = 0$, ii) for $k \geq 1$ it is increasing and concave in S, which implies that high-quality lower-ranked units sell at a lower pace. Notice remarkably that paying for the ranking allows the investor to use an estimated arrival probability with the same qualitative properties of the actual arrival rate: highly-priced market segments have low arrival rates and a low ranking that implies a low estimated arrival probability. This fact implies that the observable version of the investor's problem is affected by the same tradeoff described above for the case of perfect information. The lower steps of the ladder, with lower-priced units, sell faster, reducing the opportunity cost of liquid funds. The higher steps of the ladder are harder to sell, but yield a higher return for the investment strategy.

Given the above paragraph, the objective function of the investor is given by:

$$V = Max_{a_{j,b,l}} -P(S) - (1+r\sum_{a_{j,b,l}} f(0, X_{s,0}^{j,b,l})a_{j,b,l})\Omega + \sum_{a_{j,b,l}} a_{j,b,l} \sum_k f(k, X_{s,k}^{j,b,l}) X_{s,k}^{j,b,l} (1+r)^{-k}, \tag{18}$$

where $P(S)$ is the price of the ranking, and the investor pays for it only once. Later, we will depart from this assumption. As this problem is linear, any investor will choose a corner solution setting $a_{j,b,l}^* = 1$ to the segment of the market with the highest marginal return, which is defined by the triple $[j,b,l]^* \equiv **$ satisfying:

$$\sum_k f(k, X_{s,k}^{**}) X_{s,k}^{**} (1+r)^{-k} - rf(0, X_{s,0}^{**}) \geq \sum_k f(k, X_{s,k}^{j,b,l}) X_{s,k}^{j,b,l} (1+r)^{-k} - rf(0, X_{s,0}^{j,b,l}). \tag{19}$$

There are at least two interesting things to notice from the solution to the investor's problem: i) there is a maximal price for the ranking $\overline{P}(S)$, which defines the *value of information*. This threshold is given by the participation constraint $V \geq 0$. Formally:

$$\overline{P}(S) \equiv \sum_k f(k, X_{s,k}^{**}) X_{s,k}^{**} (1+r)^{-k} - (1 + rf(0, X_{s,0}^{**}))\Omega. \tag{20}$$

Note that the value of information equals the profits of the best option in the market. ii) The investment strategy has the form of a *cascade*: If the market is profitable (i.e.$\overline{P}(S,\Omega) > 0$), invest all your resources in the most profitable segment of the market X^{**}. Then, take all your profits $1 + \Omega + V$ and evaluate the market, excluding the previously selected segment. If the market is profitable (i.e., $\overline{P}(S, \Omega + V) > 0$), then invest $\Omega + V$ in the best option according to (18), excluding X^{**} from the menu of locations. Call the value of this project V_1. Continue until $\overline{P}(S, \Omega + \sum_n V_n) < 0$. This gives a planning problem abstracting from execution times between stages of the strategy n.

Of course, there are other pricing strategies for the value of information. Suppose the investor pays a fraction of the posted price if he sells the unit. Under this contract, he has T periods to use ranking every time he pays a fee (proportional to the price of the unit sold). For simplicity, we assume that each investment opportunity takes T periods. That is $k \leq T$. Then, the maximum value of the ranking under this pricing strategy $\overline{P}_1$ is given by:

$$1 + \overline{P}_1 = \sum_{t=1}^{nT} \sum_{j,b,l}^{\underline{j},\underline{b},\underline{l}} \frac{(1 + y(j,b,l,t))}{(1+r)^t}, \tag{21}$$

where n is the number of profitable locations under this pricing strategy (i.e., $\overline{P}(S, \Omega + \sum_n V_n) \geq 0$), and y is given by

$$y(j,b,l,t) = \frac{(1 + rf(0, X_{s,0}^{j,l,l,t}))}{\sum_k f(k, X_{s,k}^{j,b,l,t}) X_{s,k}^{j,b,l,t} (1+r)^{-k}}. \tag{22}$$

Note that $t \in \{T, 2T, ..., nT\}$ and $y(j, b, l, t)$ is the opportunity cost of location (j, b, l) that is executed at time t.

3.4 *Calibration Results*

This section is work in progress.

4. Conclusions

The findings of this study highlight the potential of machine learning (ML) to transform real estate investment strategies by improving predictive accuracy, optimizing portfolio allocations, and addressing market inefficiencies. By integrating big data and ML, particularly in analyzing unstructured data from online platforms, our model provides new insights into how investors can mitigate information asymmetry and better manage the trade-off between liquidity and price discrimination.

We identified key drivers of real estate prices, including unit size and neighborhood income, and demonstrated that ML-driven dynamic pricing models can capture these variables to enhance investment decisions. The model also reveals how ML tools can be used to identify arbitrage opportunities in real estate markets by analyzing the discrepancies between posted and demand prices.

The study suggests that future research could build on this framework by further exploring real estate price dynamics through advanced ML models. This would enhance the predictive power of price movements and inform dynamic portfolio allocation strategies, particularly for institutional investors. Additionally, investigating the market microstructure, information asymmetry, and the integration of algorithmic trading strategies could provide deeper insights into market efficiency and risk management in real estate finance. Finally, the interplay between public policies, transportation infrastructure, and temporal dynamics should be examined to understand their effects on spatial equilibrium and long-term investment strategies.

References

Aghion, Philippe and Bolton, Patrick and Harris, Christopher and Jullien, Bruno(1991):"Optimal learning by experimentation. " *Review of Economic Studies*, 58(4),621–654.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019): "Optuna: A Next-generation Hyperparameter Optimization Framework." *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Bertsimas, D., & Kallus, N. (2020): "From Predictive to Prescriptive Analytics," *Management Science*, 66(3), 1025-1044. []

Bordo, M. D., & Jeanne, O. (2002): "Boom-Busts in Asset Prices, Economic Instability, and Monetary Policy," *NBER Working Paper No. 8966*. []

Breiman, L. (2001): "Random forests." *Machine Learning*, 45(1), 5–32.

BRUNEEL-ZUPANC, C. (2020): "Learning in Real Estate Markets," *Journal of Economic Behavior and Organization*, 170, 1-14. [3, 4]

CHARPENTIER, A., ELIE, R., & HENRY, L. (2020): "Reinforcement Learning in Economics and Finance," *Machine Learning in Finance*, 90-120. [3, 4]

CLEVELAND, W. S. (1979): "Robust locally weighted regression and smoothing scatterplots." *Journal of the American Statistical Association*, 74(368), 829–836.

DURANTON, G., & PUGA, D. (2014): "The Growth of Cities," *Handbook of Economic Growth*, Vol. 2, 781-853. [3]

FAVILUKIS, J., MABILLE, P., & VAN NIEUWERBURGH, S. (2023): "Affordable Housing and City Welfare," *Review of Economic Studies*, 90, 293-330. []

FRIEDMAN, J. H. (2001): "Greedy function approximation: a gradient boosting machine." *Annals of Statistics*, 1189–1232.

GLOSTEN, L. R., & MILGROM, P. R. (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14(1), 71-100. []

GU, S., KELLY, B., & XIU, D. (2020): "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), 2223-2273. []

GUERRIERI, VERONICA AND SHIMER, ROBERT AND WRIGHT, RANDALL (2010):"Adverse selection in competitive search equilibrium " *Econometrica*, 78(6) 1823–1862

HENDERSHOTT, T., JONES, C. M., & MENKVELD, A. J. (2011): "Does Algorithmic Trading Improve Liquidity?" *Journal of Finance*, 66(1), 1-33. []

KOULOVATIANOS, CHRISTOS, MIRMAN, LEONARD, AND SANTUGINI, MARC (2010):"Investment in a Monopoly with Bayesian Learning " *Mimeo*

KUTNER, M. H., NACHTSHEIM, C. J., & NETER, J. (2004): "Applied Linear Statistical Models." 5th Edition, *McGraw-Hill*.

LUNDBERG, S. M., & LEE, S.-I. (2017): "A unified approach to interpreting model predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

MONTGOMERY, D. C., PECK, E. A., & VINING, G. G. (2021): "Introduction to Linear Regression Analysis." 5th Edition, *John Wiley & Sons*.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., DOROGUSH, A. V., & GULIN, A. (2018): "CatBoost: Unbiased Boosting with Categorical Features." *Advances in Neural Information Processing Systems*, 31, 6638–6648.

ROSENTHAL, STUART S. (2014): "Are Private Markets and Filtering a Viable Source of Low-Income Housing? Estimates from a "Repeat Income" Model" *American Economic Review*, 104 (2), 687–706.

APPENDIX A: STATISTICAL ANALYSIS OF THE DATASET

A.1 *Descriptive Analysis of Prices*

A.1.1 *Average Property Prices in Madrid by Neighborhood* Figure A.1 shows a map of Madrid's neighborhoods, where the color of each neighborhood represents the average property price per square meter, based on sales postings from Fotocasa for December 2021. The colors range from blue to yellow, with blue indicating lower prices and yellow indicating higher prices, with the scale ranging from 0 to 10,000 euros per square meter.

This visualization highlights the significant variation in property prices across different parts of Madrid. As seen in the map, the central neighborhoods tend to have higher property prices, which gradually decrease as one moves toward the outskirts of the city. This spatial distribution of property values is important for understanding the real estate market dynamics in Madrid, as well as for making informed decisions regarding property investments and urban development.

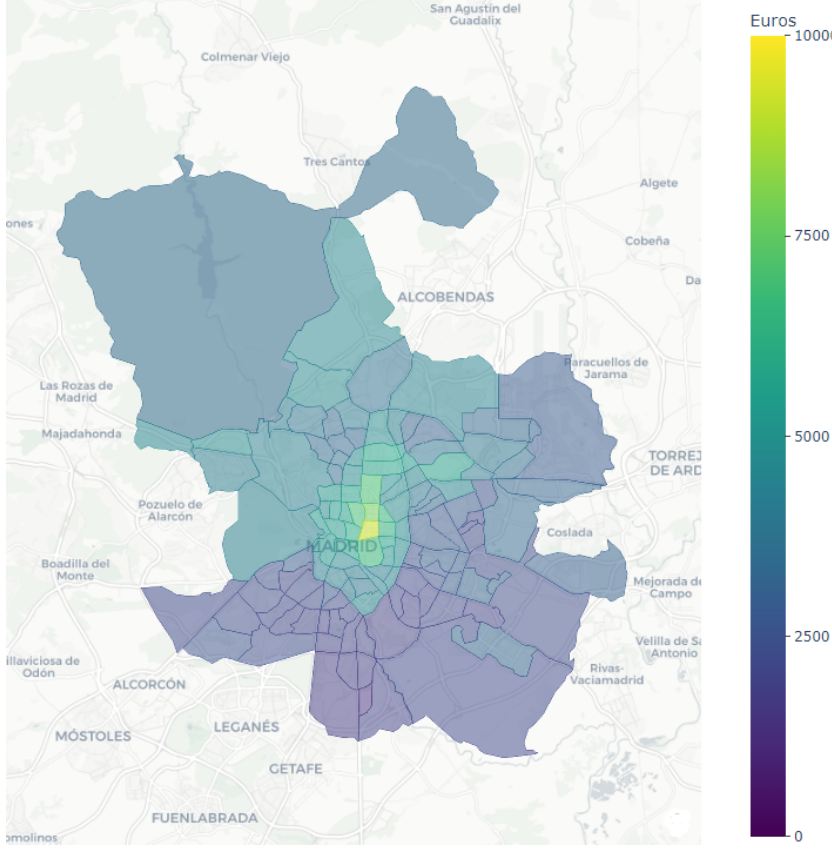


FIGURE A.1. Average property prices per square meter in Madrid by neighborhood, based on Fotocasa data from December 2021.

A.1.2 *Income Levels and Their Impact on Property Prices in Madrid* In addition to analyzing property prices across Madrid's neighborhoods, it's important to consider the role of income levels in shaping the real estate market. Figure A.2 presents a map of Madrid where each neighborhood is colored according to the average income per person. Similar to the property price map, the colors range from blue, indicating lower income levels, to yellow, representing higher income levels, with the scale ranging from 0 to 35,000 euros per year.

As expected, there is a strong correlation between income levels and property prices. Figure A.2 shows a scatter plot comparing the mean income per person against the mean property price per square meter across different neighborhoods. The positive correlation of 0.86 observed in the scatter plot highlights that neighborhoods with higher average incomes tend to have higher property prices. This relationship is particularly pronounced in central areas of Madrid, where both income levels and property prices are significantly higher compared to the outskirts of the city.

The spatial distribution of income levels and property prices underscores the socioeconomic divide within the city. Wealthier neighborhoods in the central and northern parts of Madrid, such as Salamanca and Chamartín, exhibit both high incomes and high property prices, making them some of the most expensive areas in the city. Conversely, neighborhoods on the periphery, particularly in the southern and southeastern regions, tend to have lower income levels and correspondingly lower property prices.

A.1.3 *Regression Model for Property Value Estimation: Feature engineering, advanced machine learning techniques, and hyperparameter optimization* In this section, we aim to develop a regression model to estimate the price per square meter of residential properties in Madrid. The model leverages a comprehensive set of features derived from various sources, including property characteristics, neighborhood income data, and proximity to amenities and points of interest.

The primary features used in the model can be grouped into several categories:

- *Property Characteristics*: This group includes variables that describe the physical attributes of the properties, such as their size, number of bedrooms and bathrooms, floor level, and the presence of amenities like a lift, garage, storage room, terrace, air conditioning, swimming pool, garden, and sports facilities. The condition of the property is captured by the `status` feature, and whether it is a new construction is indicated by the `new_construction` feature. We also include user engagement metrics from the real estate platform, represented by `rating_leads` and `rating_visits`.

- *Socioeconomic Characteristics*: Income data is a crucial determinant of property value. We incorporate average income per household and per person within the neighborhood (`Average Household Income` and `Average Personal Income`), which provide insights into the purchasing power of residents and, consequently, the demand for property in the area.

- *Proximity to Amenities and Points of Interest*: Using data from Google Places, we have calculated the minimum distance from each property to various types of

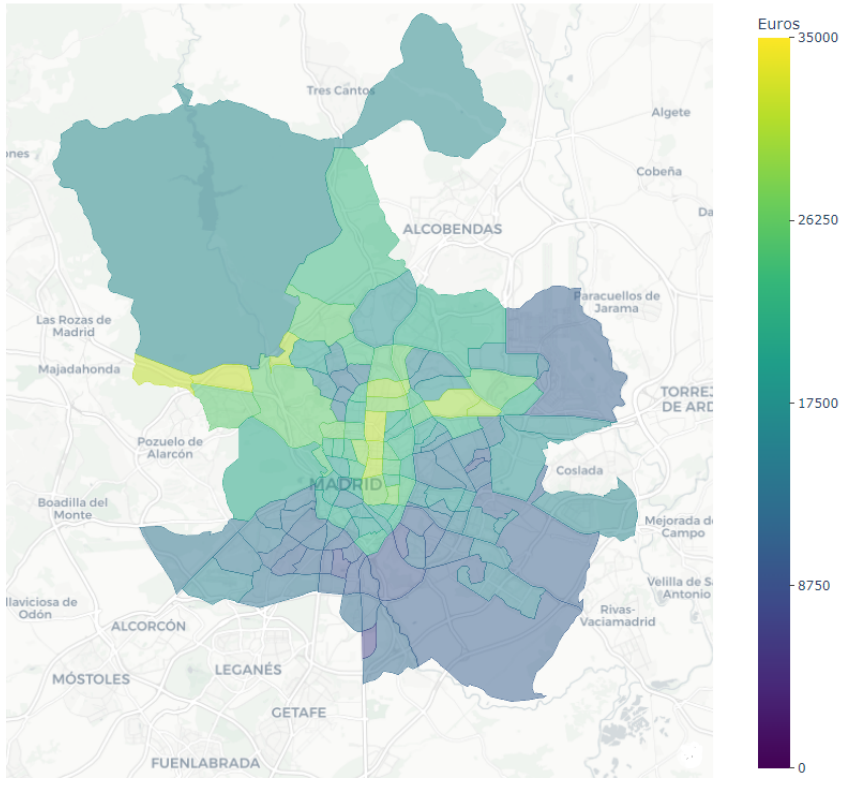


FIGURE A.2. Average income per person in Madrid by neighborhood. The data reflects the average income levels across the city, highlighting socioeconomic disparities.

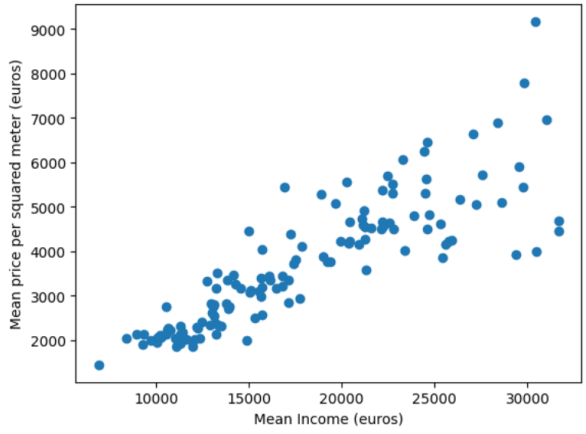


FIGURE A.3. Scatter plot showing the relationship between mean income per person and mean property price per square meter across Madrid's neighborhoods.

amenities, such as accounting offices, airports, schools, hospitals, parks, restaurants, and more (`distance_*` features). These distances help capture the convenience and accessibility of the property to important services and recreational activities. In addition to distance, we have also quantified the density of these amenities within a 1-kilometer radius (`count_*` features), reflecting the concentration of services and potentially enhancing the attractiveness of the property.

By integrating these diverse features into a regression model, we aim to gain a deeper understanding of the factors that drive property values in Madrid and to develop a predictive tool that can accurately estimate the price per square meter based on the characteristics of a given property and its surrounding environment.

In recent years, advanced machine learning algorithms such as gradient boosting methods have gained significant popularity in predictive modeling tasks due to their ability to capture complex, non-linear relationships within the data. One such method, CatBoost, is particularly effective in scenarios with a mix of categorical and numerical features and in handling datasets with high dimensionality and interaction effects (Prokhorenkova et al., 2018).

We used CatBoost, and to optimize the model, we employed Optuna, a hyperparameter optimization framework, to find the best combination of hyperparameters Akiba et al. (2019). The dataset was split into training (70%), validation (20%), and testing (10%) sets. The optimization process aimed to minimize the Root Mean Squared Error (RMSE) on the validation set.

After the optimization, the best hyperparameters were identified as:

- *Iterations*: 4535
- *Depth*: 10
- *Learning Rate*: 0.1
- *L2 Leaf Regularization*: 0.35
- *Subsample*: 0.754
- *Colsample by Level*: 0.835

The final model was trained using these hyperparameters on a combined dataset of the training and validation sets, and its performance was evaluated on the test set.

The model demonstrated strong predictive performance with the following metrics on the test set:

- *Test RMSE*: 420.73 euros per square meter
- *Test MAE*: 239.51 euros per square meter
- *Test R^2*: 0.953

The scatter plot in Figure A.4 compares the predicted property prices per square meter to the actual prices. The plot reveals a high degree of correlation between the predicted and actual values, as evidenced by the tight clustering of points along the diagonal line. This indicates that the model is accurately capturing the underlying patterns in the data.

The experiment successfully developed a model capable of predicting property prices per square meter with a high degree of accuracy. The combination of feature engineering, advanced machine learning techniques, and hyperparameter optimization proved effective in capturing the complex relationships within the data. This model can

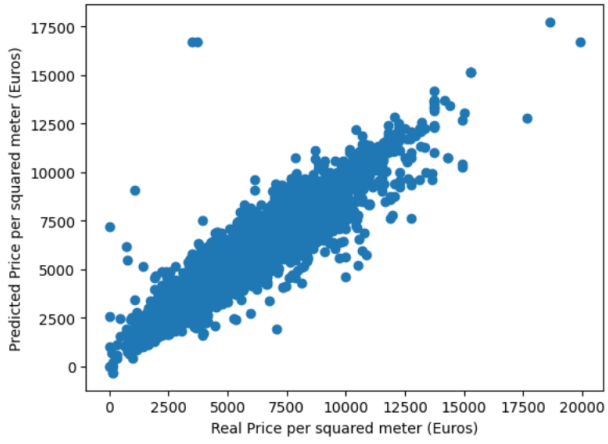


FIGURE A.4. Scatter plot comparing predicted versus actual property prices per square meter.

be further refined or integrated into larger systems for property valuation or real estate investment analysis.

A.1.4 *A Permutational Feature Importance Analysis* In order to gain insights into the most influential features affecting the prediction of property prices per square meter in Madrid, we conducted a Permutational Feature Importance analysis. This technique assesses the importance of each feature by measuring the impact on the model's predictive accuracy when the feature's values are randomly permuted.

The procedure was as follows:

1. We utilized the trained CatBoost model, which had been previously optimized using a detailed hyperparameter tuning process.
2. The Permutational Feature Importance was calculated on the test set. This involved permuting the values of each feature across all samples, one feature at a time, and measuring the decrease in model accuracy (measured by the mean decrease in accuracy) that resulted from this permutation.
3. The features were then ranked based on the mean decrease in accuracy, with higher values indicating greater importance.

The top 10 most important features, as determined by the Permutational Feature Importance analysis, are displayed in Figure A.5. The plot indicates that the most influential feature is *size*, followed by *Average Personal Income* and *counting_embassy*.

The main features obtained by this analysis were:

- *Size*: The most influential feature by a significant margin is the *size* of the property. It has the largest impact on the model's accuracy, suggesting that larger properties tend to have a direct and significant effect on price per square meter.

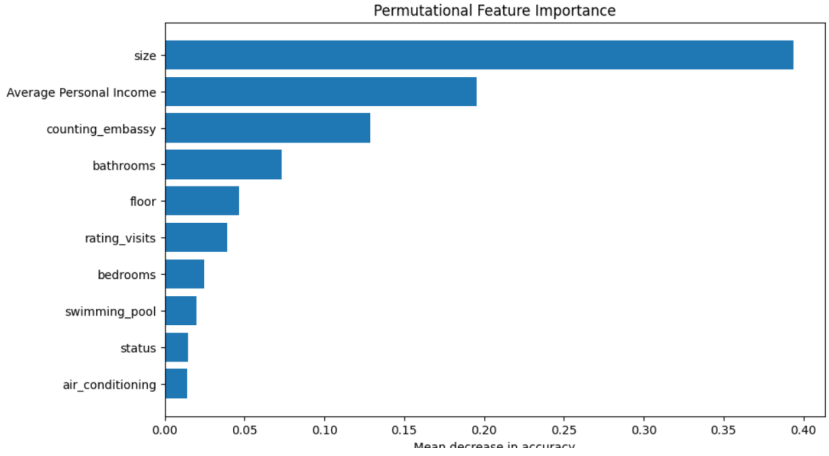


FIGURE A.5. Top 10 features by Permutational Feature Importance, measured by the mean decrease in accuracy.

- *Average Personal Income*: The second most important feature is *average personal income*, which reflects the economic conditions of the neighborhood where the property is located. Higher-income levels in the surrounding area likely correlate with higher property prices.
- *Counting Embassy*: This feature, representing the number of embassies near the property, also strongly influences property prices. Proximity to embassies may indicate desirable, prestigious neighborhoods.
- *Bathrooms and Floor*: Both *bathrooms* and *floor* are also important, as more bathrooms and properties on higher floors often correlate with higher desirability and, thus, higher prices.
- *Rating Visits*: The number of *rating visits*, which likely reflects the popularity or demand for a property, contributes notably to price prediction. Higher visit counts may indicate that a property is more sought after in the market.
- *Bedrooms and Swimming Pool*: The number of *bedrooms* and the presence of a *swimming pool* have a moderate impact. Properties with more bedrooms or a pool are typically considered more luxurious, contributing to their higher valuation.
- *Status, Air Conditioning*: Finally, *status* and the presence of *air conditioning* also affect property prices, albeit to a lesser extent. These features represent additional comforts that buyers may prioritize, especially in high-end properties.

These findings underscore the complexity of property valuation, where a mix of intrinsic property characteristics (e.g., size, number of bathrooms) and external factors (e.g., income level in the area and proximity to specific amenities) play a role in determining market prices.

Given that *size* emerged as the most important feature in the regression model for predicting the price per square meter, this analysis explores the relationship between these two variables across different neighborhoods. The scatter plot A.6 presents one point per neighborhood, where the x-axis represents the *median size* of properties in square meters and the y-axis represents the *median price per square meter*.

The **LOWESS curve** (in red) helps to identify the overall trend. Locally Weighted Scatterplot Smoothing (LOWESS) is a non-parametric regression method used to model complex relationships between variables without assuming a specific global functional form. The method fits multiple regressions on localized subsets of the data to produce a smooth curve that captures the underlying trend. By giving more weight to data points that are closer to the target point being estimated, LOWESS effectively balances flexibility and overfitting (Cleveland, 1979). It is useful wheneve a linear regression will not prove adequate enough.

We can observe the following:

- **General Trend**: The curve shows a rising trend at first, indicating that in neighborhoods with smaller median property sizes, the price per square meter tends to be higher. This suggests that smaller properties tend to be more expensive per unit of area, possibly due to demand for compact housing in denser, more central areas.
- **Diminishing Increase**: As the median property size increases beyond a certain point (around 100 m²), the price per square meter continues to rise but at a much slower rate. In some cases, it even flattens out, suggesting that larger properties may offer diminishing returns in terms of price per square meter.
- **Outliers**: Some neighborhoods, such as *Recoletos* and *Castellana*, stand out with notably high prices per square meter, whereas others like *Fuencarral* and *Fuente la Reina* show larger property sizes but with relatively lower prices per square meter.

This pattern is consistent with typical real estate market dynamics, where smaller properties, especially in highly sought-after areas, tend to have a premium price per square meter. Larger properties, particularly those in less central or less affluent neighborhoods, tend to be priced lower on a per-square-meter basis due to reduced demand or different market segments.

Thus, the decision to explore the relationship between *size* and *price per square meter* is validated by the observation that *size* indeed plays a significant role in determining the price per unit area, but its influence varies across neighborhoods and size ranges.

A.1.5 *Property Clustering* A clustering analysis was performed on the dataset of properties, focusing on intrinsic characteristics of each property. The goal of the clustering was to group properties with similar features, allowing for the identification of patterns in the types of properties available in the market.

The features used for the clustering process were as follows:

- *Size*: The total size of the property in square meters.
- *Bedrooms*: The number of bedrooms in the property.

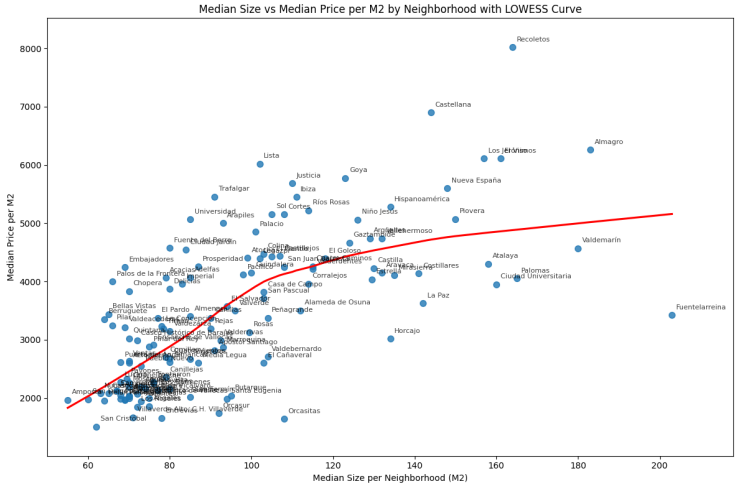


FIGURE A.6. Median Size vs Median Price per M2 by Neighborhood with LOWESS Curve

- *Bathrooms*: The number of bathrooms.
- *Lift*: Whether the property includes access to a lift.
- *Garage*: Availability of a garage.
- *Storage*: Whether the property has a dedicated storage area.
- *Terrace*: Availability of a terrace.
- *Air Conditioning*: Whether the property has air conditioning installed.
- *Swimming Pool*: Presence of a swimming pool.
- *Garden*: Whether the property has access to a garden.
- *Sports Facilities*: Availability of sports facilities, such as a gym or sports courts.
- *Rating Leads and Visits*: The demand for the property, measured by the number of leads and visits it received.
- *Floor Level*: The floor on which the property is located, encoded as an integer value.

A clustering algorithm was applied using these features, segmenting the properties into **five clusters**. Each cluster represents a group of properties that share similar intrinsic characteristics. This clustering approach enables a deeper understanding of the property market by categorizing properties based on their structural attributes and amenities, as well as their demand levels, offering insights into the different housing market segments. The clustering process was based on the intrinsic characteristics of the properties, and we classified each neighborhood according to the predominant cluster observed within it. Additionally, we analyze the distribution of property prices in each cluster using boxplots.

The neighborhoods in Madrid were clustered based on property features such as size, number of bedrooms, bathrooms, and amenities (e.g., lift, garage, swimming pool). The clustering process resulted in five distinct clusters (0 to 4), and each neighborhood was assigned to the cluster with the most properties. Figure A.7 shows a map of Madrid where each neighborhood is colored according to its predominant cluster.

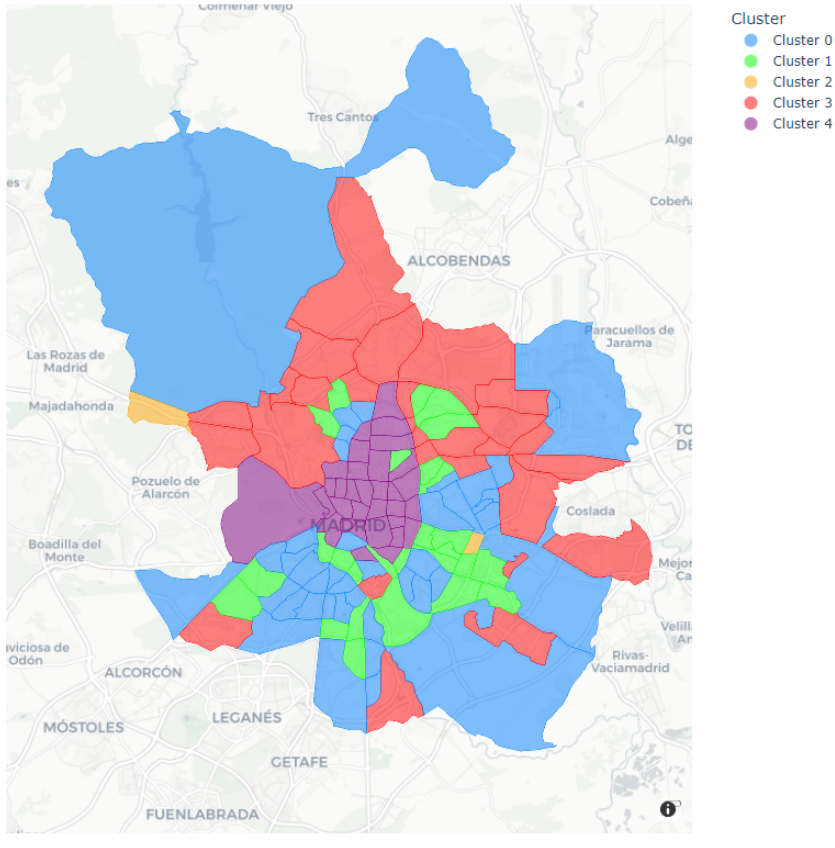


FIGURE A.7. Map of Madrid neighborhoods showing the predominant property cluster in each. Clusters range from 0 (blue) to 4 (purple).

To better understand the characteristics of each cluster, we examined the distribution of property prices within each cluster using boxplots. Figure A.8 illustrates the variation in property prices across the five clusters. The prices are represented in euros, and outliers have been excluded from this visualization to focus on the core price ranges in each cluster.

Characteristics of each cluster:

- *Cluster 0 (Blue)*: This cluster has the lowest median price, with a large concentration of properties in the 200,000 to 400,000 euros range. It likely represents areas

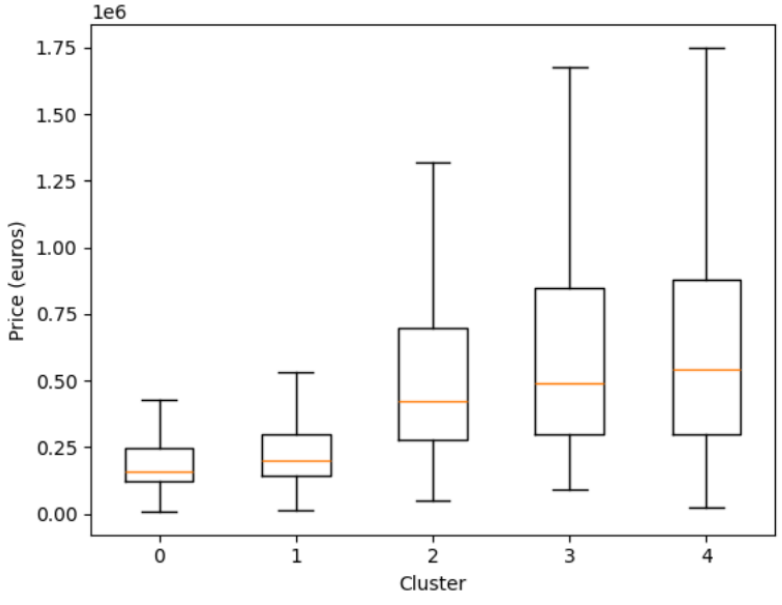


FIGURE A.8. Boxplots showing the distribution of property prices for each cluster.

with more affordable housing, possibly located further from central Madrid or in neighborhoods with fewer amenities.

- *Cluster 1 (Green)*: Properties in this cluster are slightly more expensive than those in Cluster 0, but they still remain within the affordable range. The median price is lower compared to other clusters, indicating that these neighborhoods may have slightly better infrastructure or be located in areas closer to public transportation.
- *Cluster 2 (Orange)*: This cluster exhibits a significant increase in property prices, with a wider distribution and a median price around 500,000 euros. These neighborhoods likely represent mid-market properties with better amenities and more central locations.
- *Cluster 3 (Red)*: Cluster 3 shows a broader range of property prices, with a median price approaching 1 million euros. These neighborhoods may be located in more prestigious or central areas of Madrid, offering access to premium amenities.
- *Cluster 4 (Purple)*: Cluster 4 contains the most expensive properties, with prices extending well beyond 1 million euros. The neighborhoods in this cluster are likely to be located in high-demand areas with luxury housing and extensive amenities.

The clustering analysis and the boxplots reveal distinct differences in property prices between clusters. The variation in price distributions is consistent with what would be expected based on the location and amenities of the properties. Cluster 0, with the lowest prices, is likely composed of less central and less well-equipped neighborhoods, whereas Cluster 4 represents the most expensive areas in Madrid.

Now, we analyze the mean and median values of key property features across the five clusters identified in the previous clustering analysis.

Table A.1 presents the mean and median values of key property features for each cluster, offering a detailed comparison across clusters.

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size (m²)					
Mean	82.15	81.30	156.51	174.81	369.87
Median	63.0	76.0	117.0	123.0	115.0
Bedrooms					
Mean	2.10	2.55	3.06	2.99	3.01
Median	2.0	3.0	3.0	3.0	3.0
Bathrooms					
Mean	1.14	1.32	2.19	2.32	2.08
Median	1.0	1.0	2.0	2.0	2.0
Lift (Presence)					
Mean	0.12	0.69	0.88	0.95	0.96
Median	0.0	1.0	1.0	1.0	1.0
Garage (Presence)					
Mean	0.02	0.08	0.58	0.76	0.11
Median	0.0	0.0	1.0	1.0	0.0
Storage (Presence)					
Mean	0.06	0.14	0.61	0.86	0.24
Median	0.0	0.0	1.0	1.0	0.0
Terrace (Presence)					
Mean	0.15	0.51	0.55	0.44	0.33
Median	0.0	1.0	1.0	0.0	0.0
Air Conditioning (Presence)					
Mean	0.24	0.46	0.72	0.73	0.70
Median	0.0	0.0	1.0	1.0	1.0
Swimming Pool (Presence)					
Mean	0.01	0.03	0.64	0.69	0.03
Median	0.0	0.0	1.0	1.0	0.0
Garden (Presence)					
Mean	0.00	0.00	1.00	0.00	0.00
Median	0.0	0.0	1.0	0.0	0.0

Table A.1.: Mean and Median values of key property features across clusters.

The data in Table A.1 reveals significant differences in the characteristics of properties across the clusters:

- *Cluster 0*: This cluster represents smaller properties, with an average size of 82 m² and a median of 63 m². These properties typically have fewer amenities, such as lifts (12% presence), garages (2% presence), and terraces (15% presence).
- *Cluster 1*: Properties in this cluster are similar in size to Cluster 0, but with a higher presence of amenities. Around 69% of the properties have lifts, and 51% have terraces, indicating improved features relative to Cluster 0.
- *Cluster 2*: The properties in Cluster 2 are notably larger, with an average size of 156 m² and a median of 117 m². Most of these properties have amenities such as swimming pools (64%), gardens (100%), and garages (58%).
- *Cluster 3*: Similar to Cluster 2, Cluster 3 includes properties with large sizes and extensive amenities. The average property size is 175 m², and a large proportion of properties feature garages (76%) and swimming pools (69%).
- *Cluster 4*: This cluster represents the most luxurious properties, with an average size of 370 m². While a lower percentage of properties have garages (11%) or swimming pools (3%) compared to Clusters 2 and 3, these properties are likely to be high-end based on other unlisted features.

The differences in property features, combined with the previously analyzed price distributions, highlight the distinct market segments each cluster represents. Cluster 0 likely corresponds to more affordable, smaller properties, while Clusters 2, 3, and 4 represent larger and more luxurious properties.