Cities 158 (2025) 105631

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

# Decoding the spatial dynamics of sales and rental prices in a high-pressure Portuguese housing market: A random forest approach for the Lisbon Metropolitan Area

Miguel Leal [a,b,c,*], Marina Carreiras [c], Sónia Alves [d,e]

[a] School of Science and Technology, University of Évora, Évora, Portugal
[b] Interdisciplinary Centre of Social Sciences, Pólo de Évora (CICS.NOVA.UÉvora), Évora, Portugal
[c] Centre of Geographical Studies and Associated Laboratory TERRA, Institute of Geography and Spatial Planning, Universidade de Lisboa. Lisbon, Portugal
[d] Institute of Social Sciences, Universidade de Lisboa. Lisbon, Portugal
[e] BUILD, Department of the Built Environment, Aalborg University, Aalborg, Denmark

A R T I C L E I N F O

*Keywords:*
Housing
Real estate
Sales prices
Rental prices
Random forest
Machine learning

A B S T R A C T

Sales and rental prices were analysed at parish level using random forest regression for the Lisbon Metropolitan Area. Three dependent variables (new sales, new rents, and all rents) and a set of independent variables/associated factors were used, including location, building/dwelling characteristics, socioeconomic features, and tourism. This geographically-based approach aims not to predict housing prices, but to identify relevant factors associated with sales/rents, ranking their importance. The temporal dimension is also explored by comparing new and all existing rents.

The results revealed similarities and differences between housing submarkets. New sales and new rents had similar spatial patterns and dynamics but were different from that of all rents, with different regulations over time. Strong associations were found between the dependent variables and the population's social status and urban quality. However, while location was more strongly related to new sales and new rents, revealing a greater dependence on the current dynamics of the housing market, socioeconomic features were more closely related to all rents, expressing the urban and demographic dynamics of recent decades. Different associated factors prevail inside and outside the Lisbon municipality. The results contribute to a better understanding of housing submarkets and the relationships between sales/rents and associated factors.

## 1. Introduction

While housing is part of everyday life and a necessity, housing markets are the result of decisions taken at different scales, and by different actors with different goals that jointly impact on several factors that affect prices. The special and contradictory nature of housing has been emphasized in the literature. Housing is "an essential commodity, fixed in geographic space, and a form of stored wealth that is subject to speculative activities in the market" (Knox & Pinch, 2014, p. 116).

Whitehead and Goering (2021) describe housing markets as "changing landscapes" that are the result of several demand-side and supply-side factors. Examples of demand-side factors include income growth, mortgage rates, availability of loans, level of wealth, population size and composition, cultural preferences and lifestyles, expected future house price increases, and migration rates, among others. Examples of supply-side factors include land availability, local amenities, inflation, and regulation. Land scarcity, for example, can lead to rising land prices that are unrelated to any investment or even efforts by the landowners; short-term inflation in construction costs (due to cost of materials and/or labour) can play a role in driving up the cost of building, reducing developers' profitability. Unless they can pass the costs on to buyers, this may result in less incentive to supply. The construction of housing is also influenced by the regulatory environment. Regulation of the rental submarket can restrict landlords' ability to raise rents or to change the tenure of dwellings, and regulation with respect to new building can affect aspects of tenure mix, housing costs, land availability for social housing, etc.

Even though, "like other markets, the amount and price of housing is determined by the interaction of demand and supply" (Ball et al., 2022, p. 404), housing markets are particularly complex as they comprise all

* Corresponding author at: Colégio Luis António Verney, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal
*E-mail address:* miguel.leal@uevora.pt (M. Leal).

https://doi.org/10.1016/j.cities.2024.105631
Received 30 August 2023; Received in revised form 28 June 2024; Accepted 30 November 2024
Available online 7 December 2024

the transactions of homes put up for sale, including existing homes, new homes, and conversions.

The urban housing market can thus be formed by thousands of properties and their attributes, including those of the dwelling itself (house type, size, age, conditions of kitchen, bathroom, etc.), of the building (plot size, age, etc.), and of the neighbourhood (related to location within the city and the region). Location is a key factor that influences residential property prices, determining not only the environment of the dwelling, but the sociodemographic environment and accessibility to services, such as jobs and educational and leisure facilities. The aggregate attributes of the housing stock (type of dwellings, type of tenancies, quality of housing), social characteristics (family, class, economic status, ethnicity) and neighbourhood reputation are all considered important house price determinants.

In addition, to understand the variation of housing prices across areas one must keep in mind that prices are both the outcome of the interactions of buyers and sellers in their respective markets, and are no longer determined only by local demand, but driven by national and international finance and investment (Whitehead et al., 2023). This mostly occurs in countries and cities where governments have developed a business-friendly finance-led environment through specific taxation regimes for foreign investors. This is definitely true for Portugal, which has been described as a very open country to investment markets, as it has offered strong incentives for foreign investment through investor visas, a non-habitual residential program, and an entrepreneurial urban planning approach to attracting private capital to rehabilitate buildings (Branco & Alves, 2020; Whitehead et al., 2023).

In a context of a rapid and sustained rise in housing costs for both owners and tenants (Eurofound, 2023, p. 9, Fig. 4), the choices that low- and middle-income families make about whether to rent or buy – and in which geographical areas of the ownership and rental submarkets to do so – are very much affected by the availability, quality and price of housing supply, as well as families' preferences and ability to pay (determined by income, borrowing costs, etc.). Household attitudes and behaviours towards homeownership and debt, as opposed to renting (Alves, 2022), are also influenced by government decisions – since, for example, tax and benefit systems affect different types of tenure.

In sum, the housing market differs from any other market, as housing has a unique set of characteristics that interact to create markets that are extremely sensitive to changes in demand and supply. In general prices rise when demand increases, but supply can only change slowly, as it takes time to build new housing.

Given the complex nature of housing markets, which require a considerable amount of (georeferenced) data along several dimensions to be better understood, innovative methods like machine learning techniques can be highly useful tools. Finding non-linear relationships and interactions between variables without requiring any data transformation, bringing together a wide range of variables, and detect the heterogeneity of housing markets are advantages of machine learning approaches (Lorenz et al., 2023; Ryo, 2022; Ryo & Rillig, 2017; Vaidynathan et al., 2023). For example, random forest (Breiman, 2001) is a machine learning algorithm which combines multiple decision trees that can be used for both classification and regression purposes. It is frequently used to model large spatial and spatiotemporal datasets with high dimensionality (Hengl et al., 2018). Random forest algorithms have been increasingly applied in several research fields, from natural to social, economic, and agricultural sciences. As recognized by Casali et al. (2022), housing is still an understudied topic and so more research is "urgently needed". Recently, random forest modelling and other machine learning approaches have been used in housing studies, proving its usefulness in improving on the results obtained by conventional methods such as linear models (Adetunji et al., 2021; Borde et al., 2017; Čeh et al., 2018; Choy & Ho, 2023; Dimopoulos et al., 2018; Ho et al., 2021; Hong et al., 2020; Hu, Huang, & Li, 2022; Levantesi & Piscopo, 2020; Lorenz et al., 2023; Mohd et al., 2019; Rico-Juan & de La Paz, 2021; Soltani et al., 2022; Truong et al., 2020; Waddell & Besharati-Zadeh, 2020; Wang & Wu, 2018; Yoshida et al., 2022; Zhou et al., 2019). However, these studies are mostly focused on forecasting housing prices, attributing great importance to the characteristics of buildings/dwellings and their proximity to local amenities, but often forgetting the socioeconomic features of geographical areas and populations. Despite its importance, the nature and volatility of the socioeconomic dimension entails additional difficulties in distinguishing between cause and effect in relation to housing prices. In this study, we do not aim to predict or forecast prices. Rather, we use a geographical perspective to uncover how the spatial patterning of housing relates to not just those factors traditionally understood to influence housing prices (location, neighbourhood, and characteristics of buildings and dwellings), but also the socioeconomic characteristics of a given area. To do this, a vast set of variables of several domains was used to identify which are relevant, or irrelevant, for sales and rental submarkets and for different areas. Knowing the factors that influence housing prices is essential for policymakers, investors, and homeowners (Vaidynathan et al., 2023), and the assessment of their individual importance has only been done in few studies (Čeh et al., 2018; Hong et al., 2020; Hu, Huang, & Li, 2022; Lorenz et al., 2023; Rico-Juan & de La Paz, 2021). In addition, the importance of variables can change depending on the characteristics of the study area (Grekousis et al., 2022).

Despite having been explored by a wide range of perspectives, the discussion on housing prices needs further development. Since few studies address both the sales and rental submarkets, this study aims to understand the spatial distribution of housing prices in these two submarkets in the Lisbon Metropolitan Area (LMA). These submarkets have been subject to different incentives over recent decades and have different spatial distributions, with the percentage of rented dwellings higher in urban areas. This study also considers the temporal dimension of the rental submarket, differentiating the prices of new contracts (made in a single year) and all existing contracts (made within the same year).

The significance of the LMA as a case study can be justified on several grounds. One is that this is a high pressure and rapidly changing housing market (as the empirical part of the paper will demonstrate). The other is that even though a substantial body of literature has addressed housing unaffordability issues in the LMA (Branco & Alves, 2020; Garha & Azevedo, 2021; Mendes, 2018), it has mostly been based on aggregated data, and as such has not helped uncover the patterns and tendencies in housing prices at a geographical, intra-urban level of analysis.

In an attempt to fill this knowledge gap, this study was designed to meet the following four specific objectives: 1) to detect the spatial patterns in housing prices in the LMA; 2) to identify factors relevant to sales and rental prices, and whether these factors change for two study areas (LMA outside Lisbon vs. Lisbon municipality); 3) to understand the importance of socioeconomic factors in the spatial distribution of sales and rental prices; and 4) to discuss the usefulness of random forest modelling for housing studies.

To do so, the paper is structured as follows: Section 2 reviews the existing theoretical framework on factors that influence housing prices. We bring together insights from literature on housing studies from within economics, urban geography, land-use planning, and other social sciences. Section 3 presents the case study and some features of the Portuguese housing stock and market. In Section 4 we detail the collected data and describe the methodological procedures used to explore the differences between housing submarkets (sales and rental) in the LMA and Lisbon municipality. Our results are reported in Section 5 and discussed in Section 6. Our research findings and conclusions are presented in the Section 7.

## 2. Literature review

Harvey, drawing on Marx, claimed that the production of the built environment within the capitalist profit-seeking logic, which continually aims at the realization of surplus value, has transformed housing into a commodity. As such, he predicted that this process would be intensified by capital switching to housing once the primary, productive circuit of profit is exhausted due to over-accumulation (Harvey, 1973). Aalbers (2017) and Cocola-Gant (2023) have also made the point that a wall of money has been deposited in housing markets, and that the neoliberal state has played a key role in enforcing laws that privilege the right to property over the right to housing, by deregulating financial markets and/or privatizing land.

Over the last decade the affordability of housing has become a major policy issue and increasingly a concern for governments as house prices have risen dramatically with respect to wages (Fingleton, 2008). Ryan-Collins (2018), who has critically examined the relationship between wages and conditions of access to credit and land, makes the point that whilst land is limited in supply and fixed in space, the deregulation and liberalization of advanced economy banking systems in the 1980s and 1990s has allowed the quantity of credit and investment to explode over time. The subsequent high liquidity has driven property prices at a much faster rate than rises in incomes, making housing increasingly unaffordable for a large proportion of the population.

Concern about the financialization of housing has increased in many developed countries and cities, particularly since the Global Financial Crisis, as the globalization of housing finance and property markets have allowed investors from across the world to take part in what were once narrow local housing markets (Whitehead et al., 2023, p. 17).

Whilst some note that the price increase has been greater in urban areas where supply is less elastic, for example where planning controls seek to protect historic building stock or open space from (re)development (Whitehead, 1999), others note that even in areas with large secondary housing stocks and with vacancies housing prices have consistently remained high (Hoekstra, 2009). Additionally, as emphasized by Ryan-Collins (2021), even though in the post-crisis period new macroprudential regulations repressed mortgage credit in some countries, monetary policy played an amplificatory role. The ultra-low medium-to-long-term interest rates generated by central bank quantitative easing programmes increased the attractiveness of real estate as a financial asset for capital market investment, as it came to be seen as a more profitable and safer financial investment.

Housing economists and geographers have sought to explain the variation in housing prices, or housing price formation, at the intra-urban level, based on variables such as use, proximity, location, and spatial competition (Oxley, 2004). The basic location theory, which assumes the existence of a central point of attraction in a radial urban area, predicts that a rent gradient will develop, with the highest rents per square meter in the centre of town and the lowest at the edge. It also predicts that land use will be determined by the highest monetary bid for a given location, and that higher incomes and more competition for property within an urban area will increase property values in general throughout the area (Oxley, 2004, p. 37). But Oxley (2004, p. 37) also points out that the claim that residential land values and house prices are a function of distance from places that households wish to reach is complicated by the inclusion of additional considerations including: 1) households' location preferences being influenced by factors other than work (e.g., shops, schools and recreational facilities); 2) work and other points of attraction being in non-central locations and there being several points of attraction (i.e., a polycentric spatial structure); 3) households competing for locations with other market-orientated users such as commercial firms; and 4) market-orientated residential, industrial and commercial users competing with public sector land uses such as town halls, parks, schools and hospitals (pp. 38–39).

Wittowsky et al. (2020) claim that two types of factors are important for price formation. There are those that are correlated with the dwelling itself (house type, size, age, plot size) and those that are correlated with the location and the surrounding area (neighbourhood-related factors, such as inhabitants' socioeconomic characteristics, green space or parks, distance to the central business district, etc.).

Adair et al. (1996, p. 67) claim that the housing market is not a uniform entity but is rather a set of distinctive submarkets arising from structural and locational attributes, i.e., from dwelling characteristics and neighbourhood characteristics. Oxley (2004) argued that is useful to distinguish the buying/selling submarket from the rental submarket. The market price of a house refers to “the sum of money for which the dwelling is exchanged between the buyer and the seller”. Additionally, the costs of paying loans, property taxes and maintenance may influence the sales value. The rental value refers to the price of occupying the house or consuming the services of the house over a given time of period, i.e., the rental payment (e.g., a monthly rent). Higher percentages of private ownership and uncontrolled rental are generally associated with higher housing prices. On the other hand, prices tend to be lower when there is a high proportion of social rented housing and controlled tenancies (for more details see Alves et al., 2023).

The concept of housing submarkets has been analysed as a working hypothesis by several authors who seek to understand the variety of processes that create market differentiation. For example, while Rey-Blanco et al. (2023) emphasize the importance of geography in explaining variations in real estate property prices at an intra-urban scale, Kauko et al. (2002), who define “housing market segmentation” as “the differentiation of housing due to income and preferences of the residents and administrative circumstances”, claim that various location-specific attributes cause segmentation of the housing market into submarkets. Kauko et al. (2002) make the point that “segmentation is real, when there are criteria that separate mutually distinctive submarkets from each other”, but also that “the question is, whether the most relevant partitioning criteria are directly related to the transaction price or to other, socioeconomic and physical features of the location” (p. 875).

## 3. Research framework

Prices of housing are cyclical, but influenced by features of the neighbourhood in which it is located, its construction date, and the conditions under which it was promoted (Nijman & Wei, 2020). In this section, we examine the Portuguese and LMA housing sales and rental markets[1] to provide the background that explains our empirical analyses (Section 5). To understand the fluctuation of housing prices, we reflect on the structure of housing ownership and tenure across administrative areas and, particularly, how this is related to contract duration, security of tenure and rental values.

### 3.1. A brief overview of Portuguese housing

In recent decades, the housing sector in Portugal has gone through different phases. From the 1980s until the end of the 20th century – in a context of macroeconomic changes characterized by the deregulation of capital markets, easing of financial conditions and public incentives for the acquisition of housing – suburbanization gathered pace, leading to a long-term rise in homeownership. In a context of low interest rates, the decision to direct most public incentives towards the construction and purchase of new homes, rather than to rehabilitation and renting, led to the decline of the rental sector in both quantitative and qualitative terms. Low rents resulted in disincentives to supply new rented accommodation, and in the neglect of existing rented dwellings. Support for home buyers accounted for 73 % of total Portuguese public investment in housing between 1996 and 1999 (IHRU, 2015) and encouraged indebtedness, which reinforced the financialization of the housing markets.

After the financial crisis of 2008/9, significant changes occurred in the Portuguese housing sector. In a period marked by an economic recession and austerity policies, the dynamism of the construction sector

[1] More detailed information on the Portuguese housing context can be found in several studies (e.g., Farha, 2017; Rodrigues, 2022).

declined, and access to mortgages became more difficult. On the other hand, rehabilitation, only occasional until then, increased, but it was highly driven by the tourism industry, targeting international demand made up of investors and consumers. In a context of high global housing demand by foreign investors, the Portuguese government triggered the rapid internationalization of the country's real estate market by launching tax incentives for non-habitual residents and Golden Visas. The growing demand for housing for secondary and seasonal use (Hall & Müller, 2018; Savills Research and HomeAway, 2018) became particularly visible in the Portuguese market – a fact linked to the country's attractiveness to tourists. The proportion of dwellings for seasonal use increased from 3 % in 1970 to 19 % in 2021 (for more details see Alves (2017)).

Both the tenure mix (ownership or rental) and the mix of forms of occupation (permanent, seasonal, or empty) have undergone important changes in recent decades. Even though the private rental market has benefited from rehabilitation and is thus more attractive, it became characterized by polarization in terms of types of contracts (their duration, rental price, security), the quality of the housing, and the characteristics of its occupants (see Section 5.1). On the one hand, there is a segment of old open-ended contracts (prior to 1990), which has suffered the cumulative effects of strict rent regulation and is consequently associated with low rental prices and strong security of tenure for sitting tenants who are over 65 years old, have high levels of disability or constitute low-income households. On the other hand, newer rental contracts are typically for 1 year and have low security of tenure.

It is also important to highlight the country's geographical diversity: there are significant differences between the structures of housing markets in different regions and parishes, in terms of the balance between tenure regimes (ownership, private and social rental markets), the characteristics of the built stock (buildings and dwellings) and the characteristics of neighbourhoods. These factors and their impacts on housing prices are analysed below in more detail for the LMA.

### 3.2. The LMA and the Lisbon municipality

This study takes the LMA as its main study area. It was chosen for several reasons: 1) it is one of the regions in Portugal most subject to pressure from housing prices; 2) the municipality of Lisbon – the Portuguese capital – has the highest prices in the country in the sales and rental submarkets; 3) it has a high concentration of people and administrative, financial, commercial and tourist activities; and 4) the municipality of Lisbon's economic, political and cultural importance extends to adjacent geographical areas, with the high prices in the centre diffusing out to other areas, even those outside the LMA.

The LMA is a Portuguese NUTS 2[2] that occupies 3015 km$^2$ and is composed of 18 municipalities and 118 parishes (Fig. 1). According to the 2021 Census, this region was home to 28 % of the country's population, 25 % of its dwellings, and 13 % of its buildings (Table 1). During the 2011–2021 period, the population (48,332 people; 1.7 %), number of dwellings (31,592; 7.5 %) and number of buildings (3625; 0.8 %) increased. In 2018, 22 % of the LMA's surface was occupied by built-up areas. Different socioeconomic and land use/land cover characteristics exist among the LMA's municipalities, and even within one and the same municipality, presenting a unity largely dictated by the centrality of Lisbon.

The municipality of Lisbon is the centre of the metropolitan area, occupying 100 km$^2$ (3 % of the LMA's surface) and comprising 24 parishes (Fig. 1). In 2018, 70 % of Lisbon's surface was covered by built-up areas. In 2021, Lisbon accounted for 19 % of the LMA's population, 21 % of the LMA's dwellings and 11 % of the LMA's buildings (Table 1). Contrary to what happened at the level of the LMA, Lisbon's population (−6904 people; −1.3 %), dwellings (−6649; −2.0 %) and buildings (−3473; −6.6 %) decreased during the 2011–2021 period. Lisbon stands out as an urbanized and privileged municipality in terms of accessibility, infrastructure, commerce and services, leisure facilities and culture, while being more subject to urban and real estate pressure.

Analysis of the tenure structure of housing markets in recent decades in Portugal, the LMA and Lisbon reveals different ratios between renters and owners (Table 2). Rates of ownership or co-ownership in the LMA have been lower than the Portuguese average, in contrast to rates of renting and sub-renting. In 2021, ownership/co-ownership accounted for 64 % and 50 % of dwellings in the LMA and Lisbon municipality, respectively. At national level, the rate was 70 %. Meanwhile rental rates reached 29 % for the LMA and 42 % for Lisbon municipality, while the national average was 22 %. The rental submarket in Lisbon has remained significant despite the continuing predominance of homeownership in the country as a whole and historical housing acquisition policies. The area's diversity and attractiveness may contribute to this, as it boosts the demand for housing, increasing the profitability of accommodation placed on the rental submarket and consequently the use of property as a financial asset. However, inside Lisbon's boundaries, distinct realities can be identified.

Lisbon's historic centre, which is characterized by a higher percentage of buildings built before 1945, has a greater proportion of old contracts with low rents, unlike in more peripheral areas, which have a more recent housing stock and a high rate of ownership. Ownership/co-ownership ranged between 70 % in Olivais and 24 % in Santa Maria Maior, with the highest values recorded in the most peripheral parishes of the city (Belém, Benfica, Carnide, Lumiar, Olivais, Parque das Nações and São Domingos de Benfica, all above 50 %), where urban growth occurred later. The lowest rates occur in some of the more central parishes (Misericórdia, Santa Maria Maior and São Vicente). Real estate development for rent, which was significant in older and central urban areas, remains less significant in more peripheral parishes. This fact should be taken into consideration if we are to better understand different housing dynamics, such as the different levels of availability of dwellings for sale and rent, and housing demand.

## 4. Data and methods

### 4.1. Data selection

Our data collection and analysis took the parish as the unit of analysis, as it is the smallest geographical unit for which there are available and comparable data. Two study areas are considered: 1) all parishes of the LMA with data available for both sales and rental values, except those that belong to the municipality of Lisbon ($n = 76$; 85 % of the parishes in this area) (LMA without Lisbon); and 2) Lisbon's parishes ($n = 24$; 100 % of the parishes of this municipality). These study areas are distinguished to better discuss the spatial dynamics inside and outside the municipality of Lisbon.

The data collected aimed both to characterize the current housing situation and to enable us to implement the random forest models. The median sales values per m$^2$ (*Sales_m2*), the median rental values per m$^2$ (*Rents_m2*), and the average monthly rents (*Rents*) were the three dependent variables considered; they were used to describe these sales and rental submarkets. *Sales_m2* and *Rents_m2* correspond to the values of transactions/contracts carried out in the year 2021, while *Rents* covers all values in 2021. Thus, the first two dependent variables cover the new transactions/contracts, and the third variable considers all the existing rental contracts (until 2021), allowing the inclusion of older contracts, in some cases tens of years old. The remaining variables used in this study and their respective sources can be found in Table 3. This set of variables includes information about different dimensions: urban and population dynamics (including characteristics of the built environment and population); mobility (commuting patterns and preferences regarding means of transport); migration (migration dynamics);

[2] Nomenclature of Territorial Units for Statistics.

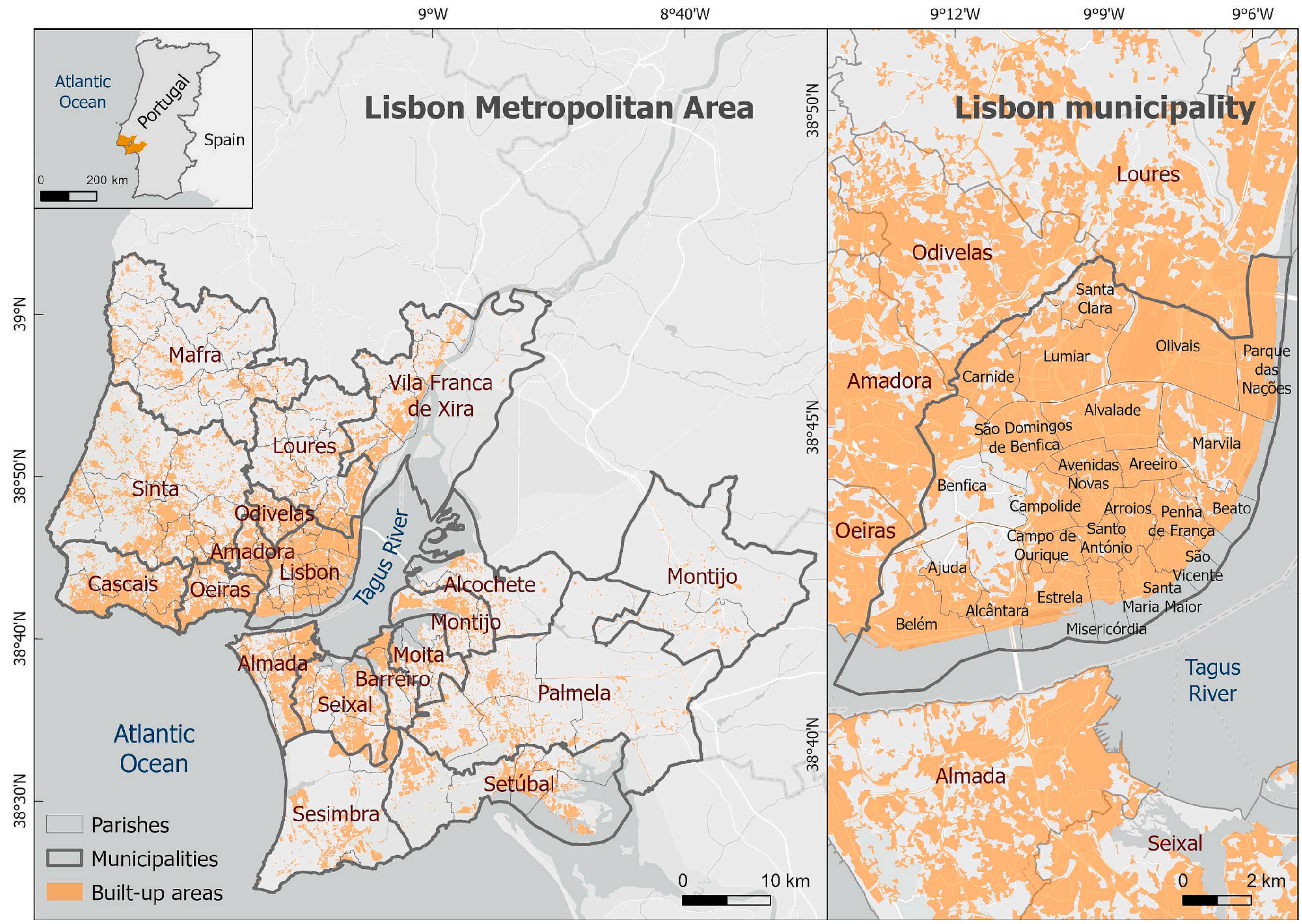


**Fig. 1.** Lisbon Metropolitan Area and Lisbon municipality.

**Table 1**
Population, dwellings, and buildings in the LMA and Lisbon municipality.

| Study area | Population | Dwellings | Buildings |
|---|---|---|---|
| LMA | 2,870,208 | 1,499,047 | 452,582 |
| Lisbon municipality | 545,796 | 320,143 | 49,223 |

**Table 2**
Proportion of dwellings by tenure status in Portugal, LMA and Lisbon.

| Tenure status | Area | 2001 | 2011 | 2021 |
|---|---|---|---|---|
| Rental market | Portugal | 20.9 | 19.9 | 22.3 |
| | LMA | 29.4 | 27.3 | 29.2 |
| | Lisbon | 48.6 | 42.3 | 42.3 |
| Ownership or co-ownership | Portugal | 75.7 | 73.2 | 70.0 |
| | LMA | 67.8 | 66.8 | 63.7 |
| | Lisbon | 47.9 | 51.8 | 50.3 |

Source: Statistics Portugal, Census 2001, 2011 and 2021.

buildings (data on density and conservation state of buildings); dwellings (size, age, type of occupation); tourism (density of short-term tourist accommodation and tourist facilities); location, accessibilities and facilities (location coefficients, existence of and distance to hospitals, rail, and road networks).

The use of quantitative variables from secondary sources associated with consolidated geographical units (parishes) has several limitations. Three are highlighted below. First, parishes are non-homogeneous geographical entities. Thus, the variables are represented by average values, and do not reflect, in several cases, these parishes' complexity and diversity, namely in terms of population and urban dynamics. For example, *Rents* (average monthly rents) do not reflect the bipolarity of rent prices, hiding the tension between those households who have been in the rental submarket for several decades and those who have recently entered it. Second, these data do not give us a direct insight into issues related to the regulation of housing market and the macroeconomic context. Third, the available information is often sparse or selective.

### *4.2. Methods*

Spatial analysis operations and mapping were performed in Geographic Information Systems (GIS) using ArcGIS Pro software.

Six random forest models (two for *Sales_m2*, two for *Rents_m2*, and two for *Rents*, considering the two study areas presented in Section 4.1) were built in RStudio software (randomForest package) to establish relationships between each dependent variable and 81 independent variables or associated factors (Table 3). We intended to identify the factors/variables (i.e., independent variables) that are related to the spatial behaviour of the three dependent variables and identify their individual importance. The elimination of irrelevant independent variables is the first step to improve the performance of the random forest models. Feature selection allows relevant variables to be kept and those that do not contribute to, or impair, the models' quality discarded (Gregorutti et al., 2017). This was done through an iterative process of discarding the variables with negative values of mean squared error (MSE), which is a measure of variable importance: the higher the Percentage Increase in MSE (%IncMSE), the higher the importance of a given variable

**Table 3**
List of variables used and the respective characteristics.

| Acronym | Variable name | Year/ Period | Unit | Source | Average [max; min] |
|---|---|---|---|---|---|
| **Dependent variables** | | | | | |
| ***Sales_m2*** | Median sales value per m$^2$ | 2021 | Euros | 1) | 2141 [5435; 980] |
| ***Rents_m2*** | Median rental value per m$^2$ of new contracts | 2021 | Euros | 1) | 8.43 [13.33; 4.8] |
| ***Rents*** | Average monthly rents (includes all existing contracts until 2021) | 2021 | Euros | 1) | 379 [669; 187] |
| **Urban and population dynamics** | | | | | |
| *BuiltUp* | Built-up areas | 2018 | % | 2) | 46.0 [98.6; 1.7] |
| *PopDens* | Population density | 2021 | Nr./km$^2$ | 1) | 3646 [17,017; 7] |
| *VarPop* | Population change variation | 2011–2021 | % | 1) | 2.3 [32.2; −26.0] |
| *SizeHouseholds* | Average size of households | 2021 | Nr. | 1) | 2.4 [2.8; 2.0] |
| *Overcrowded* | Proportion of households in overcrowded classic family dwellings | 2021 | % | 1) | 6.2 [12.7; 2.9] |
| *IndivHouse* | Proportion of population residing in a house | 2021 | % | 1) | 30.9 [94.2; 0.7] |
| *Young* | Proportion of young population (< 15 years) | 2021 | % | 1) | 14.3 [19.3; 8.4] |
| *Elderly* | Proportion of elderly population (≥ 65 years) | 2021 | % | 1) | 21.9 [38.3; 13.0] |
| *LiveAlone* | Proportion of population living alone | 2021 | % | 1) | 27.6 [45.9; 17.2] |
| *LiveAlone2564* | Proportion of population aged 25 to 64 living alone | 2021 | % | 1) | 14.3 [27.4; 8.2] |
| *LiveAlone65* | Proportion of elderly population living alone | 2021 | % | 1) | 12.9 [22.0; 7.6] |
| *NoEducat* | Proportion of population (> 15 years) with no completed level of education | 2021 | % | 1) | 4.6 [12.6; 2.2] |
| *HigherEducat* | Proportion of population with higher education | 2021 | % | 1) | 27.0 [61.1; 8.5] |
| *ActivePop* | Proportion of active population | 2021 | % | 1) | 47.9 [54.5; 39.2] |
| *Employed* | Employment rate | 2021 | % | 1) | 51.2 [60.0; 39.3] |
| *SocialSupport* | Proportion of population (> 15 years) with social support (unemployment benefits, social integration income, illness benefits, maternity license) | 2021 | % | 1) | 4.7 [10.8; 2.6] |
| *YearsResid* | Average years of residence of the household | 2021 | Nr. | 1) | 18.3 [25.4; 12.2] |
| *OneYearResid* | Proportion of population that a year before lived in another geographical unit or abroad | 2021 | % | 1) | 13.1 [23.0; 8.3] |
| **Mobility** | | | | | |
| *WorkParish* | Proportion of population (< 65 years) whose place of work/study is the parish where they live | 2021 | % | 1) | 20.3 [42.3; 10.0] |
| *WorkOtherParish* | Proportion of the population (< 65 years) whose place of work/study is another parish in the municipality where they live | 2021 | % | 1) | 21.1 [46.5; 4.1] |
| *WorkOtherMunic* | Proportion of population (< 65 years) whose municipality where they work/study is different from the municipality where they live | 2021 | % | 1) | 29.6 [47.2; 8.3] |
| *Commuting* | Average duration of commuting | 2021 | Minutes | 1) | 24.8 [33.2; 18.5] |
| *TransportWalk* | Proportion of the population that walks as main means of transport | 2021 | % | 1) | 16.0 [37.9; 5.0] |
| *TransportBike* | Proportion of the population using bicycle as main means of transport | 2021 | % | 1) | 0.6 [3.7; 0.1] |
| *TransportIndivid* | Proportion of the population using car/motorcycle as main means of transport | 2021 | % | 1) | 58.5 [80.0; 24.0] |
| *TransportCollectiv* | Proportion of the population using public/collective transport as their main means of transport | 2021 | % | 1) | 24.4 [39.9; 14.0] |
| **Migration** | | | | | |
| *ForeignPop* | Proportion of the resident population with foreign nationality | 2021 | % | 1) | 8.5 [33.3; 1.57] |
| *VarForeignPop* | Variation in the proportion of foreign population | 2011–2021 | % | 1) | 1.9 [15.5; −2.1] |
| *PopEnterPTWork* | Proportion of population that entered Portugal after 2010 with the purpose to work over the total foreign population | 2021 | % | 1) | 19.5 [57.3; 9.7] |
| *PopEnterPTEduc* | Proportion of population that entered Portugal after 2010 with purpose to access education over the total foreign population | 2021 | % | 1) | 9.4 [21.7; 1.9] |
| *PopEnterPTResid* | Proportion population that entered Portugal after 2010 to establish residence over the total foreign population | 2021 | % | 1) | 17.4 [32.8; 3.1] |

(continued on next page)

Table 3 (*continued*)

| Acronym | Variable name | Year/ Period | Unit | Source | Average [max; min] |
|---|---|---|---|---|---|
| *EuropePop* | Proportion of population from elsewhere in Europe | 2021 | % | 1) | 2.0 [7.5; 0.4] |
| *AfricaPop* | Proportion of population from Africa | 2021 | % | 1) | 1.9 [8.4; 0.1] |
| *NorthAmericaPop* | Proportion of population from North America (United States and Canada) | 2021 | % | 1) | 0.1 [0.6; 0] |
| *BrazilPop* | Proportion of population from Brazil | 2021 | % | 1) | 3.0 [6.6; 0.7] |
| *AsiaPop* | Proportion of population from Asia | 2021 | % | 1) | 1.4 [20.5; 0] |
| *VarEuropePop* | European population rate of change | 2011–2021 | % | 1) | 15.9 [331.0; −53.2] |
| *VarAfricaPop* | African population rate of change | 2011–2021 | % | 1) | 44.2 [800.0; −58.5] |
| *VarNorthAmericaPop* | North American population rate of change | 2011–2021 | % | 1) | 87.2 [1000.0; −100.0] |
| *VarBrazilPop* | Brazilian population rate of change | 2011–2021 | % | 1) | 73.4 [390.2; −41.5] |
| *VarAsiaPop* | Asian population rate of change | 2011–2021 | % | 1) | 631.5 [30,750.0; −33.3] |
| **Buildings** | | | | | |
| *BuildFloors* | Average number of floors per building | 2021 | Nr. | 1) | 2.8 [7.2; 1.1] |
| *BuildRepair* | Proportion of buildings in need of repair | 2021 | % | 1) | 36.1 [63.1; 13.3] |
| **Dwellings** | | | | | |
| *VarDwellings* | Dwellings rate of change | 2011–2021 | % | 1) | 1.3 [14.9; −28.4] |
| *LicensedDwellings* | Density of licensed dwellings in new housing constructions | 2021 | Nr./km$^2$ | 1) | 21.6 [131.1; 0.01] |
| *SizeDwellings* | Average surface of dwellings | 2021 | m$^2$ | 1) | 100.4 [141.3; 68.3] |
| *Dwellings100y* | Proportion of dwellings in buildings with 100 years or older | 2021 | % | 1) | 3.2 [54.5; 0] |
| *Dwellings21cent* | Proportion of dwellings in buildings of the 21st century | 2021 | % | 1) | 18.7 [56.8; 2.9] |
| *NumbRooms* | Average number of rooms per dwelling | 2021 | Nr. | 1) | 4.5 [5.2; 3.9] |
| *Dwellings200m2* | Proportion of dwellings with 200 m$^2$ or more | 2021 | % | 1) | 5.8 [21.4; 0.5] |
| *Dwellings30m2* | Proportion of dwellings with <30 m$^2$ | 2021 | % | 1) | 2.1 [10.4; 0.6] |
| *DwellingsSeason* | Proportion of dwellings with seasonal use | 2021 | % | 1) | 10.5 [59.3; 4.0] |
| *DwellingsEmpty* | Proportion of empty dwellings | 2021 | % | 1) | 11.6 [33.1; 5.1] |
| *DwellingsRented* | Proportion of rented dwellings | 2021 | % | 1) | 28.9 [69.0; 8.0] |
| *ChargesHousing* | Average monthly charges due to home ownership (mortgage payments) | 2021 | 2021 Euros | 1) | 416 [620; 271] |
| *RentSupport* | Proportion of dwellings with rental support | 2021 | % | 1) | 5.1 [29.0; 0] |
| *ParkingGarage* | Proportion of dwellings with parking space or garage | 2021 | % | 1) | 40.4 [84.7; 5.4] |
| *CentralHeat* | Proportion of dwellings with central heating | 2021 | % | 1) | 7.9 [32.5; 1.7] |
| *NonCentralHeat* | Proportion of dwellings with noncentral heating | 2021 | % | 1) | 58.2 [77.2; 44.1] |
| *NoHeat* | Proportion of dwellings with no heating | 2021 | % | 1) | 33.9 [53.3; 14.8] |
| **Tourism** | | | | | |
| *TouristAccom* | Density of short-term tourist accommodation (e.g., Airbnb) | 2021 | Nr./km$^2$ | 3) | 74.6 [1613.7; 0] |
| *TouristFacilit* | Density of tourist facilities (e.g., hotels) | 2021 | Nr./km$^2$ | 3) | 1.1 [46.3; 0] |
| **Location, accessibilities and facilities** | | | | | |
| *LocationMax* | Maximum location coefficient* | 2023 | Nr. | 4) | 1.9 [3.5; 1.0] |
| *LocationMin* | Minimum location coefficient* | 2023 | Nr. | 4) | 1.2 [2.7; 0.4] |
| *LocationMean* | Mean location coefficient* | 2023 | Nr. | 4) | 1.6 [3.1; 0.9] |

(*continued on next page*)

**Table 3** (*continued*)

| Acronym | Variable name | Year/ Period | Unit | Source | Average [max; min] |
|---|---|---|---|---|---|
| *CenterMinDist* | Minimum distance to the city centre (Marquês de Pombal square) | n.a. | Meters | n.a. | 12,732 [36,292–0] |
| *CenterMaxDist* | Maximum distance to the city centre (Marquês de Pombal square) | n.a. | Meters | n.a. | 18,994 [57,439; 1270] |
| *RestMichelin* | Existence of Michelin restaurants** | 2023 | Binary | 5) | 0.1 [1; 0] |
| *RestInsider* | Existence of restaurants in TheFork INSIDER category*** | 2023 | Binary | 6) | 0.3 [1; 0] |
| *TrainMinDist* | Minimum linear distance to the nearest rail network | 2023 | Meters | 7) | 1139 [14,782; 0] |
| *TrainMaxDist* | Maximum linear distance to the nearest rail network | 2023 | Meters | 7) | 5458 [24,004; 772] |
| *TrainStatMinDist* | Minimum linear distance to the nearest train station | 2023 | Meters | 7) | 1445 [15,014; 0] |
| *TrainStatMaxDist* | Maximum linear distance to the nearest train station | 2023 | Meters | 7) | 6066 [24,154; 1118] |
| *DensRoads* | Main road network density | 2023 | Nr./$km^2$ | 7) | 1.5 [7.4; 0] |
| *PubHosp* | Existence of public hospital | 2023 | Binary | 8) | 0.1 [1; 0] |
| *PubHospMinDist* | Minimum linear distance to the nearest public hospital | 2023 | Meters | 8) | 4690 [25,796; 0] |
| *PubHospMaxDist* | Maximum linear distance to the nearest public hospital | 2023 | Meters | 8) | 10,114 [42,575; 1149] |
| *PrivHosp* | Existence of private hospital | 2023 | Binary | 9) | 0.1 [1; 0] |
| *PrivHospMinDist* | Minimum linear distance to the nearest private hospital | 2023 | Meters | 9) | 5968 [31,305; 0] |
| *PrivHospMaxDist* | Maximum linear distance to the nearest private hospital | 2023 | Meters | 9) | 11,761 [44,402; 1321] |
| *PubPrivHosp* | Existence of public or private hospital | 2023 | Binary | 8), 9) | 0.2 [1; 0] |

1) Population and building Census, Statistics Portugal; 2) Land Cover Map, Portuguese Directorate-General for Territory; 3) Portuguese National Tourism Institute; 4) Portuguese Tax and Customs Authority; 5) the MICHELIN Guide; 6) TheFork Portugal; 7) Infraestruturas de Portugal, S.A.; 8) Portuguese National Health Service; 9) Portuguese Association of Private Hospitals; n.a. - not applicable.

* The location coefficient is one element of the formula for calculating the municipal tax on urban buildings. It considers accessibilities, proximity to social facilities, and public transports, ranging from 0.4 and 3.5.

** Restaurants awarded by the Michelin Guide (reference guide for high quality hotels and restaurants).

*** Trendiest and highest rated restaurants on the online platform TheFork.

(Friedman, 2001). When the variables with negative %IncMSE are removed, the explained variance increases, improving the predictive capacity of a model. The selected variables were validated by the results of the Boruta algorithm (Boruta package), which was also used in other studies for variable selection (Degenhardt et al., 2019; Speiser et al., 2019). So, we chose to collect and use, at an initial stage, as many independent variables as possible and then let the random forest and the Boruta algorithm choose which variables are relevant in each model. This means that the final set of independent variables used in each random forest model was different depending on the number of relevant variables. The train function (caret package) was used to adjust the mtry parameter to find out the optimal number of random variables used in each decision tree of the random forest model. The mtry parameter controls the randomness of the decision tree creation process. The random forest models were validated using a k-fold cross validation (5 folds) and errors were expressed by Mean Cross-Validation Error (MCVE). Lastly, the partial dependence plots show the marginal effect of each relevant independent variable on the predicted outcome of the model (Friedman, 2001), allowing the identification of positive and negative relationships between independent and dependent variables (Canadas et al., 2023; Ribeiro et al., 2021). Positive relationships mean that the values of both dependent and independent variables increase. Negative relationships mean that when the values of the independent variable increase, the values of the dependent variable decrease, or vice-versa.

## 5. Results

Following the brief overview of housing in Portugal and our methodology, we present and discuss our results in two stages. The first addresses the spatial patterns of sales and rental prices. The second focuses on determining the most relevant factors for the spatial distribution of *Sales_m2*, *Rents_m2* and *Rents*.

### 5.1. Values of housing sales and rents in the LMA and Lisbon municipality

In recent years, the values of housing prices in Portugal have significantly and consistently increased. Values in Lisbon and in the LMA are substantially higher than the national numbers, revealing the attractiveness of, and real estate pressure in, the country's main metropolitan area (Fig. 2). According to Statistics Portugal, between 2017 and 2022, the median value of sales per $m^2$ increased from €1264 to €1939 in the LMA and from €2334 to €3582 in the Lisbon municipality (Fig. 2A). Both represented a relative increase of 53 % in five years. In the same period, the median monthly house rental values per $m^2$ for new rental agreements grew from €6.70 to €9.86 (47 %) in the LMA and from €10.64 to €12.88 (21 %) in Lisbon (Fig. 2B). These values also reveal a widening of the gap between the LMA and Lisbon in sales and a narrowing of the gap in rents over the 2017–2022 period. Additionally, a break can be identified in Lisbon's rental values in 2020 and 2021, which may be a COVID-19 effect that is not significant for the LMA as a whole. In 2022, the highest values in the study area were recorded in the parish of Santo António: €5753/$m^2$ (sales) and €16.34/$m^2$ (rents).

A deeper analysis reveals similarities and differences among the

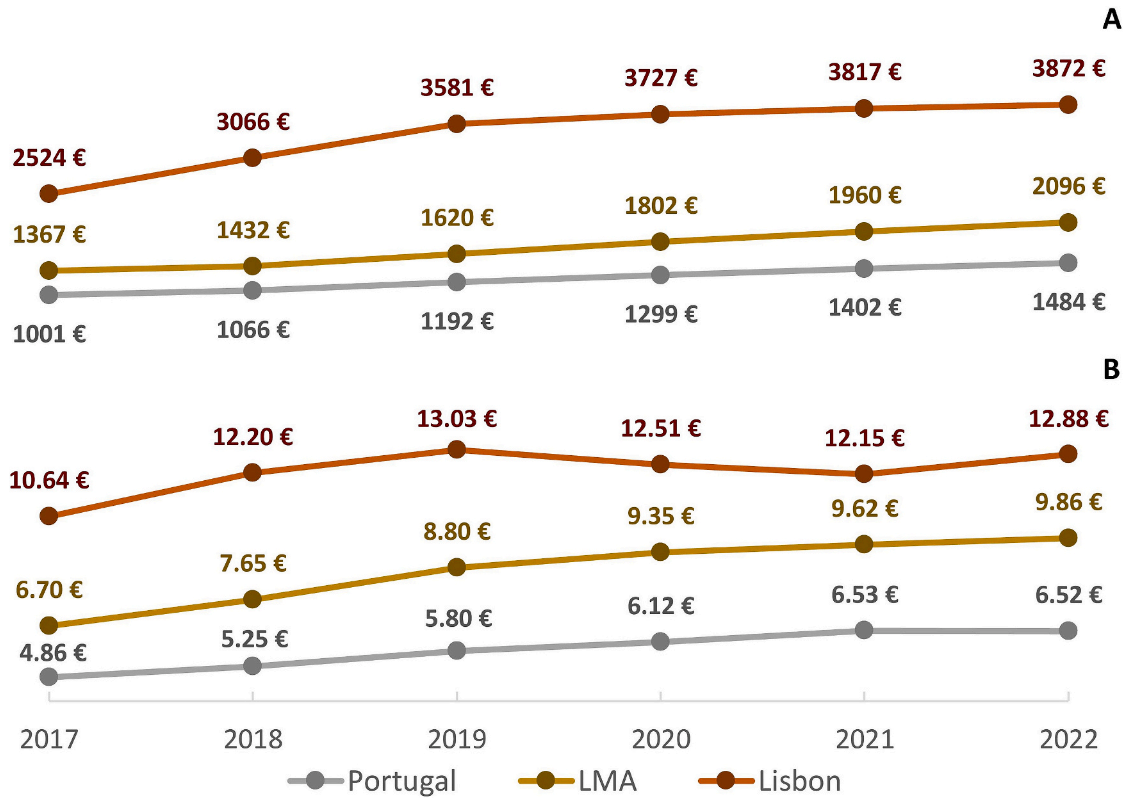


**Fig. 2.** Change over time of the median value of sales per $m^2$ (A) and the median monthly house rental values per $m^2$ of new rental agreements (B) during the 2017–2022 period in Portugal, LMA and Lisbon municipality. All values are presented in 2022 Euros.

three dependent variables used in this study (*Sales_m2*, *Rents_m2* and *Rents*, Table 3). The correlation coefficients between *Sales_m2* and *Rents_m2* for LMA without Lisbon ($r = 0.79$) and the parishes of Lisbon alone ($r = 0.85$) point to strong positive associations between these variables (Fig. 3A). The higher the *Sales_m2*, the higher the *Rents_m2*. This trend confirms that the values of new transactions and new rental contracts share similar dynamics in the LMA with both variables recording most of their highest values in the Lisbon parishes (Table 4 and Fig. 4). On the other hand, the correlation coefficients between new (*Rents_m2*) and all (*Rents*) rental contracts are much less significant: 0.58 and 0.52 for the LMA without Lisbon and the Lisbon parishes, respectively (Fig. 3B). This is expressed by the distinct spatial patterns of these two variables (Fig. 4), demonstrating noticeable differences in the rental submarket, especially inside the Lisbon boundaries. The contrast between *Rents_m2* (y axis of Fig. 3B) in the LMA and Lisbon is notable. This difference is much less noticeable for *Rents* (x axis of Fig. 3B), whose values are more highly dispersed in Lisbon (standard deviation = €129 in Lisbon vs. €78 in the LMA without Lisbon). Both the highest and lowest values of rents are recorded in the Lisbon's parishes.

The highest values of *Sales_m2* (≥€3000/$m^2$) and *Rents_m2* (≥€11/$m^2$) were concentrated in most of Lisbon's parishes and along the southern coast of the Cascais municipality (Fig. 4A and C), an area closely associated with occupation by upper classes. The difference between the parishes with the highest and the lowest values in the LMA was €4455/$m^2$ for *Sales_m2* and €8.53/$m^2$ for *Rents_m2* (Table 4). Taking a more detailed view of Lisbon, the highest values were recorded in the city center (Santo António), while the lowest values occurred in more peripheral parishes with less diversity of urban functions (Fig. 4B and D). There was a range of €2934/$m^2$ (*Sales_m2)* and €4.74/$m^2$ (*Rents_m2*) between the parishes with the highest and lowest values in Lisbon (Table 4).

The spatial pattern of *Rents* in the LMA has a less structured pattern with less spatial continuity when compared to the more concentrated patterns of *Sales_m2* and *Rents_m2*, which is mainly evident in Lisbon. Similar spatial patterns were found by Hu, He, and Su (2022). For *Rents*, there were fewer parishes in Lisbon and more parishes outside Lisbon included in the top category (≥€500), namely in the Mafra and Oeiras municipalities (Fig. 4E). There were four parishes with values greater than or equal to €600 in Lisbon, with Avenidas Novas standing out as the parish with the highest value: €669 (Fig. 4F). The difference between the parishes with the highest and the lowest values in the LMA was €483 (Table 4). A substantial contrast between *Rents* and *Sales_m2/Rents_m2* can be seen in certain parishes in the Lisbon municipality, which have low values even for the LMA context. The lowest value within the LMA was even found in a Lisbon parish: Marvila (€187). Part of the explanation for this reality may lie in two factors: the low values of social housing rents, and the low values of old rents (leasing contracts initiated prior to 1990).

According to Statistics Portugal, in 2015 Lisbon accounted for approximately half of the LMA's rented social housing units (49 %) and €82 was its average monthly rent for social housing. These units are not evenly distributed across the area. 25 % of social housing units in the city of Lisbon were in the parish of Marvila.[3] In 2021, rental contracts initiated prior to 1990 accounted for 20 % and 24 % of the total in the LMA and Lisbon, respectively. In these older contracts, rents up to €100 accounted for 39 % of these contracts in the LMA and 34 % in Lisbon, while rents up to €200 accounted for 72 % in LMA and 62 % in Lisbon. In contracts initiated after 1990, rents up to €200 represented only 15 % of these contracts in the LMA and 21 % in Lisbon.

[3] Source: Gebalis (the company that manages municipal housing in Lisbon). Data obtained from: https://lxhabidata.iscte-iul.pt/pt/data/

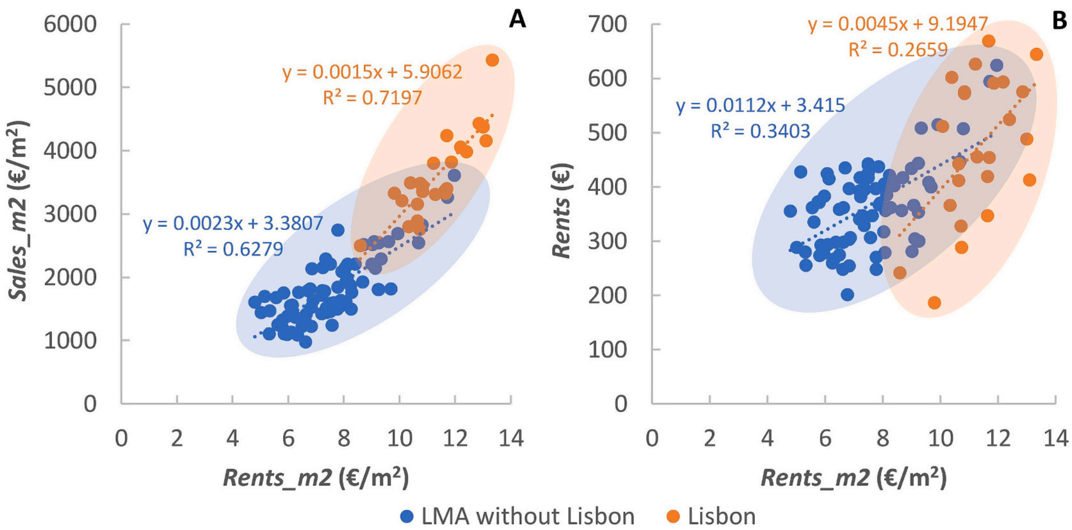


**Fig. 3.** *Sales_m2* vs. *Rents_m2* (A) and *Rents* vs. *Rents_m2* (B) in the parishes of the LMA and Lisbon in 2021.

**Table 4**
Descriptive statistics of *Sales_m2*, *Rents_m2*, and *Rents* in the LMA municipalities in 2021.

| Municipality | *Sales_m2* (€/m²) | | | | *Rents_m2* (€/m²) | | | | *Rents* (€) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | Max | Min | StdDev | Avg | Max | Min | StdDev | Avg | Max | Min | StdDev |
| Alcochete | 1732 | 2143 | 1321 | 411 | 6.84 | 6.84 | 6.84 | 0 | 351 | 397 | 310 | 36 |
| Almada | 1969 | 2548 | 1618 | 311 | 8.45 | 10.68 | 6.78 | 1.27 | 338 | 447 | 201 | 88 |
| Amadora | 1880 | 2214 | 1769 | 151 | 8.72 | 9.68 | 7.77 | 0.70 | 346 | 401 | 271 | 46 |
| Barreiro | 1213 | 1282 | 1092 | 73 | 6.91 | 7.57 | 6.32 | 0.51 | 289 | 307 | 268 | 16 |
| Cascais | 2994 | 3616 | **2521** | 466 | 10.49 | 11.96 | **8.70** | **1.38** | **511** | 624 | **409** | 99 |
| Lisbon | **3592** | **5435** | 2501 | **634** | **11.31** | **13.33** | 8.59 | 1.12 | 472 | **669** | 187 | **129** |
| Loures | 1812 | 2547 | 1100 | 400 | 7.31 | 9.24 | 5.33 | 1.35 | 323 | 444 | 255 | 66 |
| Mafra | 1565 | 2297 | 1024 | 361 | 6.89 | 9.33 | 5.79 | 1.32 | 388 | 509 | 303 | 61 |
| Moita | 1067 | 1110 | 980 | 61 | 5.93 | 6.62 | 5.31 | 0.54 | 260 | 293 | 220 | 28 |
| Montijo | 1355 | 1547 | 1096 | 190 | 6.63 | 6.63 | 6.63 | 0 | 305 | 363 | 249 | 47 |
| Odivelas | 2129 | 2207 | 1995 | 80 | 8.09 | 8.99 | 7.24 | 0.62 | 373 | 435 | 318 | 43 |
| Oeiras | 2565 | 2833 | 2208 | 209 | 9.44 | 10.79 | 8.40 | 0.83 | 415 | 515 | 282 | 88 |
| Palmela | 1348 | 1497 | 1217 | 117 | 6.54 | 8.25 | 5.62 | 1.21 | 350 | 417 | 275 | 52 |
| Seixal | 1540 | 1698 | 1355 | 125 | 6.62 | 7.50 | 5.15 | 0.93 | 386 | 428 | 330 | 43 |
| Sesimbra | 1995 | 2750 | 1554 | 536 | 6.47 | 7.78 | 5.56 | 0.95 | 345 | 425 | 248 | 73 |
| Setúbal | 1478 | 1788 | 1227 | 198 | 6.79 | 7.19 | 6.50 | 0.29 | 326 | 435 | 253 | 63 |
| Sintra | 1683 | 2294 | 1444 | 284 | 6.99 | 7.87 | 4.80 | 1.05 | 385 | 443 | 289 | 51 |
| Vila Franca de Xira | 1459 | 1720 | 1134 | 210 | 6.78 | 7.63 | 6.10 | 0.53 | 316 | 393 | 260 | 43 |

Notes: Number of parishes with data available in the LMA: a) *Sales_m2*–113; b) *Rents_m2*–100; and c) *Rents* – 118. Avg – average; Max – maximum; Min – minimum; StdDev – standard deviation.

### 5.2. *The associated factors of sales and rents*

The random forest results suggest associations as well as the contribution made by each independent variable to the model's predictive capacity. However, it is important to point out that a greater association does not necessarily mean causality – i.e., that the dependent variable is a consequence of the independent variables. This means that a given independent variable should not be considered necessarily as a driver of the dependent variables based on the random forest results. There may be a bidirectional effect, which it is not possible to assess with these models, in which these dependent and independent variables influence each other. For example, housing prices and income may influence each other, impacting the socioeconomic characteristics of a given parish. People with greater financial capacity seek attractive areas that meet their needs. Growing interest in these locations could lead to increased demand, driving up housing prices and driving away people with lower incomes. In turn, the presence of wealthier populations can generate investments and the emergence of other types of commerce and services, and contribute to the creation/improvement of infrastructure and local amenities. This chain makes these areas even more sought after, further increasing housing prices. For this reason, and because random forest models do not reveal causality, the independent variables used in this study are called associated factors and not determinants, drivers, predictors, or explanatory variables.

The six random forest models developed for the three dependent variables (*Sales_m2*, *Rents_m2* and *Rents*) and for two study areas (LMA without Lisbon and Lisbon parishes) are presented in Fig. 5. This shows the top 30 independent variables in terms of importance, whenever there are 30 variables identified as relevant. The green and red bars of this figure represent positive and negative associations with the dependent variables, respectively. The grey bars of this figure represent unclear trends (e.g., there are two or more trends associated with the increasing values of *Sales_m2*/*Rents_m2*/*Rents*). The positive, negative, and unclear trends were identified from the partial dependence plots. Some examples are shown in Fig. 6; these confirm the non-linear patterns of the relationships between some dependent and independent variables.

The first model (*Sales_m2* LMA without Lisbon – Fig. 5A) determined 25 relevant independent variables and was able to explain 73.1 % of the variance (MCVE = 75,019). The most important variables are

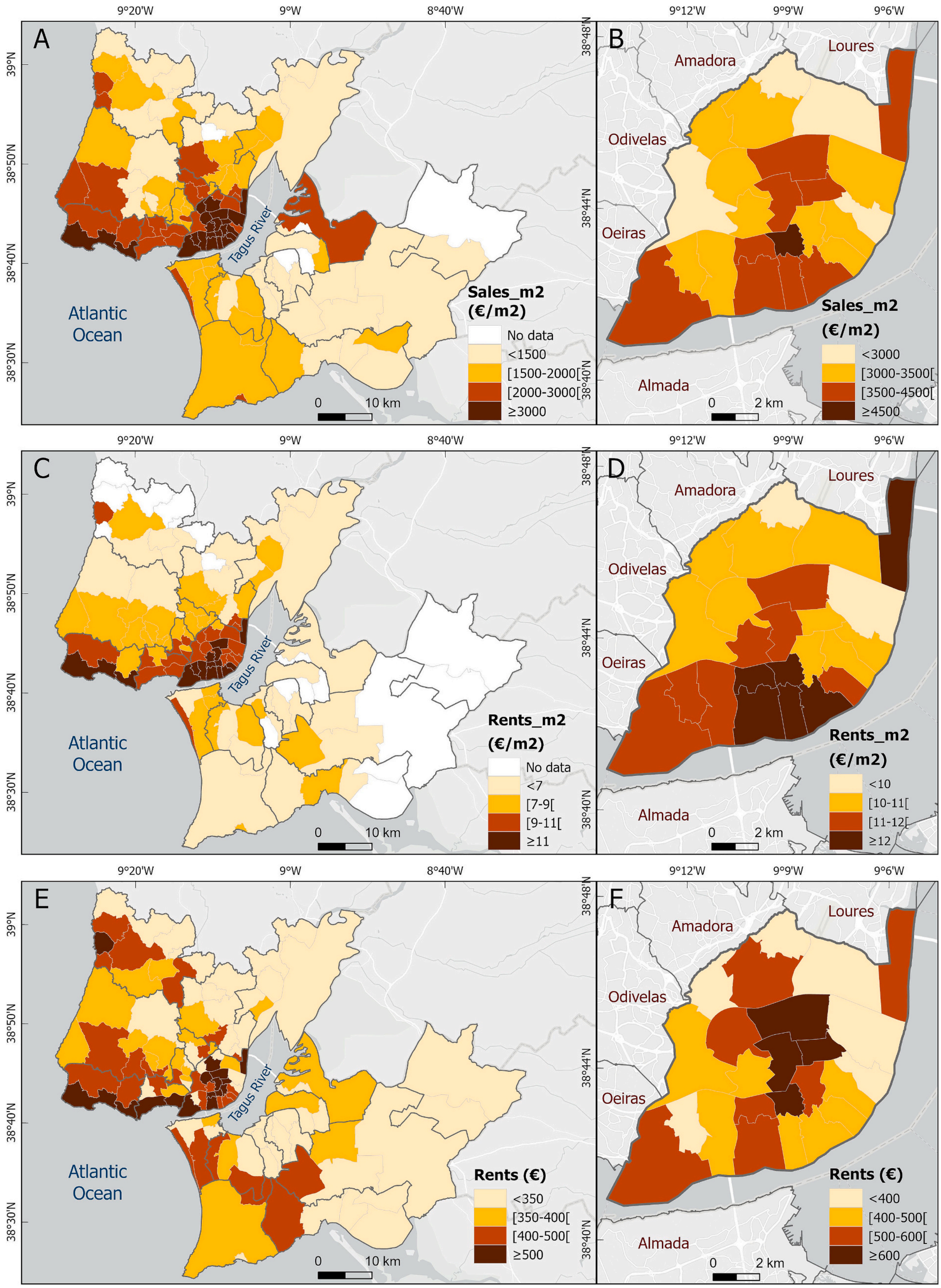


**Fig. 4.** *Sales_m2* (A and B), *Rents_m2* (C and D) and *Rents* (E and F) in the LMA parishes (A, C and E) and the Lisbon parishes (B, D and F) in 2021.

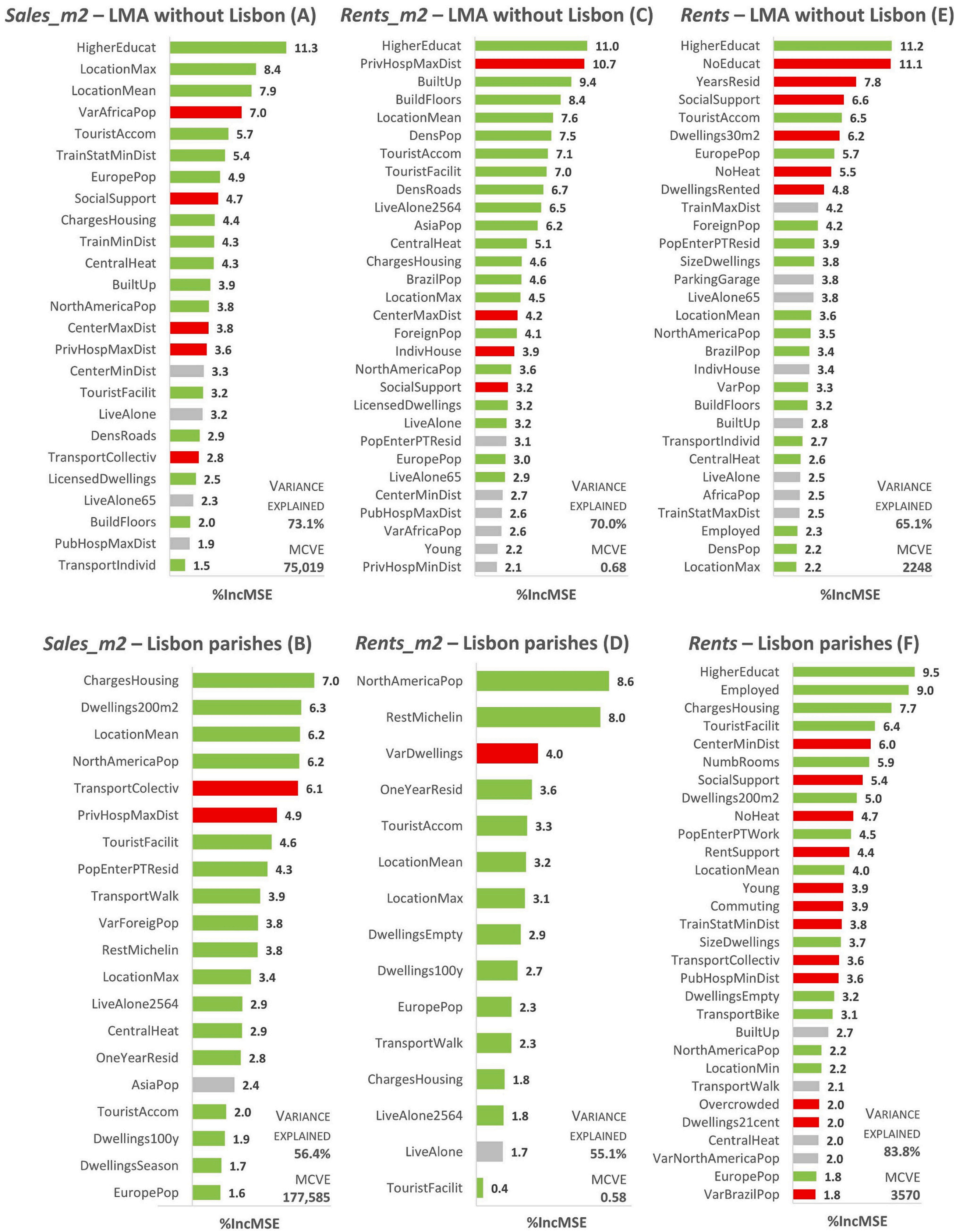


**Fig. 5.** Variable importance for the random forest models for *Sales_m2* (A and B), *Rents_m2* (C and D) and *Rents* (E and F) in the LMA without Lisbon (top) and Lisbon (bottom). Values describe the contribution of each independent variable to improving the mean squared error (MSE) of the model. Green, red, and grey bars identify positive, negative, and unclear marginal effects of the variable on *Sales_m2*/*Rents_m2*/*Rents*, respectively. MCVE corresponds to Mean Cross-Validation Error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

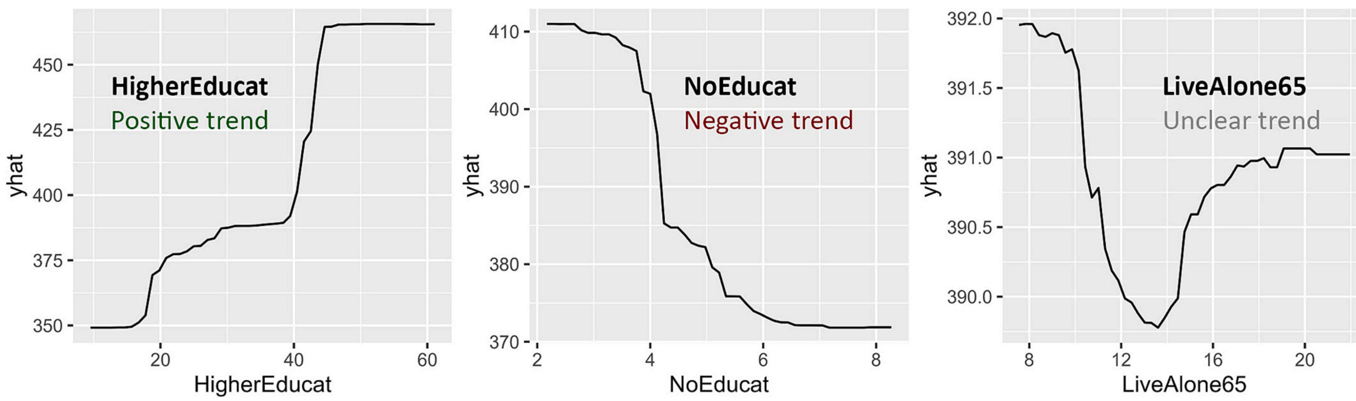


**Fig. 6.** Examples of partial dependence plots obtained from the *Rents* model of Lisbon.

*HigherEducat*, *LocationMax* and *LocationMean*, all with positive effects. The second model (*Sales_m2* Lisbon – Fig. 5B) determined 20 relevant independent variables and was able to explain 56.4 % of the variance (MCVE = 177,585). The three most important variables are *ChargesHousing*, *Dwellings200m2* and *LocationMean*, all with positive effects.

Regardless of the study areas, the *Sales_m2* values tend to be higher in parishes with: 1) greater transport accessibility and proximity to social facilities; 2) less use of public/collective transport; 3) higher mortgage payments; 4) more European and North-American population; 5) more tourism; 6) less energy poverty; 7) more population living alone.

However, some differences arise between areas. Outside Lisbon, there are additional relevant factors associated with high *Sales_m2* that do not emerge within Lisbon: 1) proximity to city centre; 2) higher education levels; 3) loss of African population; 4) less social support; 5) more construction; and 6) less distance to train.

The important factors associated with high *Sales_m2* unique to Lisbon appear in parishes with: 1) larger and older dwellings, and with seasonal occupation; 2) proximity to high-quality services (e.g., *RestMichelin*); 3) a preference for walking as the main means of transport; 4) more of the working-age population living alone; and 5) more foreign residents, namely from Asia.

The third model (*Rents_m2* LMA without Lisbon – Fig. 5C) determined 30 relevant independent variables and explains 70 % of the variance (MCVE = 0.68). The most important variables are *HigherEducation*, *PrivHospMaxDist* and *BuiltUp*. The fourth model (*Rents_m2* Lisbon – Fig. 5D) determined 15 relevant independent variables and was able to explain 55.1 % of the variance (MCVE = 0.58). The most important factors are *NorthAmericaPop*, *RestMichelin* and *VarDwellings*.

The higher values of *Rents_m2* for both outside and inside Lisbon occur in urban central areas, namely in parishes with: 1) more tourism; 2) higher mortgage payments; 3) more European and North-American population; 4) greater transport accessibility and proximity to social facilities; and 5) more people living alone. These associated factors also appear in the *Sales_m2* models for both study areas.

The results of *Rents_m2* model for the LMA without Lisbon show the following exclusive factors related to high rental prices: 1) higher population density and built-up areas, including less population living in a house; 2) more educated population; 3) less social support; 4) more foreign population, namely Asian and Brazilian; 5) less energy poverty; 6) less distance to Lisbon and to private hospitals; 7) and better road accessibilities.

On the other hand, Lisbon presents some exclusive variables linked to high values of *Rents_m2*: 1) less residence time; 2) a preference for walking as the main means of transport; 3) less built-up areas; 4) more older and empty dwellings; and 5) existence of high-standard restaurants.

The fifth model (*Rents* LMA without Lisbon – Fig. 5E) determined 31 relevant independent variables and was able to explain 65.1 % of the variance (MCVE = 2248). The three most important variables are *HigherEducat*, *NoEducat*, and *YearsResid*. The last model (*Rents* Lisbon – Fig. 5F) determined 31 relevant independent variables and was able to explain 83.8 % of the variance (MCVE = 3570). The most important variables are *HigherEducat*, *Employed* and *ChargesHousing*, all with positive effects.

Regardless of the area, *Rents* tend to be higher in parishes with: 1) more population with higher education; 2) more tourism; 3) larger dwellings; 4) less social support; 5) more European and North-American population; 6) lower levels of energy poverty; 7) proximity to social facilities; and 8) higher levels of employment.

Outside Lisbon, these are the important factors associated with high *Rents* that are not significant in Lisbon: 1) lower levels of population without education; 2) shorter average household residence; 3) fewer rented dwellings; 4) more foreign population and more foreign population for the purpose of residence; 5) more Brazilian population; 6) higher population growth and population density; and 7) higher number of floors per building.

The factors associated with high *Rents* that are significant only in Lisbon parishes are: 1) higher charges due to the acquisition of housing; 2) proximity to city centre; 3) newer and/or empty dwellings; 4) more foreign population with the purpose of working; 5) less rent support; 6) less young population; 7) more soft mobility (cycling and walking) and less duration of commuting; 8) more overcrowded dwellings; and 9) negative variations of Brazilian population.

A more detailed comparison between the *Sales_m2*, *Rents_m2* and *Rents* models of LMA and Lisbon (Table 5) allows us to draw several conclusions. First, there are only three of the 81 variables (4 %) that come up in all models (*EuropePop*, *LocationMean* and *NorthAmericaPop*). However, *HigherEducat* is the most important variable in four models, being non-relevant for *Sales_m2* and *Rents_m2* in Lisbon. Second, there are 17 variables (21 %) that are represented in both *Sales_m2*, *Rents_m2* and *Rents* models in one of the study areas. Third, there are three variables (*VarForeignPop*, *DwellingsSeason* and *TrainMinDist*) one (*VarDwellings*) that are exclusive to *Sales_m2* and *Rents_m2* models, respectively. On the other hand, there 23 variables that only appear in the *Rents* models, which demonstrates the different nature of this independent variable when compared to the other two. Fourth, there are 12 variables (15 %) that appear in the three models for LMA and there are six variables (7 %) that appear in the three models for Lisbon. Lastly, there are 15 variables (19 %) considered as irrelevant or not included in the top 30 in all models: *SizeHouseholds*, *Elderly*, *ActivePop*, *WorkParish*, *WorkOtherParish*, *WorkOtherMunic*, *PopEnterPTEduc*, *VarEuropePop*, *VarAsiaPop*, *BuildRepair*, *NonCentralHeat*, *RestInsider*, *PubHosp*, *PrivHosp* and *PubPrivHosp*.

**Table 5**
Position of each independent variable in random forest models for *Sales_m2*, *Rents_m2*, and *Rents* in the LMA without Lisbon and in Lisbon.

| Variable | *Sales_m2* | | *Rents_m2* | | *Rents* | | Variable | *Sales_m2* | | *Rents_m2* | | *Rents* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LMA | Lisbon | LMA | Lisbon | LMA | Lisbon | | LMA | Lisbon | LMA | Lisbon | LMA | Lisbon |
| BuiltUp | 12 | | 3 | | 22 | 21 | BuildFloors | 23 | | 4 | | 21 | |
| DensPop | | | 6 | | 29 | | VarDwellings | | | | 3 | | |
| VarPop | | | | | 20 | | LicensedDwellings | 21 | | 21 | | | |
| Overcrowded | | | | | | 25 | SizeDwellings | | | | | 13 | 16 |
| IndivHouse | | | 18 | | 19 | | Dwellings100y | | 18 | | 9 | | |
| Young | | | 29 | | | 13 | Dwellings21cent | | | | | | 26 |
| LiveAlone | 18 | | 22 | 14 | 25 | | NumbRooms | | | | | | 6 |
| LiveAlone2564 | | 13 | 10 | 13 | | | Dwellings200m2 | | 2 | | | | 8 |
| LiveAlone65 | 22 | | 25 | | 15 | | Dwellings30m2 | | | | | 6 | |
| NoEducat | | | | | 2 | | DwellingsSeason | | 19 | | | | |
| HigherEducat | 1 | | 1 | | 1 | 1 | DwellingsEmpty | | | | 8 | | 19 |
| Employed | | | | | 28 | 2 | DwellingsRented | | | | | 9 | |
| SocialSupport | 8 | | 20 | | 4 | 7 | RentSupport | | | | | | 11 |
| YearsResid | | | | | 3 | | ParkingGarage | | | | | 14 | |
| OneYearResid | | 15 | | 4 | | | CentralHeat | 11 | 14 | 12 | | 24 | 27 |
| Commuting | | | | | | 14 | NoHeat | | | | | 8 | 9 |
| TransportWalk | | 9 | | 11 | | 24 | TouristAccom | 5 | 17 | 7 | 5 | 5 | |
| TransportBike | | | | | | 20 | TouristFacilit | 17 | 7 | 8 | 15 | | 4 |
| TransportIndivid | 25 | | | | | | LocationMax | 2 | 12 | 15 | 7 | 30 | |
| TransportCollectiv | 20 | 5 | | | 23 | 17 | LocationMin | | | | | | 23 |
| ChargesHousing | 9 | 1 | 13 | 12 | | 3 | LocationMean | 3 | 3 | 5 | 6 | 16 | 12 |
| ForeignPop | | | 17 | | 11 | | CenterMinDist | 16 | | 26 | | | 5 |
| VarForeigPop | | 10 | | | | | CenterMaxDist | 14 | | 16 | | | |
| PopEnterPTWork | | | | | | 10 | RestMichelin | | 11 | | 2 | | |
| PopEnterPTResid | | 8 | 23 | | 12 | | TrainMinDist | 10 | | | | | |
| EuropePop | 7 | 20 | 24 | 10 | 7 | 29 | TrainMaxDist | | | | | 10 | |
| AfricaPop | | | | | 26 | | TrainStatMinDist | 6 | | | | | 15 |
| NorthAmericaPop | 13 | 4 | 19 | 1 | 17 | 22 | TrainStatMaxDist | | | | | 27 | |
| BrazilPop | | | 14 | | 18 | | DensRoads | 19 | | 9 | | | |
| AsiaPop | | 16 | 11 | | | | PubHospMinDist | | | | | | 18 |
| VarAfricaPop | 4 | | 28 | | | | PubHospMaxDist | 24 | | 27 | | | |
| VarNorthAmericaPop | | | | | | 28 | PrivHospMinDist | | | 30 | | | |
| VarBrazilPop | | | | | | 30 | PrivHospMaxDist | 15 | 6 | 2 | | | |

Notes: The numbers correspond to the respective position in each chart of Fig. 5. The colours are associated with the numbers/positions, ranging from dark green (most important variables) to dark red (least important). Grey is applied when a given variable is not relevant or does not appear in Fig. 5. The non-relevant variables, or those that do not appear in the top 30 of the six models, were not presented in this table.

## 6. Discussion

### 6.1. Usefulness, suitability, advantages, and limitations of random forest modelling

At this point it is helpful to discuss the usefulness of the random forest algorithm for regression purposes and its suitability for this study.

Several studies have emphasized that random forest regression modelling can provide better results than, for example, linear regression because it adjusts for non-linear behaviours, handles large amounts of data, does not require scaling of independent variables, and deals with outliers and noise well (Biau & Scornet, 2016; Blanchet et al., 2020; Donick & Lera, 2021; Sarker, 2021). Considering these advantages and the goals of this study, random forest modelling is the most suitable method, although some weaknesses need to be considered.

First, the number of cases in both study areas (76 and 24) is low when compared to databases with thousands of entries, and second, there are few independent variables with non-linear relationships with the dependent variables. These facts affect the performance of the random forest models and mean that we are not able to fully leverage some of their main virtues (Gregorutti et al., 2017; Han et al., 2021; Oller et al., 2021). Six linear regression models for *Sales_m2*, *Rents_m2* and *Rents* were developed to compare the results with those obtained using random forest models.[4] If the goal of this study was to predict housing prices, linear regression would be a better solution, ensuring a better fit with the values of the dependent variables. However, we are mainly interested in identifying all factors relevant to housing prices in different submarkets and study areas, and random forest models allow us to efficiently measure their importance (Saarela & Jauhiainen, 2021). In this sense, our approach differs from most studies that use random forest or linear regression, in which some associated factors must be discarded due to collinearity so as not to harm the predictive capacity of the model. In this study, we aim to identify all relevant factors (the irrelevant ones were discarded in the final models), even if they are highly correlated, because these results will help researchers and decision-makers to improve their knowledge of the housing submarkets. In the LMA, *HigherEducation*, *ChargesHousing* and *LocationMean* have correlation coefficients above 0.75; if we removed highly correlated variables to prevent multicollinearity, we would retain just one of these variables, thereby losing two variables that express completely different realities. The results demonstrate that these three variables are relevant for several models (Fig. 5 and Table 5), but, at the same time, their relevance is not of a similar degree, i.e., for LMA outside Lisbon, *HigherEducation* is the most important variable in the *Rents* model whereas *ChargesHousing* is considered irrelevant. These connections would remain undetected lost if highly correlated variables were not included in the models. In short, variable selection improves the models' performance, but, considering the goals of this research, only irrelevant variables should be removed and not highly correlated ones.

[4] Linear regression obtained better results in terms of predictive capacity for all six models. For example, the predictive capacity of the Rents_m2 model for the LMA without Lisbon is 70.0 % using random forest and 85.7 % using linear regression. The variable selection process in linear regression was done by removing the highly correlated independent variables ($r \geq 0.75$).

### 6.2. Spatial differences in housing submarkets and the role of associated factors

#### 6.2.1. Dependent variables: Characteristics and spatial behaviour

The three dependent variables aim to express the different realities of the housing market in the two study areas, allowing comparisons to be made between sales and rents, and, more specifically, between new rents and all rents. However, these variables have different characteristics. *Sales_m2 and Rents_m2* correspond to the values of houses sold and rented, respectively, in a single year, while *Rents* reflects all existing contracts up to 2021. In addition, *Rents* is not weighted by the size of the dwellings, and this is something that can contribute to differences between parishes, i.e., dwellings with different dimensions may have similar rental prices in different parishes. Other data limitations and uncertainties also need to be considered. Taking the parish as a unit of analysis, in many cases, does not represent the housing reality with the necessary detail. Not all dependent and independent variables are homogeneous within a parish, with variations occurring on both a broader and a finer scale. It is also important to consider both the variations across geographical regions and the context of each dependent variable. The Lisbon municipality is exposed to stronger urbanization and population pressures than the surrounding areas. Moreover, dependent variables also reflect economic, social, and political contexts. Important factors such as public policies, real estate dynamics, macroeconomic realities, and housing regulation decisions also influence the housing submarkets and their prices over time, but these operate at a level broader than that of the parish.

The difference in the models' predictive capacity allows us to assume that the independent variables play a more important role in explaining the spatial distribution of new transactions/contracts per $m^2$ (*Sales_m2* and *Rents_m2*) in the wider LMA than in Lisbon. The opposite is true of *Rents*. A wide range of independent variables was only able to explain little more than half of the variance of *Sales_m2* and *Rents_m2* within Lisbon, something that can be explained in two ways. First, there are other relevant factors that were not considered in this study (e.g., income), data on which are not available at parish level. Second, some of the variation in *Sales_m2* and *Rents_m2* values may be due to factors that cannot easily be captured as quantitative variables (e.g., urban image/perception, feelings or quality of life, urban management, and land use planning). In other words, there may not be the necessary spatial detail at parish level to improve the models' predictive capacities. However, examining the parish level provides valuable information on a wide range of associated factors and their relationship with the dependent variables.

*Sales_m2* and *Rents_m2* are strongly and positively correlated, and have similar spatial distributions. *Rents* has a different spatial distribution, reflecting the role played by rental contracts prior to 1990 and social housing. *Sales_m2* and *Rents_m2* have a more concentrated spatial pattern within Lisbon that contrasts with that of *Rents*, which is characterized by a more scattered and interspersed pattern that is shaped by several factors, such as rent regulation. The co-existence of social and private rental markets in some areas (neighbourhoods, parishes), which are associated with more affordable rental values and allocation by local/central authorities, explains why *Rents* is more compartmentalized and less continuous than *Sales_m2/Rents_m2*. The bipolarity of the rental dependent variables (*Rents_m2* vs. *Rents*) in the LMA, and particularly within Lisbon, is reflected in some geographical pockets that have houses at lower prices. Rents are lower in parishes with greater social and housing precariousness. Additionally, the greater fluctuation of the rental market may be explained by appeal to path-dependency theories, which claim that earlier policy decisions set the housing stock on distinct trajectories that are later difficult to reverse (Bengtsson & Ruonavaara, 2010).

#### 6.2.2. Independent variables: Comparisons with other contexts and contrasts between submarkets and study areas

The results showed that housing prices were associated with location, dwelling/building features, urban and population dynamics, mobility, migration, and tourism. This diversity of data is absent from housing studies that use machine learning techniques. The association of housing prices with dwelling/building features is common in this type of research. Building age and dwelling area are top factors in several studies (Čeh et al., 2018; Hong et al., 2020; Hu, Huang, & Li, 2022; Lorenz et al., 2023; Rico-Juan & de La Paz, 2021). Although not equally relevant, other factors such as the number of apartments in a building, the type of house, and the dwelling/building features (quality of construction, existence/number of elevators, renovations, number of bathrooms and orientation) are also mentioned as important. Due to the different nature of the cases (dwellings vs. parishes) it is not always possible to establish direct comparisons between the independent variables. However, some parallel can be drawn with our results. The age of buildings (expressed in our study by *Dwellings100y* and *Dwellings21cent*) does not seem to be as important as it is in other contexts, but we would need a more informative, continuous variable to draw definitive conclusions. Mayer et al. (2019) and Lorenz et al. (2023) found a U-shaped pattern in which the newest and oldest buildings are associated with the highest prices, a pattern that would also seem plausible for a historical city like Lisbon. On the other hand, the valuation of both old and new buildings by the real estate market can vary depending on their conservation status and the quality of their construction. The importance of the size of the dwelling (expressed here by *SizeDwellings*, *NumbRooms*, *Dwellings200m2*, and *Dwellings30m2*) in our study is more in line with the results of other studies. This is a linear factor with a positive relationship with prices (Dumm et al., 2016, 2018; Lorenz et al., 2023; Stamou et al., 2017).

The importance of location is also frequently highlighted in the literature, which includes variables such as distance/proximity to city centre, infrastructures, facilities, services or transports, neighbourhood, and views. The variables related to proximity to or existence of facilities and services (e.g., *LocationMean*, *RestMichelin* or *PrivHospMaxDist*) were correlated with housing prices, demonstrating that higher prices are associated with areas better served by these facilities. On the other hand, distances to city centre, to transport, and to infrastructure had different relationships with housing prices. First, distance to city centre was mostly negatively associated with prices, with sharp price drops in the first 10/15 km. Similar results were found by Osland (2010) and Zheng et al. (2016). Nevertheless, nonlinear trends were also identified in some cases, meaning that, outside Lisbon municipality's boundaries, prices fluctuate as the distance to the centre increases. Second, distance to railway lines or to train stations showed different trends. Prices increase significantly across the first 5 km from a railway line and then stabilize. Distance to train stations has a similar pattern outside Lisbon, but exhibits the opposite pattern within Lisbon, where there is a negative relationship with prices, i.e., prices decrease sharply up to 1 km away from train stations and then decrease more smoothly. This means that, in Lisbon, proximity to train stations is seen as an advantage, potentially outweighing the negative effect of noise and urban barriers. Third, density of main roads was positively related to housing prices.

Less attention is given in the literature to social issues and population/urban dynamics, although some studies recognize the association between areas with high purchasing power and more expensive housing (Lorenz et al., 2023; Rico-Juan & de La Paz, 2021). This is in line with the positive associations established in this study between prices and variables such as level of higher education and the size of the European and North American population, which may be understood as proxies for income. Location, land availability and urban consolidation/infrastructure are highly relevant for *Sales_m2* and *Rents_m2*. Thus, both sales and rental submarkets share common spatial patterns (Fig. 4) and many important associated factors (Fig. 5), which shows that these two submarkets are subject to similar dynamics when only the new transactions/contracts are considered. In this case, the highest and lowest prices currently tend to occur in the same parishes and are associated with (and probably determined by) the same variables/factors. Nevertheless, some differences were found between sales and new rents. Compared to the *Sales_m2* model, *Rents_m2* seems to have a stronger link to variables related to urban and population diversity and density. The high values of new rental contracts are associated with the characteristics of cosmopolitan urban areas, where population transformation and the supply of services, commerce and mobility are more intense. In contrast, *Sales_m2* seems to be more linked to the physical quality of housing and the importance of the banking system in home buying.

By contrast, socioeconomic features of the populations/parishes are more closely associated with *Rents* than with *Sales_m2* and *Rents_m2*. The *Rents* variable adds a temporal perspective to the rental submarket by considering all the existing contracts, showing a substantially different spatial distribution when compared to the new rents and highlighting the role played by other associated factors, namely those related to the socioeconomic characteristics of the population/parishes. This means that *Rents_m2* and *Rents* express two distinct ways to address the rental submarket. In this sense, current and past rental prices are not always spatially coincident in the LMA. Their comparison points to the need to deepen the discussion on issues of urban planning options, mobility policies, urban consolidation and qualification, regulation of the rental market, among others.

The two study areas have distinct associated factors related to the dependent variables. There are several relevant variables in Lisbon that are less relevant or even irrelevant outside Lisbon and vice-versa. The variables related to the age of the building and the dwelling area are mainly relevant in Lisbon. These are not the most important factors in this context, but, as in other studies, these are more relevant in the main city (i.e., Lisbon municipality) compared to the surrounding areas. Most variables related to urban density and construction were identified as more important outside Lisbon (e.g., *BuiltUp*, *DensPop*, *BuildFloors*, and *LicensedDwellings*). This means that housing prices outside Lisbon are higher in the most urban parishes, which are in most cases along the railway/road axes of urban expansion in the Portuguese capital. Meanwhile, the higher proportion of new residents and the lower rate of change in the number of dwellings are associated with higher housing prices in Lisbon. This reveals both greater demand and lower housing supply within the city compared to other municipalities in the LMA, which makes housing more unaffordable by pushing up rents and mortgage payments (as expressed by *ChargesHousing*). Gentrification and touristification also substantially reduce the availability of dwellings on the market (Cocola-Gant & Gago, 2019). The proportion of rented dwellings is highly relevant the LMA outside Lisbon, while the proportions of empty dwellings and those with seasonal occupation are mostly relevant within Lisbon. A greater proportion of dwellings without a full-time residential function reflects real estate speculation and harms the supply of housing, increasing prices and compromising access to housing for younger people, who tend to have more difficulty paying high rents and accessing home loans.

The use of transport is clearly different inside and outside Lisbon. The level of individual transport, density of main roads and distances to city centre, train stations, railways and public hospitals are particularly relevant in municipalities of the LMA outside Lisbon, while soft mobility (walking and cycling) is only relevant in Lisbon. In a city like Lisbon, where circulation and access to services and facilities is "taken for granted", proximity to infrastructure is a less differentiating factor in the valuation of housing. The population that settles in more central areas seeks and adopts different lifestyles, valuing proximity to daily amenities and services – particularly high-quality ones, as expressed by *RestMichelin*. Greater centrality also allows for smooth mobility and the avoidance of public transport for daily trips.

#### 6.2.3. Featured independent variables

Given the large number of factors considered, it is pertinent to discuss some of them in more detail, highlighting four in particular: 1) education; 2) population changes, particularly in terms of nationality; 3) support for vulnerable populations; and 4) tourism.

Rate of higher education is the most important independent variable in four models. Let us then hypothesize that higher education facilitates access to housing in the most expensive parishes. There is a clear positive relationship between rate of higher education and housing prices. However, if this is true for the LMA outside Lisbon, the same cannot be said for Lisbon's parishes, where *HigherEducation* is not a relevant variable for new transactions/rental contracts. For example, Lumiar is the parish in Lisbon with the second highest rate of higher education (60.6 %), but it only occupies the 19th position in *Sales_m2* and the 22nd position in *Rents_m2*. If higher education may eventually give access to better salaries and greater purchasing power, this may not currently be enough to buy/rent a house in the most expensive parishes in Lisbon. When it comes to the rental submarket, the duality in Lisbon is notable because higher education is considered as both irrelevant for new rents and the most important variable for all rents. Another perspective may point to a compromise/balance between price, location, size, dwelling/building age, among others, depending on how each individual/family value each factor. Thus, the choice of which parish to live in may be based on the preference or needs of the buyer/tenant.

The importance of the level of foreign population confirms the increasing internationalization of the Portuguese real estate market (Rodrigues, 2022). The increase in the North American and European population stands out here, and the variables related to these groups were considered as relevant in all models. The demand among retirees and investors to live in Portugal has been increasing significantly in recent years. Their high purchasing power allows them to buy or rent dwellings in the most expensive parishes of the LMA. According to Statistics Portugal, in the fourth quarter of 2021, transactions made in the LMA by buyers domiciled abroad were 131 % higher than those made by Portuguese citizens: 4283 €/m$^2$ vs. 1858 €/m$^2$. However, if some groups of foreigners have the economic capacity to live in the most valued and sought-after residential areas, others are displaced (e.g., the African population and/or their descendants). Even in the European population, purchasing power can change depending on nationality, which is reflected in different spatial distributions across the LMA. Nationals of the 15 pre-2004 European Union Member States,[5] the most developed countries in Europe, tend to seek parishes with high urban quality than those sought by people from Eastern European countries.

The existing data allow us to debate the worsening of both socio-spatial segregation and lack of access to the city for the most vulnerable populations. The highest values of *Rents_m2* and *Sales_m2* outside Lisbon are associated with parishes where the proportion of social support is lower. The proportion of the population that benefits from rent support is only relevant within the Lisbon parishes for *Rents*. These results point to a segmentation of the rental and sales markets and to the difficulty of resolving socio-spatial inequalities, especially in the residential sector.

When it comes to tourism, there is a spatial contrast between variables (*TouristAccom* and *TouristFacilit*). The density of tourist facilities is more significant in Lisbon for both *Sales_m2* and *Rents*, expressing the fact that tourist demand in the city makes these parishes more expensive. On the other hand, the association between high housing values and the density of short-term tourist accommodation is particularly high outside Lisbon. There is more short-term tourist accommodation within Lisbon, but its relationship with the spatial distribution of housing prices is not as evident as it is outside of Lisbon, where higher densities of such accommodation seem to be associated with more attractive areas in terms of landscape and/or heritage values. These findings are in line with research that demonstrates the impact of tourism on housing affordability in Portugal and European cities (Franco & Santos, 2021; Reichle et al., 2023).

### 6.3. Policy implications

This study has highlighted the diversity of the housing market and associated factors, addressing differences in the patterning of rental and sales markets across different geographical areas, as other studies have

[5] Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom.

noted (Hu, He, & Su, 2022; Nelson et al., 2024). The methodology used has the potential to be integrated into the monitoring of and the political decision-making about the housing market. Given that housing regulation results from the social and political relationships that develop within a society (Alves, 2022; Lord et al., 2018), our results contribute to the development of well-informed and joined-up policies. Several examples of how findings can be used to inform policies can be found below.

The internationalization of housing, namely through the demand of foreign groups with purchasing power, is associated with higher housing prices. This may justify a more careful assessment of policies that benefit, fiscally or otherwise, groups of foreigners buying homes in Portugal.

Given the particular importance of the rental market in central urban areas and for smaller households, there may be a justification for policies aimed at different audiences and areas. For example, specific policies could be aimed at areas with high tourist demand, while others might be aimed at stimulating the market in areas with a lower percentage of rented accommodation.

The differentiation between factors associated with new rental contracts (*Rents_m2*) and factors associated with all rents (*Rents*) exposes how restricted access to the current rental market in the LMA has become: high rental values increasingly exclude the lower and middle classes from this market. This could be the starting point for developing strategies to stimulate or regulate certain sectors of the rental market.

Housing prices should not be analysed only at a local scale, as they are affected by strong socio-spatial dynamics at a wider scale that relate to mobility, labour organization, and broader population dynamics. However, in response to this situation, policy design and implementation are providing fragmented responses. It should be noted that the legal and organisational framework for housing planning or strategies is increasingly designed at municipal level. Therefore, this study supports the development of a coordinated approach between municipalities at the metropolitan level.

## 7. Conclusions

Understanding the drivers of housing prices remains a key challenge. Many factors influence prices and shape their spatial patterns. The high demand for, and low supply of, housing in the LMA generates an environment where there is strong real estate competition. This reinforces segmentation. This pattern is underpinned by the fact that housing is both a need and an investment asset.

This study has addressed both sales and rental submarkets in two study areas in the LMA. In addition, the analysis of the two dependent variables related to rental prices allows us to explore the temporal dimension of the housing market. Using the parish as the unit of analysis instead of dwellings/buildings allowed us to add a vast set of variables from several domains to the data analysis. The random forest made it possible to identify the associated factors relevant to each submarket and study area and quantify their individual importance, uncovering, confirming, or rejecting relationships between dependent and independent variables. This is a geographically-based approach that has allowed for a more direct connection with the characteristics of study areas and their populations – a direct connection that distinguishes this study from others that use machine learning algorithms to predict prices. This offers a new perspective on housing studies, providing a tool for better understanding housing dynamics and supporting political decision-making processes.

The results demonstrate strong associations between the three dependent variables and factors associated with urban quality and the social status of the population. The current prices of sales and rental submarkets show similar spatial patterns, i.e., the most expensive or cheapest parishes for buying or renting a house are generally the same. Sales and new rents also tend to share common associated factors, mostly related to location variables. On the other hand, when all existing rental contracts are considered instead of just new rents, the results show a different spatial distribution and a stronger link with the socio-economic characteristics of the parishes. So, current housing prices are more associated with the value of the place and the advantages of a given location, while the prices of all rents are more related to the population features (e.g., education or employment), expressing the evolution of urban growth and the appropriation of space over the decades. These aspects contribute to and are highlighted in the results by study area, exposing local particularities.

Although analysing the LMA has made it easier to place our results in a local and national context, extending the scope of the analysis to other cities in Portugal or other countries would enable greater clarity on the differences between the housing submarkets. Future research could add a temporal component to these results; this would allow it to capture the evolution of housing prices in relation to different associated factors.

## CRediT authorship contribution statement

**Miguel Leal:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marina Carreiras:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis. **Sónia Alves:** Writing – review & editing, Writing – original draft, Validation, Investigation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Marina Carreiras was supported by the Portuguese Foundation for Science and Technology (FCT) through grant PD/BD/114270/2016. Sónia Alves has received funding from the FCT under the DL 57/2016/CP1441/CT0017 and from the Marie Skłodowska-Curie Actions (MSCA) research and innovation programme of the European Union's Horizon Europe under the grant agreement ID 101086488: Delivering sAfe and Social housing (DASH). This research was also supported by the Interdisciplinary Centre of Social Sciences - CICS.NOVA (UIDB/04647/2020 and UIDP/04647/2020), the Centre of Geographical Studies (UIDB/00295/2020 and UIDP/00295/2020), the Institute of Social Sciences (LA/P/0051/2020) and the Associate Laboratory TERRA (LA/P/0092/2020).

## Data availability

The data used in this research are publicly available on the websites of the institutions mentioned in the data and methods section.

## References

Aalbers, M. B. (2017). The variegated Financialization of housing. *Internacional Journal of Urban and Regional Research, 41*(4), 542–554. https://doi.org/10.1111/1468-2427.12522

Adair, A. S., Berry, J. N., & McGreal, W. S. (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research, 13*(1), 67–83. https://doi.org/10.1080/095999196368899

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2021). House Price prediction using random Forest machine learning technique. *Procedia Computer Science, 199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Alves, S. (2017). Poles apart? A comparative study of housing policies and outcomes in Portugal and Denmark. *Housing, Theory and Society, 34*(2), 221–248. https://doi.org/10.1080/14036096.2016.1236036

Alves, S. (2022). Divergence in planning for affordable housing: A comparative analysis of England and Portugal. *Progress in Planning, 156*, Article 100536. https://doi.org/10.1016/j.progress.2020.100536

Alves, S., Azevedo, A. B., Mendes, L., & Silva, K. (2023). Urban regeneration, rent regulation and the private rental sector in Portugal: A case study on Inner-City Lisbon's social sustainability. *Land, 12*(8), 1644. https://doi.org/10.3390/land12081644

Ball, M., Shepherd, E., & Wyatt, P. (2022). The relationship between residential development land prices and house prices. *Town Planning Review, 93*(4), 401–421. https://doi.org/10.3828/tpr.2021.27

Bengtsson, B., & Ruonavaara, H. (2010). Introduction to the special issue: Path dependence in housing. *Housing, Theory and Society, 27*(3), 193–203. https://doi.org/10.1080/14036090903326411

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Blanchet, L., Vitale, R., van Vorstenbosch, R., Stavropoulos, G., Pender, J., Jonkers, D., … Smolinska, A. (2020). Constructing bi-plots for random forest: Tutorial. *Analytica Chimica Acta, 1131*(2020), 146–155. https://doi.org/10.1016/j.aca.2020.06.043

Borde, S., Rane, A., Shende, G., & Shetty, S. (2017). Real estate investment advising using machine learning. *International Research Journal of Engineering and Technology (IRJET), 4*(3), 1821–1825.

Branco, R., & Alves, S. (2020). Urban rehabilitation, governance, and housing affordability: Lessons from Portugal. *Urban Research & Practice, 13*(2), 157–179. https://doi.org/10.1080/17535069.2018.1510540

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Canadas, M. J., Leal, M., Soares, F., Novais, A., Ribeiro, P. F., Schmidt, L., … Santos, J. L. (2023). Wildfire mitigation and adaptation: Two locally independent actions supported by different policy domains. *Land Use Policy, 124*, Article 106444. https://doi.org/10.1016/j.landusepol.2022.106444

Casali, Y., Aydin, N. Y., & Comes, T. (2022). Machine learning for spatial analyses in urban areas: A scoping review. *Sustainable Cities and Society, 85*, Article 104050. https://doi.org/10.1016/j.scs.2022.104050

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random Forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information, 7*(5), 168. https://doi.org/10.3390/ijgi7050168

Choy, L. H. T., & Ho, W. K. O. (2023). The use of machine learning in real estate research. *Land, 12*(4). https://doi.org/10.3390/land12040740

Cocola-Gant, A. (2023). Place-based displacement: Touristification and neighborhood change. *Geoforum, 138*, Article 103665. https://doi.org/10.1016/j.geoforum.2022.103665

Cocola-Gant, A., & Gago, A. (2019). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. *Environment and Planning A: Economy and Space, 53*(7), 1671–1688. https://doi.org/10.1177/0308518X19869012

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics, 20*(2), 492–503. https://doi.org/10.1093/bib/bbx124

Dimopoulos, T., Tyralis, H., Bakas, N. P., & Hadjimitsis, D. (2018). Accuracy measurement of random forests and linear regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. *Advances in Geosciences, 45*, 377–382. https://doi.org/10.5194/adgeo-45-377-2018

Donick, D., & Lera, S. C. (2021). Uncovering feature interdependencies in high-noise environments with stepwise lookahead decision forests. *Scientific Reports, 11*(1), 1–12. https://doi.org/10.1038/s41598-021-88571-3

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2016). Price variation in waterfront properties over the economic cycle. *Journal of Real Estate Research, 38*(1), 1–26. https://doi.org/10.1080/10835547.2016.12091435

Dumm, R. E., Sirmans, G. S., & Smersh, G. T. (2018). Sinkholes and residential property prices: Presence, proximity, and density. *Journal of Real Estate Research, 40*(1), 41–68. https://doi.org/10.1080/10835547.2018.12091491

Eurofound. (2023). *Unaffordable and inadequate housing in Europe*. Publications Office of the European Union.

Farha, L. (2017). Report of the Special Rapporteur on adequate housing as a component of the right to an adequate standard of living, and on the right to non-discrimination in this context, on her mission to Portugal: Mission to Portugal. https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/34/51/Add.2.

Fingleton, B. (2008). Housing supply, housing demand, and affordability. *Urban Studies, 45*(8), 1545–1563. https://doi.org/10.1177/0042098008091490

Franco, S. F., & Santos, C. D. (2021). The impact of Airbnb on residential property values and rents: Evidence from Portugal. *Regional Science and Urban Economics, 88*, Article 103667. https://doi.org/10.1016/j.regsciurbeco.2021.103667

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Garha, N. S., & Azevedo, A. B. (2021). Population and housing (Mis)match in Lisbon, 1981–2018. A challenge for an aging society. *Social Sciences, 10*(3), 102. https://doi.org/10.3390/socsci10030102

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Grekousis, G., Feng, Z., Marakakis, I., Lu, Y., & Wang, R. (2022). Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. *Health and Place, 74*(135). https://doi.org/10.1016/j.healthplace.2022.102744

Hall, C. M., & Müller, D. (Eds.). (2018). *The Routledge handbook of second home tourism and mobilities*. Routledge. https://doi.org/10.4324/9781315559056.

Han, S., Williamson, B. D., & Fong, Y. (2021). Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making, 21*(1), 1–9. https://doi.org/10.1186/s12911-021-01688-3

Harvey, D. (1973). *Social justice and the city*. University of Georgia Press.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ, 6*(e5518). https://doi.org/10.7717/peerj.5518

Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hoekstra, J. (2009). Two types of rental system? An exploratory empirical test of Kemeny's rental system typology. *Urban Studies, 46*(1), 45–62. https://www.jstor.org/stable/43198460.

Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach. *International Journal of Strategic Property Management, 24*(3), 140–152.

Hu, C., Huang, R., & Li, H. (2022). Prediction and analysis of rental price using random forest machine learning technique take Shanghai and Wuhan for example. In G. V. Bhau, Y. Shvets, & H. Mallick (Eds.), *International conference on mathematical statistics and economic analysis*. https://doi.org/10.2991/978-94-6463-042-8_84

Hu, L., He, S., & Su, S. (2022). A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. *Computers, Environment and Urban Systems, 94*, Article 101775. https://doi.org/10.1016/j.compenvurbsys.2022.101775

IHRU. (2015). *1987–2011: 25 anos de esforço do Orçamento do Estado com habitação*. IHRU, Portuguese Institute of Housing and Urban Rehabilitation. https://www.portaldahabitacao.pt/documents/20126/58203/Esforco-do-Estado-em-Habitacao.pdf/732f4eae-210a-8df0-045b-e62d9c663d78?t=1549877472657.

Kauko, T., Hooimeijer, P., & Hakfoort, J. (2002). Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies, 17*(6), 875–894. https://doi.org/10.1080/02673030215999

Knox, P., & Pinch, S. (2014). *Urban social geography: An introduction*. Routledge.

Levantesi, S., & Piscopo, G. (2020). The importance of economic variables on London real estate market: A random forest approach. *Risks, 8*(4), 1–17. https://doi.org/10.3390/risks8040112

Lord, A., Dunning, R., Dockerill, B., Burgess, G., Carro, A., Crook, T., … Whitehead, C. (2018). *The incidence, value and delivery of planning obligations and community infrastructure levy in England in 2016–17*. London: Ministry of Housing, Communities and Local Government. https://www.cchpr.landecon.cam.ac.uk/system/files/documents/DownloadTemplate_27.pdf.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics, 51*(5), 1178–1208. https://doi.org/10.1111/1540-6229.12397

Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research, 12*(1), 134–150. https://doi.org/10.1108/JERER-08-2018-0035

Mendes, L. (2018). Tourism gentrification in Lisbon: The panacea of touristification as a scenario of a post-capitalist crisis. In D. Isabel (Ed.), *Crisis, austerity and transformation: How disciplinary neoliberalism is changing Portugal* (pp. 25–48). Lexington.

Mohd, T., Masrom, S., & Johari, N. (2019). Machine learning housing price prediction in Petaling jaya, Selangor, Malaysia. *International Journal of Recent Technology and Engineering, 8*(2 special issue 11), 542–546. https://doi.org/10.35940/ijrte.B1084.0982S1119

Nelson, R., Warnier, M., & Verma, T. (2024). Housing inequalities: The space-time geography of housing policies. *Cities, 145*, Article 104727. https://doi.org/10.1016/j.cities.2023.104727

Nijman, J., & Wei, Y. D. (2020). Urban inequalities in the 21st century economy. *Applied Geography, 117*, Article 102188. https://doi.org/10.1016/j.apgeog.2020.102188

Oller, P., Baeza, C., & Furdada, G. (2021). Empirical α-β runout modelling of snow avalanches in the Catalan Pyrenees. *Journal of Glaciology, 67*(266), 1043–1054. https://doi.org/10.1017/jog.2021.50

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house Price modeling. *Journal of Real Estate Research, 32*(3), 289–320. https://doi.org/10.1080/10835547.2010.12091282

Oxley, M. (2004). *Economics, planning and housing*. Palgrave Macmillan.

Reichle, P., Fidrmuc, J., & Reck, F. (2023). The sharing economy and housing markets in selected European cities. *Journal of Housing Economics, 60*, Article 101914. https://doi.org/10.1016/j.jhe.2023.101914

Rey-Blanco, D., Arbués, P., López, F. A., & Páez, A. (2023). Using machine learning to identify spatial market segments. A reproducible study of major Spanish markets. *Environment and Planning B: Urban Analytics and City Science, 51*(1), 89–108. https://doi.org/10.1177/23998083231166952

Ribeiro, P. F., Santos, J. L., Canadas, M. J., Novais, A., Moreira, F., & Lomba, Â. (2021). Explaining farming systems spatial patterns: A farm-level choice model based on socioeconomic and biophysical drivers. *Agricultural Systems, 191*, Article 103140. https://doi.org/10.1016/j.agsy.2021.103140

Rico-Juan, J. R., & de La Paz, P. T. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*, Article 114590. https://doi.org/10.1016/j.eswa.2021.114590

Rodrigues, P. (2022). *The real estate market in Portugal: Prices, rents, tourism and accessibility*. Fundação Francisco Manuel dos Santos.

Ryan-Collins, J. (2018). *Why can't you afford a home?* John Wiley & Sons.

Ryan-Collins, J. (2021). Breaking the housing–finance cycle: Macroeconomic policy reforms for more affordable homes. *Environment and Planning A: Economy and Space, 53*(3), 480–502. https://doi.org/10.1177/0308518X19862811

Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture, 6*, 257–265. https://doi.org/10.1016/j.aiia.2022.11.003

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere, 8*(11). https://doi.org/10.1002/ecs2.1976

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences, 3*(2), 1–12. https://doi.org/10.1007/s42452-021-04148-9

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Savills Research and HomeAway. (2018). *Second homes: Global trends in ownership and renting* (Vol. 44, Issue 0). https://pdf.euro.savills.co.uk/global-research/second-homes—2019.pdf.

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

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 134*(3), 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Stamou, M., Mimis, A., & Rovolis, A. (2017). House price determinants in Athens: A spatial econometric approach. *Journal of Property Research, 34*(4), 269–284. https://doi.org/10.1080/09599916.2017.1400575

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price prediction via improved machine learning techniques. *Procedia Computer Science, 174*(2019), 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Vaidynathan, D., Kayal, P., & Maiti, M. (2023). Effects of economic factors on median list and selling prices in the U.S. housing market. *Data Science and Management, 6*(4), 199–207. https://doi.org/10.1016/j.dsm.2023.08.001

Waddell, P., & Besharati-Zadeh, A. (2020). A comparison of statistical and machine learning algorithms for predicting rents in the San Francisco Bay Area. In *98th annual meeting of the transportation research board*. https://doi.org/10.48550/arXiv.2011.14924

Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Science, 4*(6), 165–171. https://doi.org/10.20852/ntmsci.2018.327

Whitehead, C. (1999). Urban housing markets: Theory and policy. In *, Vol. 3. Handbook of regional and urban economics* (pp. 1559–1594). https://doi.org/10.1016/S1574-0080(99)80009-1

Whitehead, C., & Goering, J. (2021). Local affordable housing dynamics in two global cities: Patterns and possible lessons? *International Journal of Urban Sciences, 25* (sup1), 241–265. https://doi.org/10.1080/12265934.2020.1828147

Whitehead, C., Scanlon, K., Voigtländer, M., Karlsson, J., Blanc, F., & Rotolo, M. (2023). *Financialization in 13 cities - an international comparative report*.

Wittowsky, D., Hoekveld, J., Welsch, J., & Steier, M. (2020). Residential housing prices: Impact of housing characteristics, accessibility and neighbouring apartments – A case study of Dortmund, Germany. *Urban, Planning and Transport Research, 8*(1), 44–70. https://doi.org/10.1080/21650020.2019.1704429

Yoshida, T., Murakami, D., & Seya, H. (2022). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. *Journal of Real Estate Finance and Economics, 69*, 1–28. https://doi.org/10.1007/s11146-022-09929-6

Zheng, S., Hu, W., & Wang, R. (2016). How much is a good school worth in Beijing? Identifying Price premium with paired resale and rental data. *The Journal of Real Estate Finance and Economics, 53*(2), 184–199. https://doi.org/10.1007/s11146-015-9513-4

Zhou, X., Tong, W., & Li, D. (2019). Modeling housing rent in the Atlanta metropolitan area using textual information and deep learning. *ISPRS International Journal of Geo-Information, 8*(8), 349. https://doi.org/10.3390/ijgi8080349