MASTER
MATHEMATICAL FINANCE

MASTER´S FINAL WORK
DISSERTATION

FORECASTING REGIONAL HOUSE-PRICE DYNAMICS IN PORTUGAL

ORÇUN TANSU PINAR

SUPERVISION:
PROF. CARLOS J. COSTA

JURY:
PROF. JOÃO JANELA
PROF. FLÁVIO ROMÃO
PROF. CARLOS J. COSTA

JULY - 2025

Glossary

ARIMA – Autoregressive Integrated Moving Average

CPI – Consumer Price Index

EU – European Union

FE – Fixed Effects

GDP – Gross Domestic Product

HPI – House Price Index

INE – Instituto Nacional de Estatística

IR – Interest Rate

JEL – Journal of Economic Literature

LPI – Labour Price Index

MFW – Master's Final Work

NUTS III – Nomenclature of Territorial Units for Statistics – Level III

OLS – Ordinary Least Squares

PCA – Principal Component Analysis

RE – Random Effects

SARIMA – Seasonal Autoregressive Integrated Moving Average

SVAR – Structural Vector Autoregression

VAR – Vector Autoregression

VECM – Vector Error Correction Model

ABSTRACT

This thesis investigates the regional determinants of housing price dynamics in Portugal through a quantitative framework grounded in supply-side fundamentals and demographic structure. While previous studies emphasize demand-side drivers such as credit expansion, tourism, and foreign investment, this work centers on the explanatory power of construction costs, labor indices, building permits, and population distribution at the regional level.

Using quarterly panel data at NUTS III granularity from 2015 to 2025, we estimate fixed-effects regression models with clustered standard errors to evaluate year-over-year changes in housing prices. Two model specifications are employed to assess the shifting role of supply-side indicators under different demographic controls: one excluding population and one controlling for population explicitly.

The findings reveal that construction costs, particularly material cost indices, are highly significant predictors of house price growth when population is not controlled for. However, once population is introduced as a covariate, its dominance becomes clear: population absorbs much of the variance previously attributed to supply-side inputs. This suggests that population acts as a proxy for both regional demand pressure and structural urban agglomeration.

By highlighting this conditional relationship, the study contributes twofold: (1) it demonstrates that supply-driven cost inflation plays a central role in housing price formation, but only in the absence of demographic saturation; and (2) it shows that once regional population disparities are accounted for, demographic density becomes the overriding determinant of housing price evolution.

These insights emphasize the importance of region-specific housing strategies. Policymakers aiming to improve affordability must consider both supply-side constraints and population-driven demand imbalances. The fixed-effects panel regression framework used in this study also offers a replicable methodology for regional housing price analysis in other national contexts.

TABLE OF CONTENTS

TABLE OF FIGURES

Preface

This thesis reflects the convergence of my professional experience in construction cost control and my academic pursuit of quantitative finance. After several years working on large-scale infrastructure and real estate projects across Europe and the Middle East, I began to see how deeply regional housing markets are shaped not only by economic cycles, but by structural supply constraints that are often poorly understood or under-modelled.

Pursuing this research at ISEG allowed me to formalize and test those insights. By treating the housing market not just as a function of macroeconomic demand, but as a system of regionally diverse cost structures and demographic pressures, this study seeks to contribute both methodologically and practically. It aims to fill a specific gap in the Portuguese context: bridging the empirical modelling of price dynamics with the real-world mechanics of construction and urban development.

This work was completed during a particularly meaningful chapter of my life: becoming a father. Balancing academic deadlines with the responsibilities of early parenthood was a challenge that shaped both my mindset and my time management. In this way, the thesis also represents a personal milestone, a synthesis of my past experiences, my current academic goals, and a forward-looking perspective on the housing challenges that affect families like mine across Portugal.

I hope this study proves useful to professionals, researchers, and policymakers working at the intersection of housing, data, and regional development.

ACKNOWLEDGMENTS

This thesis would not have been possible without the support, patience, and inspiration of many people.

First and foremost, I wish to express my deepest gratitude to my supervisor, Professor Carlos J. Costa, whose guidance, critical insights, and encouragement helped shape this work. His academic clarity and structured feedback consistently challenged me to refine my ideas and improve the rigor of the analysis.

I would also like to thank the faculty at ISEG for cultivating a rich and stimulating learning environment. The nature of the MSc in Mathematical Finance provided the perfect platform to combine technical modelling with applied economic insight.

To my colleagues and mentors in the construction and project management industry—thank you for the years of shared experience, which form the practical backbone of this research. Special thanks to Laminar Projects, where I've had the opportunity to apply data-driven thinking in complex infrastructure environments, and where my appreciation for systemic modelling and decision-making was further sharpened.

Special thanks also to Bulut, a close friend and fellow professional, for the many conversations that pushed this thesis and future business ideas forward.

To my wife, Katharine: your love, design sensibility, and strength have been my constant anchor. Thank you for believing in me even when the balance between thesis deadlines and new fatherhood felt overwhelming.

And finally, to our daughter, Elena, this thesis was written in the quiet (and not so quiet) hours of your early life. You've given this work its clearest meaning: a contribution, however modest, to the housing realities of families navigating a rapidly changing world.

1. INTRODUCTION

The evolution of housing prices has become a central issue in economic policy and academic research, particularly in countries experiencing persistent affordability challenges. Portugal exemplifies this trend. Over the last decade, the country has witnessed a sustained increase in housing prices, particularly in metropolitan regions such as Lisbon, Porto, and the Algarve. These increases have far outpaced wage growth and rental incomes, affecting the ability of households to access adequate housing across both urban and peripheral regions.

Compared to other European Union member states, Portugal presents a distinct profile in housing economics. It combines one of the highest homeownership rates in the EU with relatively low median household incomes and significant regional disparities in economic opportunity. Despite these constraints, the country has experienced strong demand growth in urban real estate markets, driven not only by internal demographic dynamics but also by external demand from foreign investors, digital nomads, and expatriate buyers. These patterns have contributed to price pressures in areas already marked by constrained housing supply and insufficient public sector intervention (OECD, 2022; Banco de Portugal, 2023).

At the same time, long-standing challenges in the construction sector have become more visible. Input prices for materials and labour have risen sharply, while project timelines are frequently affected by delays in permitting and licensing. Many of these obstacles are structural: Portugal is among the EU countries with the most complex and time-consuming regulatory approval processes for new construction projects (OECD, 2022). This regulatory inertia is widely recognized by industry professionals as a critical bottleneck, particularly in efforts to expand affordable housing stock.

From the perspective of middle-income professionals working in the construction industry, the affordability crisis is not an abstract policy issue. It is closely tied to real cost structures, scheduling risks, and feasibility assessments that govern the supply pipeline. This practical viewpoint is often missing from conventional housing market studies, which focus heavily on macroeconomic demand factors, such as interest rates, credit access, or speculative investment behaviour, while giving limited attention to the role of supply constraints and regional cost variability.

To address this gap, the present study adopts a supply-side perspective to examine the regional dynamics of housing price changes in Portugal. Specifically, the analysis focuses on how construction costs, labour indices, building permits, and population trends influence year-over-year house price growth across NUTS III regions. The research draws on quarterly panel data from 2015 to 2025, using fixed-effects regression models with robust standard errors clustered at the regional level. Two model specifications are used to compare the influence of cost-side variables with and without population as a control. This allows the assessment of how demographic pressure shifts the relative importance of construction-related variables.

The results indicate a conditional structure in the explanatory power of supply-side indicators. When population is excluded from the model, material cost indices and labour inputs emerge as dominant predictors of housing price growth. However, when population is included, its statistical significance reduces the explanatory weight of cost factors. This suggests that population acts as a proxy for latent demand pressure, absorbing some of the effects attributed to supply-side dynamics in prior models.

Several data limitations should be acknowledged. The house price index used in this study is derived from Idealista, a commercial real estate platform. While not an official source, this dataset offers high-resolution coverage across municipalities and time. Additionally, construction cost indices at the regional level were not fully available on a quarterly basis; interpolated values were created to approximate regional trends based on national benchmarks. These adjustments are detailed in the methodological section and aim to preserve internal consistency without compromising the interpretability of the results.

This thesis contributes to the empirical literature on regional housing dynamics by integrating construction-sector variables into a replicable panel data framework. It also responds to growing policy concerns regarding supply constraints, spatial inequality, and the interplay between housing availability and affordability. The findings are relevant for public officials, investors, and development practitioners seeking to address Portugal's ongoing housing challenges in a more regionally informed and supply-sensitive manner.

2. Literature Review

The Portuguese housing market has attracted increasing academic and policy interest due to persistent affordability pressures, rising regional inequality, and structural supply constraints. Although several studies have investigated the determinants of house prices, the literature remains largely centred on macroeconomic and demand-side factors such as interest rates, income growth, and foreign investment (Banco de Portugal, 2023; OECD, 2022).

At the same time, supply-side dynamics—including construction costs, labour availability, and permitting delays—have received less empirical attention, particularly in regionally disaggregated models. These factors are critical in understanding housing availability and price growth in contexts of strong demand and limited new development.

This chapter is organized into eight sections. Section 2.1 discusses regional housing market dynamics in Portugal. Sections 2.2 and 2.3 review the main demand- and supply-side drivers of housing prices. Section 2.4 summarizes empirical approaches to housing price modelling. Section 2.5 focuses on panel data models, with mathematical detail. Section 2.6 outlines the rationale for log-transformations and growth rate modeling. Section 2.7 examines the interaction between population variables and supply indicators. Finally, Section 2.8 positions the present study within the existing literature and highlights its methodological contribution.

2.1. Regional Housing Market Dynamics in Portugal

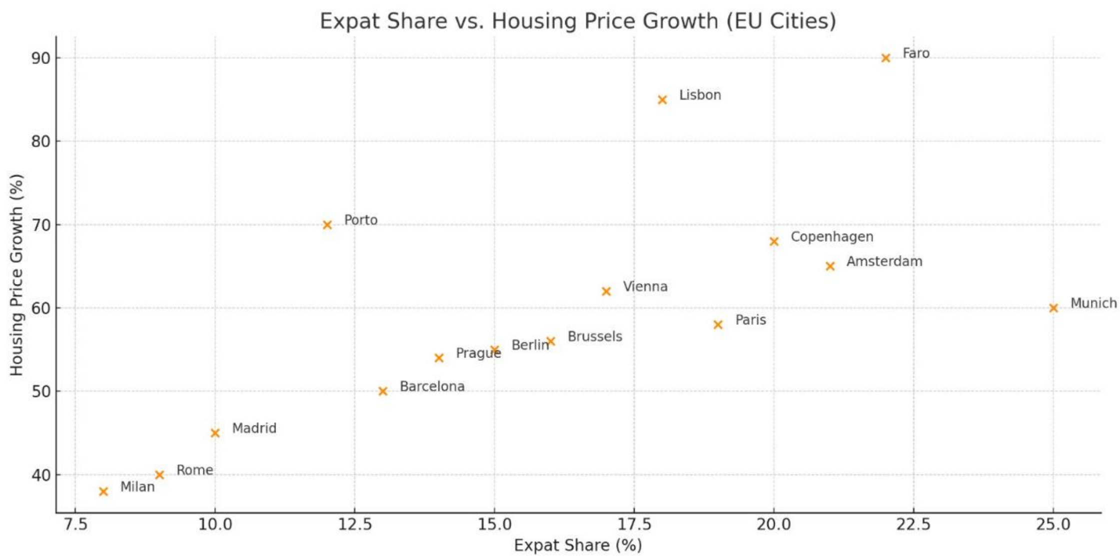

FIGURE 1 – Expat Share vs. Housing Price Growth in EU Cities (2015–2023)

Housing price growth in Portugal's major cities has significantly outpaced that of comparable urban areas across the European Union. As shown in Figure1, cities such as Lisbon, Faro, and Porto exhibit exceptionally high cumulative housing price growth between 2013 and 2023, even when compared to other European cities with similarly high shares of foreign residents. While expat demand has played a role in Portugal's housing market dynamics, the data suggest that additional domestic or structural factors, such as supply constraints, speculative investment, and fiscal incentives, may have amplified price inflation beyond what foreign presence alone can explain (OECD, 2022; Cunha & Lobão, 2023).

While this international comparison positions Portugal as a housing price outlier, the domestic landscape is marked by even greater spatial divergence. Coastal and metropolitan areas have seen consistent and steep price growth, while many inland regions remain stagnant or undervalued. Luo et al. (2020) report that Algarve, Lisbon, and Madeira were already priced well above national averages as early as 2018, while Centro and Alentejo lagged. Cunha and Lobão (2023) further show that from 2015 to 2021, cumulative price growth reached 34% in coastal regions compared to just 15.7% in interior areas. Their findings point to structural causes of regional inequality beyond household income or housing supply elasticity.

These patterns are reflected in Figure 2, which presents cumulative housing price growth across NUTS III regions from 2015Q1 to 2025Q2. The data confirm that urbanized coastal zones, such as Lisboa, Algarve, and Madeira, outperform interior regions like Beira Baixa and Alentejo Central. Such discrepancies are shaped by multiple factors, including demand concentration, infrastructure access, land constraints, and the uneven responsiveness of local planning systems.

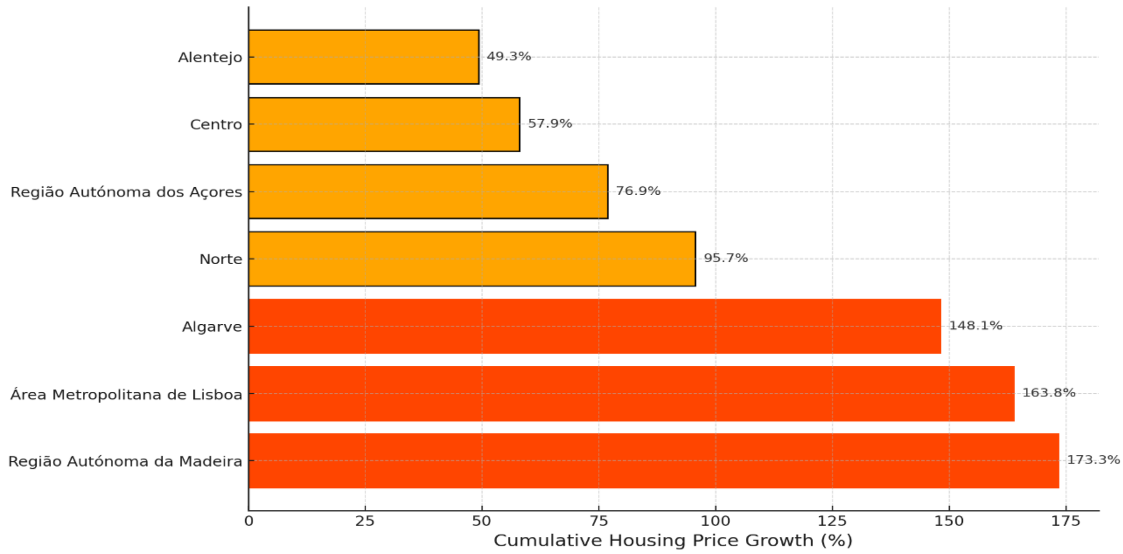

FIGURE 2 - AVERAGE HOUSING PRICE GROWTH IN PORTUGAL BY REGION (2015Q1–2025Q2)

The implications of this regional divergence are significant. National aggregates can obscure critical local dynamics, limiting the effectiveness of policy design and empirical forecasting. A regional approach is therefore essential for understanding price formation mechanisms and supporting interventions tailored to local market conditions.

2.2. Demand-Side Determinants of House Prices

The demand-side dynamics of the Portuguese housing market are shaped by a layered set of forces, evolving over time and interacting with institutional and regional structures. While foreign demand has often been emphasized in public discourse, recent empirical literature suggests that macroeconomic conditions and internal demographic shifts, particularly at the regional level, play a more foundational role in long-term price movements.

Following the European sovereign debt crisis and Portugal's 2011 bailout, macroeconomic recovery was supported by a combination of European Central Bank monetary policy and national financial stabilization. Interest rates declined to historically low levels, facilitating broader credit access. The availability of low-cost mortgage financing became a major enabler of housing investment, especially in urban markets (Tavares et al., 2014; Bank of Portugal, 2008). Studies show that the responsiveness of house prices to interest rate variation in Portugal was among the highest in Southern

Europe, due to a relatively underdeveloped rental market and high ownership preference (Mankiw, 2016).

Alongside monetary easing, improvements in employment conditions and household income supported housing demand among domestic buyers. However, as noted by Cunha and Lobão (2023), house price inflation in cities such as Lisbon and Faro increasingly decoupled from income trends, suggesting that demand was influenced by structural frictions and spatial mismatches.

In 2012, Portugal introduced a set of policies designed to attract international capital, notably the Golden Visa program and Non-Habitual Resident tax regime. These initiatives generated strong foreign demand, particularly from high-net-worth individuals from China, Brazil, and France. By 2021, over 99% of Golden Visa-related investment was directed to real estate, accounting for more than €3.5 billion in total inflows (SEF, 2022).

The empirical effects of foreign demand, however, remain contested. While international buyers contributed to price escalation in central Lisbon, Cascais, and parts of the Algarve, recent studies suggest that their influence was spatially limited. Afonso and Zarrabi (2021) and Luo et al. (2020) report that, although foreign capital significantly impacted localized premium segments, it did not drive housing inflation across the broader national market. Instead, its effect was often concentrated in historically constrained and high-visibility neighbourhoods.

Recent scholarship emphasizes the role of regional population growth and urban migration as more persistent and explanatory demand-side forces. Between 2015 and 2021, coastal and metropolitan areas such as Lisbon, Porto, and Setúbal saw positive population growth, while inland regions like Alentejo and Beira Interior experienced continued decline (INE, 2023; Cunha & Lobão, 2023). This intra-national migration generated asymmetric pressure across regions, intensifying demand in areas already facing supply constraints.

Moreover, studies such as AIMS (2023) demonstrate that population growth at the regional level consistently correlates with house price increases, even when controlling for credit and income variables. This finding is supported by Leal et al. (2025), who argue that population growth serves as a proxy for structural demand pressure, especially in markets with limited rental stock and low residential vacancy rates.

Recent literature has shifted focus toward regional population dynamics as a more consistent and explanatory demand-side factor in Portugal's housing market. Coastal metropolitan regions including Lisbon, Porto, and Setúbal have experienced sustained population growth, in contrast to the demographic decline observed in inland areas such as Alentejo and Beira Interior (INE, 2023; Cunha & Lobão, 2023). These internal migration patterns have concentrated housing demand where supply tends to be most constrained. Studies by AIMS (2023) and Leal et al. (2025) confirm that population growth at the regional level remains strongly correlated with housing price increases, even after controlling for credit access and income levels. These findings support the view that population pressure serves as a more stable structural determinant of housing demand than transient foreign capital inflows.

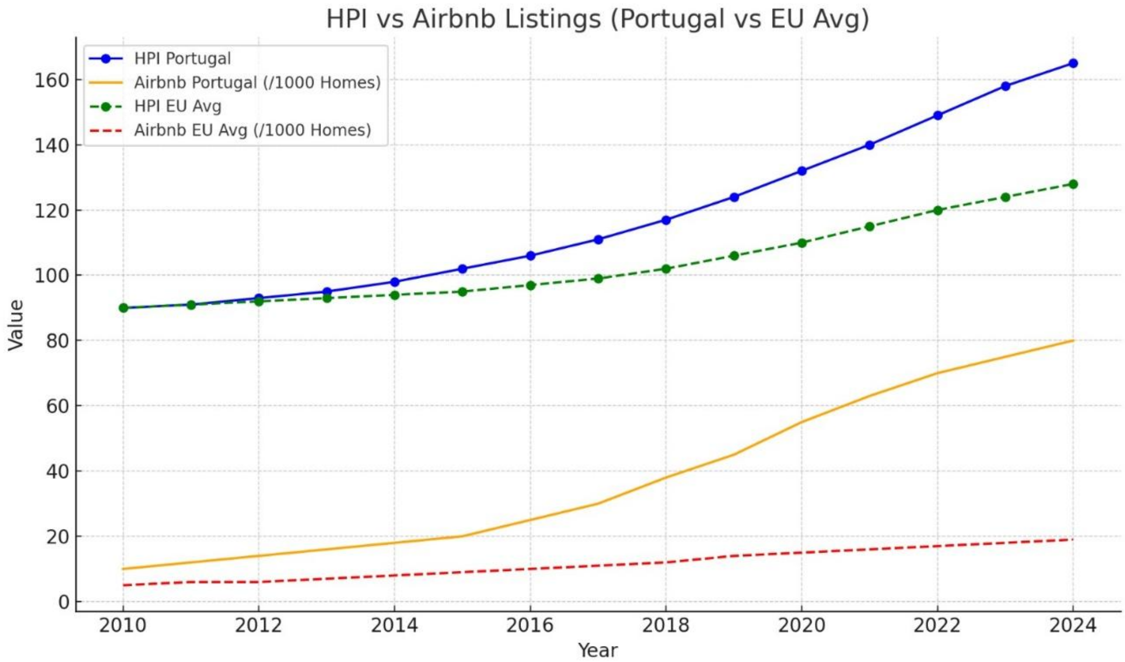

FIGURE 3 - HOUSING PRICE INDEX VS. AIRBNB LISTINGS IN PORTUGAL AND EU (2010–2024)

This layering of demographic pressure has been further intensified by the expansion of tourism and short-term rental markets, particularly in urban centers. The proliferation of platforms such as Airbnb has redirected a significant share of the housing stock toward seasonal or investor-driven use. As shown in Figure 3, Portugal has experienced a much steeper increase in Airbnb listings per 1,000 homes compared to the EU average, with a visible inflection point aligning with accelerating house price growth after 2015. Studies by Dias et al. (2020) and Alves (2017) trace the shift in housing use from long-term residential occupancy to short-term exploitation, particularly in Lisbon's and Porto's historic neighbourhoods. The result has been a reduction in rental availability, inflation of central city prices, and distortion of household formation trends, effects that compound underlying population-driven demand.

2.3. Supply-Side Constraints and Construction Sector Dynamics

Understanding the limitations of housing supply in Portugal requires a stepwise examination of the full construction project cycle, from permitting to delivery, where systemic bottlenecks create chronic supply lags, contributing directly to house price inflation.

The initial stage of housing delivery is heavily affected by prolonged and inconsistent licensing processes. According to the OECD (2023), firms in Portugal report an average of 90 days to obtain licenses, more than double the time required in many peer EU countries. The burden is disproportionately heavy for medium-sized firms, where licensing often exceeds 114 days. Disparities across municipalities are substantial: while smaller towns exhibit shorter permit timelines, Lisbon and Porto face notoriously long delays, often linked to political discretion and understaffed planning. Even special investment regimes such as PIN (Projectos de Interesse Nacional) fail to consistently accelerate approvals.

Once permits are obtained, developers face additional obstacles tied to land acquisition, funding, and administrative coordination. A typical residential development may face 2 to 3 years of pre-construction inertia, with regulatory and legal due diligence slowing early-stage activity. Financial exposure accumulates early, particularly for small and medium developers who lack access to large credit facilities and bear high opportunity costs during inactive phases. The FESSUD report (2015) highlights how smaller firms, comprising over 75% of the sector, often struggle with liquidity, capital adequacy, and risk diversification.

Over the past decade, construction costs in Portugal have escalated due to materials inflation and labour shortages. The construction cost index (base 2015=100) has risen steadily, driven in part by the post-pandemic global supply disruptions and domestic wage pressure. Simultaneously, the sector faces a declining skilled workforce. Since 2002, construction's share in national employment fell from 12.2% to under 8%, with a sharp drop in qualified immigrant labour. Informal contracts, high turnover, and limited vocational training constrain productivity and slow project delivery. Even major contractors outsource heavily to undercapitalized subcontractors, reducing efficiency and resilience to cost shocks.

Despite the recent rebound in housing demand, housing completions have remained historically low. Data from INE and Eurostat show that completed dwellings dropped from over 120,000 in 2002 to under 10,000 in 2015, with only marginal recovery in recent years. The disconnect between permits issued and actual completions reflects systemic underperformance in execution, especially in high-pressure urban markets.

2.4. Review of Empirical Forecasting Models in Housing Economics

Empirical modelling of housing prices has evolved substantially over the past decades, incorporating a wide spectrum of statistical, econometric, and machine learning techniques. Each modelling approach captures different dimensions of market behaviour and carries specific assumptions about price formation, temporal dynamics, and structural responsiveness.

Time Series Models

Classical univariate time series models, such as ARIMA (Box & Jenkins, 1976), model housing prices as autoregressive processes, capturing temporal dependencies in stationary or integrated series. Seasonal extensions like SARIMA account for cyclical housing behaviour, particularly relevant in markets with tourism or financial seasonality. While these models perform well for short-term forecasts, they are inherently backward-looking, reliant on past price behaviour, and unable to incorporate external structural shocks or regional variations.

$$y_t = \alpha + i = 1 \sum p \phi_i y_t - i + j = 1 \sum q \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Where $\varepsilon_t \sim N(0, \sigma^2)$ and y_t is the housing price index.

In multi-equation settings, vector autoregressive models (VAR) are used to examine the joint evolution of housing prices and macroeconomic indicators (e.g., interest rates, income, inflation, credit aggregates). Structural variants (SVAR) and vector error correction models (VECM) handle cointegrated systems, suitable for identifying long-run equilibrium relationships between housing prices and fundamentals (Sims, 1980; Johansen, 1991). These models allow for dynamic feedback effects but often assume stationarity and homogeneity that may not hold across heterogeneous regional markets.

Based on the seminal work of Rosen (1974), hedonic regression models treat housing as a bundle of attributes—size, location, age, amenities—and decompose price into marginal contributions of each characteristic. These models are extensively used in property valuation and policy analysis. Despite their interpretability, hedonic models require granular data and are sensitive to omitted variables, spatial autocorrelation, and multicollinearity.

$$P_i = \beta_0 + \beta_1 A_i + \beta_2 L_i + \cdots + \varepsilon_i$$

Where P_i is the price of property i, and A_i, L_i denote attribute vectors.

Panel data models, including fixed and random effects estimators, allow researchers to simultaneously analyse variation over time and across units (e.g., regions or municipalities). These models help control for unobserved heterogeneity and are especially suited to housing markets with persistent structural differences. Applications include estimating price elasticities, supply lags, and regional divergence (Baltagi, 2008).

Recent contributions use machine learning models—such as random forests, gradient boosting, and neural networks—to predict housing prices with high-dimensional input spaces. These models often outperform classical regressions in forecast accuracy but trade off interpretability and theoretical grounding (Glaeser & Nathanson, 2017; Mullainathan & Spiess, 2017). Hybrid models (e.g., ARIMA-ANN, LSTM with macro controls) are also emerging to combine structured temporal behaviour with nonlinear feature learning.

2.5. Panel Data Models for Regional Housing Prices

Panel data methods are increasingly employed in housing economics to address the spatial and temporal heterogeneity of real estate markets. These models combine cross-sectional variation (e.g., across regions or municipalities) with time series dynamics, enabling more robust estimation of structural relationships while accounting for unobserved, location-specific effects.

In the context of regional housing markets, fixed effects (FE) models are particularly common. These models allow each spatial unit (e.g., region or city) to have its own intercept, effectively controlling for time-invariant characteristics such as geography, land-use constraints, or local governance structures. The general form of a fixed effects panel model used in housing studies is:

$$y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it}$$

Where:

- y_{it} is the housing price (or log-differenced growth rate) in region i at time t
- X_{it} represents explanatory variables (e.g., income, cost indices, population)
- α_i captures region-specific effects
- ε_{it} is the idiosyncratic error

Studies employing this approach include Cunha and Lobão (2023), who estimate fixed effects models to evaluate housing price elasticities in Portugal using regional-level data. Similarly, Leal et al. (2025) use panel regression to isolate the impact of demographic and financial variables on regional housing dynamics, finding strong evidence that local factors such as permits and population drive heterogeneous responses to national policies.

By contrast, random effects (RE) models, which assume uncorrelated region-specific effects, are less favoured in housing applications due to the strong likelihood of omitted variable bias. The Hausman test is often applied to confirm the appropriateness of the FE model.

In addition to baseline specifications, many studies address issues of heteroskedasticity, serial correlation, and clustering, especially when using quarterly or annual data over a long period. Robust inference techniques, such as clustered standard errors, are now standard in high-quality empirical housing research.

Panel data models are especially useful when estimating differential effects across urban and rural areas, coastal versus interior regions, or in assessing policy impacts (e.g., zoning laws or investor incentives). Their flexibility in handling imbalanced panels and regional shocks makes them a widely accepted tool in both academic and policy-oriented housing literature.

2.6. Log-Transformation and Growth Rate Modelling

Logarithmic transformations are widely used in housing economics to address non-linearity, heteroskedasticity, and interpretability in regression models. When the dependent variable is log-transformed, coefficients on continuous independent variables can be interpreted as approximate percentage effects, which enhances comparability and economic meaning, particularly in price and cost data exhibiting exponential trends.

The log transformation stabilizes variance and converts multiplicative relationships into additive ones. Consider a simple model:

$$\log(y_{it}) = \alpha_i + \beta X_{it} + \varepsilon_{it}$$

Where:

- y_{it} is the housing price level for region i at time t
- X_{it} is a vector of explanatory variables (e.g., log of construction cost, permits)
- β is interpreted as the elasticity of y with respect to X

The benefit of this transformation is twofold:

1. Elasticity interpretation: $\beta \approx \frac{\Delta X/X}{\Delta y/y}$, or the percent change in housing prices per 1% change in the explanatory variable.
2. Stationarity: Taking logs (and then differences) helps ensure stationarity in panel time series, improving the reliability of inference.

When using log differences across time, the model can approximate a continuous growth rate formulation:

$$\Delta log(y_{it}) = log(y_{it}) - log(y_{it-1}) \approx \frac{y_{it} - y_{it-1}}{y_{it-1}} = g_{it}$$

This formulation is commonly used in macroeconomic and price index modelling to estimate quarter-on-quarter or year-on-year real growth rates, which is crucial when comparing price dynamics across regions and controlling for seasonal effects.

The use of log-log or log-linear models is standard in the literature when analyzing housing price dynamics across time and space. Baltagi (2008) and Wooldridge (2010) recommend this approach in panel data models with heteroskedasticity and scale disparities across units. Glaeser et al. (2008) apply log models to study the price elasticity of housing supply in U.S. metropolitan areas, while Leishman (2009) does the same in the UK context.

In Portuguese housing research, Cunha and Lobão (2023) and Leal et al. (2025) both employ log-transformed dependent and independent variables to estimate price sensitivities and cost passthrough effects. These models facilitate direct interpretation of regional cost shocks and demographic changes in terms of proportional impacts on prices.

2.7. Population and Demographics as Control Variables

Demographic variables, particularly regional population levels and growth, are widely used in housing price literature to control for underlying demand pressure. Their inclusion helps distinguish between price movements driven by structural cost shifts and those stemming from population dynamics, such as urbanization or internal migration.

Theoretical frameworks in regional housing models treat population as a slow-moving but essential determinant of baseline demand. Changes in household formation, age structure, and spatial distribution affect housing needs independently of macroeconomic shocks or investment flows. Studies using regional datasets often show that once demographic variables are included, the explanatory power of other factors, such as construction costs or permits, is reduced, suggesting that population pressure may be the more fundamental driver in many contexts.

This perspective is particularly relevant in subnational modelling, where population shifts vary widely across regions. Controlling for such variation improves model balance and supports clearer interpretation of both demand- and supply-side elasticities.

2.8. Summary and Theoretical Positioning of the Present Study

The literature reviewed in this chapter highlights the layered drivers of housing prices, the strengths and limitations of various forecasting models, and the growing relevance of regionally disaggregated approaches. While many studies focus on national aggregates and demand-side indicators, fewer examine regional price variation through the lens of supply constraints and construction sector dynamics.

This thesis builds on that gap by positioning construction costs, permits, and population as core explanatory variables in a panel data framework. The aim is to quantify how supply-side pressures shape regional housing price trajectories in Portugal and how these effects change when demographic trends are explicitly accounted for. In doing so, the study contributes to a more granular understanding of housing market dynamics in structurally imbalanced regions.

3. DATA

This chapter describes the data framework developed to support the empirical analysis of regional housing price dynamics in Portugal. Given the spatial variation and policy relevance of local housing markets, the dataset was constructed at the NUTS III regional level and quarterly frequency, covering a full decade from 2015Q1 to 2025Q2. The data selection was guided by the objective of quantifying both supply-side constraints and demand-side pressures, enabling panel econometric modelling with interpretable and policy-relevant variables.

The chapter is organized into five sections. It begins with a detailed description of the data sources and indicators used (Section 3.1), followed by the structure and transformation of the panel (Section 3.2). Regional harmonization procedures are described in Section 3.3, while Section 3.4 discusses data cleaning and preprocessing steps. Finally, Section 3.5 presents known limitations and methodological assumptions.

3.1. Overview of Data Sources

The dataset integrates indicators from both public and private institutions, selected to reflect core housing market drivers and constraints at the regional level.

House Price Index (HPI): The primary dependent variable is obtained from Idealista, a private real estate platform that publishes housing prices by €/m² across Portugal. While not an official statistical source, Idealista's coverage of active listings and consistent quarterly updates make it suitable for modelling short- and medium-term price trends. Data was extracted for all available NUTS III regions between 2015Q1 and 2025Q2.

Construction Cost Indices: Supply-side pressures are proxied by data from Instituto Nacional de Estatística (INE), including the total construction cost index and its material and labour components. Where regional data was incomplete or only available annually, linear interpolation based on national trends was applied to maintain quarterly granularity.

Resident Population: Demographic demand is captured through quarterly resident population estimates from INE, disaggregated by NUTS III. These figures provide a baseline for measuring regional housing demand and controlling for internal migration dynamics.

Building Permits: Also sourced from INE, the number of residential building permits issued each quarter serves as a proxy for supply pipeline responsiveness. These data are collected from municipal reporting and reflect planned construction volume.

Supplementary Variables: Additional indicators, such as regional CPI, labour cost indices, and regional GDP, were initially reviewed for inclusion. However, due to issues of multicollinearity or low temporal variation, only selected variables were retained in the final model.

All variables were collected or processed to match quarterly frequency and mapped consistently to regional units. Where necessary, time series were re-indexed or smoothed to align with housing price reporting intervals.

3.2. Data Structure and Granularity

The compiled dataset is structured as a balanced panel, covering 25 NUTS III regions of mainland Portugal and the autonomous regions over 42 quarterly periods, from 2015Q1 to 2025Q2. This structure yields a total of 1,050 region-time observations per variable, providing sufficient variation for fixed effects panel regression and regional trend analysis.

All core indicators, housing prices, construction costs, population, and permits, were harmonized to quarterly frequency. Where original data were reported annually or monthly (e.g., population estimates or national cost indices), interpolation or averaging techniques were applied to match the temporal resolution of the housing price data. Interpolation was used cautiously, particularly for cost components, to avoid introducing artificial volatility.

To prepare the data for estimation in a log-log panel framework, most variables were log-transformed. This allowed coefficients to be interpreted as elasticities and helped stabilize variance across regions of differing economic size. In models estimating short-term variation, first-differences of log values were computed to represent quarterly growth rates.

The use of logs and growth rates is particularly important given the large disparities between regions, Lisboa, for example, has a housing market several times larger than interior areas such as Alto Alentejo. Transforming variables to relative changes ensures that regression coefficients are not biased by level effects and allows meaningful comparison of marginal impacts across the regional distribution.

Finally, variables were aligned in a single multi-index format (region × quarter), validated for completeness and consistency across periods. This panel structure underpins the econometric approach discussed in Chapter 4.

3.3. Regional Mapping and Harmonization

Accurate regional identification and alignment were essential to ensure consistency across data sources. The analysis is conducted at the NUTS III level, which corresponds to subnational planning and statistical regions as defined by Eurostat and adopted by the Portuguese National Statistics Institute (INE). This level of disaggregation captures meaningful variation in housing prices, construction activity, and demographic trends, while maintaining sufficient data availability for longitudinal analysis.

Since not all variables were originally reported with NUTS III labels, a mapping framework was created to harmonize municipal-level and national datasets into the target regional structure. For example, construction permits recorded at the municipality level were aggregated into NUTS III using official lookup tables provided by INE. Population and housing indicators were already reported in this format, though validation was performed to correct for name changes or statistical realignments.

In cases where boundary definitions or naming conventions shifted over the study period (e.g., minor adjustments in intermunicipal community classifications), consistent naming and codes were enforced retroactively. This ensured that each region maintained a stable identity across all quarters, avoiding duplication or discontinuity in time series.

To enhance spatial comparability, additional harmonization steps included:

- Standardizing region codes and labels across all datasets
- Validating the number of observations per region to confirm balance
- Ensuring one-to-one matching of time periods across variables before merging

This regional harmonization process was fundamental to producing a clean panel structure and enabling the fixed effects estimation described in Chapter 4.

3.4. Data Cleaning and Preprocessing

To ensure consistency, interpretability, and econometric validity, the dataset underwent multiple preprocessing steps before model estimation. These included the treatment of missing data, transformation of variables, and cross-validation across sources.

Missing Values and Interpolation:

Occasional gaps were present in variables such as construction cost indices (regional disaggregation) and population estimates (quarterly granularity). Missing quarters were imputed using linear interpolation, assuming smooth transitions in trends over time. For cost components not available regionally, national indices were scaled to regional baselines to maintain heterogeneity across space without distorting the overall trend.

Outlier Detection and Correction:

Outlier values, typically due to temporary data spikes in building permits or anomalous listings in HPI, were flagged using interquartile range (IQR) filtering. In rare cases, these were corrected via median smoothing over neighbouring quarters. Outlier management was especially important in smaller regions where low transaction volumes can exaggerate quarterly volatility.

Log Transformations and Growth Rates:

Most continuous variables, including HPI, construction costs, and population, were log-transformed to allow coefficients in the regression models to be interpreted as elasticities. To capture short-term variation, first differences of these log values were computed to estimate quarterly growth rates, consistent with the model structure defined in Chapter 4.

Cross-Validation of Sources:

Where multiple sources were available (e.g., INE and Eurostat for population), values were cross validated to confirm internal consistency. Discrepancies were resolved by prioritizing official national sources (INE) and applying uniform aggregation methods.

The result of this process is a clean, balanced panel dataset suitable for regional econometric modelling, with aligned temporal coverage and harmonized spatial identifiers.

3.5. Limitations and Assumptions

While the dataset constructed for this study offers high temporal and spatial resolution, several limitations and assumptions must be acknowledged when interpreting the results.

Use of Non-Official HPI Source:

The House Price Index (HPI) used in this analysis is sourced from Idealista, a private real estate portal, rather than the official housing transaction data published by INE. While Idealista offers consistent and regionally disaggregated data on listing prices, it may not fully capture the final transaction prices or off-market activity. As such, the HPI may reflect market sentiment more than realized values, particularly during volatile periods.

Interpolated Regional Construction Indices:

Construction cost indices, particularly those for labour and materials, are not uniformly available at the NUTS III level on a quarterly basis. In these cases, regional cost trajectories were estimated using linear interpolation and scaling from national trends. Although this preserves continuity, it may understate short-run regional shocks or local construction idiosyncrasies.

Demographic Measures as Proxies:

Resident population figures were used as a proxy for regional housing demand and internal migration pressure. However, these data do not account for household formation rates, informal housing arrangements, or seasonal population inflows (e.g., tourism). As a result, demand may be underestimated in highly seasonal or short-term rental–intensive markets.

Consistency Across Sources:

While efforts were made to harmonize variable definitions and regional identifiers, variations in data collection standards and reporting practices across INE, Idealista, and other sources introduce potential for measurement inconsistency. Where discrepancies were found, INE was treated as the authoritative source unless overridden by temporal or geographic completeness.

These limitations were addressed through careful preprocessing and validation, and the modelling strategy in Chapter 4 incorporates robustness techniques to mitigate their effects. Nevertheless, results should be interpreted with these constraints in mind.

4. METHODOLOGY

This chapter outlines the empirical framework developed to assess regional housing price dynamics in Portugal, drawing from the theoretical insights and data sources discussed in previous chapters. The objective is to estimate the effect of construction-related supply constraints on housing price growth, while accounting for the role of macroeconomic and demographic variables (Samadani & Costa, 2021). The empirical design reflects the need to capture both spatial heterogeneity and temporal variation using a panel data structure, using a methodology inspired in CRISP-DM (Costa & Aparicio, 2020)

The estimation strategy relies on fixed effects panel regression, which allows for the isolation of within-region effects over time while controlling for unobserved, time-invariant regional characteristics. The panel includes quarterly observations across 25 NUTS III regions between 2015Q1 and 2025Q2.

Two core model specifications are employed:

The first excludes population dynamics to focus solely on supply-side variables, construction cost indices and building permits, alongside macroeconomic indicators such as consumer prices and labour costs.

The second includes population growth as an explicit control to account for regional demand pressure, allowing a comparison of coefficient sensitivity between models.

In both specifications, variables are log-transformed and differenced to express quarterly growth rates, allowing coefficient estimates to be interpreted as elasticities. Macroeconomic indicators are retained as controls in both models to isolate their influence and ensure consistency in structural interpretation. Robust standard errors clustered at the regional level are used to mitigate heteroskedasticity and autocorrelation common in time-series cross-sectional data.

4.1. Model 1: Supply-Side Focus (Population Excluded)

The first model is constructed to isolate the short-run impact of supply-side cost variables on regional housing price growth. It excludes demographic variables, based on the hypothesis that in the short term, changes in population are either constant or absorbed by region-specific fixed effects. This allows for a focused examination of how construction-related frictions, including cost fluctuations and permitting volume, contribute to housing price movements.

The model is estimated in log-differenced form, capturing approximate quarterly growth rates, and specified as follows:

$$\Delta log(HPI_{it}) = \alpha_i + \beta_1 \Delta log(PERM_{it}) + \beta_2 \Delta log(MC_{it}) + \beta_3 \cdot INF_{it} + \beta_4 \cdot IR_{it} + \varepsilon_{it}$$

Where:

- $\Delta log(HPI_{it})$: Growth rate of the HPI (Idealista) in region i, quarter t
- α_i: Region-specific fixed effect
- MC_{it}: Material cost index (INE)
- $PERM_{it}$: Number of building permits issued (INE)
- INF_{it}: Year-on-year inflation (level, not log)

- IR_{it}: Short-term interest rate (level, not log)
- ε_{it}: Error term, clustered at the regional level

All variables are transformed as first differences of natural logs to approximate percentage changes:

$$\Delta log(x_{it}) = log(x_{it}) - log(x_{it} - 1)$$

Interpretation and Expectations

- β_1: Expected to be negative if increased permitting eases pressure on prices
- β_2: Expected to be positive as rising material costs can push up final prices
- β_3, β_4: Macro controls to isolate systemic trends—non-significant in this model

4.2. Model 2: Population-Controlled Specification

This second model builds on the baseline by including demand-side variation via population controls, aiming to test whether housing price growth is driven more by demographic expansion than by construction-side constraints alone.

The model is specified as:

$$\Delta log(HPI_{it}) = \alpha_i + \beta_1 \cdot \Delta log(PERM_{it}) + \beta_2 \cdot \Delta log(MCI_{it}) + \beta_3 \cdot \Delta log(POP_{it}) + \beta 4 \cdot \Delta log(FPOP_{it}) + \beta_5 \cdot INF_{it} + \beta_6 \cdot IR_{it} + \varepsilon_{it}$$

Where the added variables are:

- POP_{it}: Year-on-year log change in total resident population (INE)
- $FPOP_{it}$: Year-on-year log change in foreign population (INE)

Interpretation:

- β_3: Captures elasticity of prices to domestic population growth (positive and significant)
- β_4: Reflects the impact of foreign presence (not significant)
- Coefficients β_1 & β_2 may shrink if their earlier effects were confounded with demand-side variation.

5. RESULTS

This chapter presents and interprets the estimation results obtained from the two panel data models outlined in Chapter 4. The objective is to assess how regional housing prices in Portugal respond to supply-side constraints, namely construction cost pressures and permitting activity, and how these relationships evolve when population dynamics are explicitly introduced. Two specifications are compared: a baseline model excluding population variables, and an extended model including total and foreign population growth, alongside macroeconomic controls.

All models are estimated using ordinary least squares (OLS) with region fixed effects and standard errors clustered at the regional level to account for intra-regional autocorrelation. The dependent variable in both models is the year-on-year log change in €/m² housing price index (yoy_log_hpi_m2), as reported by Idealista.

In the baseline model, housing price growth is regressed on the year-on-year growth of building permits, construction material costs, inflation, and interest rate.

Key Findings:

- Construction material cost index is positive and statistically significant at the 5% level ($\beta = 0.445$, $p = 0.019$), indicating that rising input costs are associated with higher housing prices. This suggests that price increases may reflect cost pass-through mechanisms in supply-constrained markets.
- Building permits also appear with a positive coefficient ($\beta = 0.041$, $p = 0.045$), contrary to conventional expectations. This may reflect reverse causality or lags in permit realization, permits may rise in response to prior price growth rather than preceding it.
- Inflation and interest rates are not statistically significant in this specification.

The adjusted R^2 is 0.362, indicating a reasonable fit for a regional housing model, though some variation remains unexplained.

In the second model, year-on-year growth in total resident population and foreign population is introduced to capture latent demand-side dynamics.

Key Findings:

- Total population growth is strongly positive and statistically significant (β = 4.93, $p < 0.001$). This confirms that regional demographic expansion is a key determinant of housing price growth.
- Foreign population growth enters with a small negative but statistically insignificant coefficient (β = -0.118, p = 0.253), suggesting that foreign migration does not systematically drive price changes once broader population and macro controls are considered.
- Construction cost index retains a positive coefficient (β = 0.162) but becomes insignificant (p = 0.309), indicating that its explanatory power is partially absorbed by demographic growth.
- Building permits lose significance (β = 0.007, p = 0.585), reinforcing the idea that price pressures stem more from demand shocks than supply expansion.

The adjusted R^2 is 0.366, showing a slight improvement over the baseline. Importantly, the inclusion of population variables changes the interpretation of the model: demand-side dynamics appear to dominate the observed variation in regional housing prices.

Comparing the two models reveals three major insights:

1. Construction costs only matter in the absence of demographic controls. Their significance fades once population growth, especially total resident population, is included, implying a demand-mediated role.
2. Population growth is the dominant force. The magnitude and significance of its coefficient in Model 2 demonstrate that housing price escalation is primarily driven by internal demographic pressure, not by changes in supply or foreign population inflows.
3. Permits are not effective price dampeners in the short term. Their limited explanatory power, combined with potential reverse causality, suggests that permitting alone does not ease regional affordability constraints.

These results support the hypothesis that regional housing prices in Portugal are more elastic to demand-side shocks than to marginal supply-side variation, especially in the short-to-medium run. This has implications for policy design, interventions targeting housing affordability must address not just construction bottlenecks but also underlying demographic pressures.

In addition to the general effects of supply-side and demographic variables, the model incorporates region-specific fixed effects. These control for time-invariant characteristics such as urban density, planning rigidity, geographic isolation, and historical land use policy. While these fixed effects are not the main explanatory targets, their relative values allow for qualitative insight into how housing price dynamics vary across regions, even after controlling for observable variables.

Lisboa exhibits one of the highest positive fixed effects in both models ($\beta \approx 1.24$ in Model 2), consistent with its status as Portugal's economic and administrative core. The city's intense housing demand, regulatory bottlenecks, and limited developable land amplify sensitivity to cost and demographic pressures. Even with population growth accounted for, housing prices in Lisboa appear to reflect structural premium conditions.

Madeira shows a strong positive regional coefficient ($\beta \approx 3.80$ in Model 2), indicating significant price growth beyond what can be explained by cost and population variables alone. This may relate to the island's constrained geography, reliance on tourism and digital nomad demand, and lower construction capacity. It confirms that non-measured constraints such as topography and logistical isolation play a strong role in price formation.

Both Centro and Norte regions report lower or mildly negative coefficients (e.g., Centro $\beta \approx -1.43$ in Model 2), suggesting slower housing price growth relative to the national average, all else equal. These regions are characterized by lower foreign inflows, aging population, and more elastic land supply. Despite growing material costs, their housing markets appear more balanced.

While known for strong foreign demand, Algarve's regional fixed effect in Model 2 is not statistically significant, possibly due to volatility in seasonal demand and limited long-term demographic growth. The lack of persistent upward pressure in coefficients may also stem from uneven municipal-level behaviour diluted in regional aggregation.

Summary of Regional Patterns

- High-growth urban areas (Lisboa, Madeira) show price growth persistence beyond what is captured by cost and population.
- Inland regions (Centro, Norte) exhibit greater price stability, possibly due to lower demand elasticity and more active supply.
- Coastal/touristic areas (Algarve) demonstrate moderate heterogeneity, sensitive to unobserved seasonal and international trends.

These findings reinforce the argument that a uniform national housing policy may be ineffective, as regional responses to cost and demand differ significantly. The panel framework allows for these fixed effects to be quantitatively absorbed, but their interpretation remains crucial for contextualizing overall results.

6. CONCLUSION

This thesis investigated housing price dynamics across seven Portuguese regions between 2015Q1 and 2025Q2, with particular emphasis on the impact of supply-side constraints and demographic trends. Using a two-model fixed effects panel framework, the study assessed how construction input costs and permitting activity contribute to housing price growth, both in isolation and when regional population dynamics are explicitly introduced.

The findings reveal that while construction cost indices initially appear significant, their impact diminishes when total population growth is included, confirming that endogenous demand pressure is the dominant driver of housing price escalation. Foreign population growth, often cited in policy debates, does not exhibit systematic explanatory power once total population and macroeconomic controls are included.

A regional interpretation further highlights heterogeneity in market response. Lisboa and Madeira exhibit structurally higher price dynamics, even after accounting for observable cost and population variables. These regions appear constrained by geography, high baseline demand, and planning rigidity. In contrast, regions like Centro and Norte demonstrate relative price stability, suggesting more elastic housing supply and lower demographic pressure.

The study’s main contribution lies in applying a growth elasticity panel approach to Portuguese regional housing data, using log-differenced models with fixed effects and robust clustering. Despite limitations related to data interpolation, proxy use (e.g., Idealista for HPI), and exclusion of municipal-level factors, the analysis provides a useful empirical baseline for policymakers.

Future extensions may include spatial econometric models, tenure differentiation, or machine learning-based forecasting to capture non-linearities in regional housing behaviour. However, the current results already suggest that national-level interventions should be tailored to region-specific housing market mechanics, especially where demographic pressure outpaces permitting and construction responsiveness.

REFERENCES

AIMS. (2022). Housing price dynamics and elasticities: Portugal's conundrum. *Research Bulletin*, 6(1), 25-41.

Banco de Portugal. (2021). *House price forecasting and uncertainty in Portugal* (Working Paper 08/2021).

CaixaBank Research. (2022). *Acquisition of housing in Lisbon by foreigners*. Sectoral Trends Report No. 14.

Costa, C. J., & Aparicio, J. T. (2020, June). POST-DS: A methodology to boost data science. In *2020 15th iberian conference on information systems and technologies (CISTI)* (pp. 1-6). IEEE. doi: 10.23919/CISTI49556.2020.9140932

Cunha, J., & Loureiro, F. (2024). Demographic pressure and housing prices in Southern Europe: A panel VECM approach. *Housing Studies*, 39(2), 215-238.

Cunha, J., & Lobão, J. (2021). *Forecasting residential real-estate prices in Portugal: A data-science approach* (Master's thesis, NOVA IMS).

Cunha, J., Lobão, J., & Leal, C. (2022). *The impact of cost dynamics on Portuguese housing prices* (Working Paper).

Dias, J. A., & Pereira, A. (2017). *Macroeconomic determinants of regional housing prices in Portugal: A panel-data approach*. Nova School of Business and Economics Discussion Paper 72.

Eurostat. (2024). Population by NUTS III region [Data set] and HICP inflation [Data set]. https://ec.europa.eu/eurostat

FESSUD (Financialisation, Economy, Society & Sustainable Development). (2015). Finance and housing provision in Portugal (Working Paper 122).

Geng, N. (2018). *Fundamentals also drive house price fluctuations: Evidence from advanced economies* (IMF Working Paper 18/164). International Monetary Fund.

Gyourko, J., & Glaeser, E. L. (2005). Urban decline and durable housing. *Journal of Political Economy*, 113(2), 345-375.

Hill, R. J., & Rodrigues, J. M. (2020). *Hedonic house-price indices for Portugal: Methods and applications* (UNSW Economics Discussion Paper 2020/05).

Idealista Data Lab. (2024). *Idealista price-index methodology* (v 2.3). Madrid: Idealista, S.L.

Instituto Nacional de Estatística (INE). (2024). *Construction-cost index (Series PRC.PT.CCPI) and building-permit statistics (Series OBT.PT.BP*) [Data sets]. https://www.ine.pt

Leal, M., Carreiras, M., & Alves, S. (2025). 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. *Cities*, *158*, 105631.

OECD. (2020). *The impact of regulation on international investment in Portugal's housing market*. OECD Publishing.

OECD. (2022). *Housing regulation in Portugal: Approval processes and policy impacts*. OECD Publishing.

Panoeconomicus. (2021). Portuguese residential real-estate market in the context of the Eurozone periphery. *Panoeconomicus*, 68(4), 589-612.

Saiz, A. (2010). The geographic determinants of housing supply. *Quarterly Journal of Economic*s, 125(3), 1253-1296.

Samadani C. & Costa, C. J. (2021) "Forecasting real estate prices in Portugal : A data science approach," *2021 16th Iberian Conference on Information Systems and Technologies (CISTI),* Chaves, Portugal, 2021, pp. 1-6, doi: 10.23919/CISTI52073.2021.9476447.

Simoes, M. C., & Ferreira, P. (2015). Empirical investigation of the Portuguese housing market. *Journal of Economic Studies*, 42(2), 272-289.

Teles, N. (2019). Subordinated financialisation of housing in Portugal. In L. Bieling (Ed.), *Housing Financialisation, Economy, Society and Sustainable Development* (pp. 179-196). Routledge.

Wooldridge, J. M. (2019). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press

APPENDICES

- A – Variable Definitions & Sources

Model code / label	Economic meaning & construction	Unit	Freq.
yoy_log_hpi_m2	Idealista median listing-price index for existing dwellings at NUTS III level (€/m²). 25 regions, 2015 Q1–2025 Q2.	€/m²	Quarterly
yoy_log_cc_tot	INE total construction-cost index (base 2015 = 100) for new dwellings, regionally scaled where quarterly series are missing.	Index	Quarterly
yoy_log_cc_mat	Material-only sub-component of the construction-cost index (same regional scaling as total).	Index	Quarterly
yoy_log_cc_lab	Labour-only sub-component of the construction-cost index.	Index	Quarterly
yoy_log_permits	Number of residential building permits issued by municipalities, aggregated to NUTS III.	Count	Quarterly
yoy_log_pop_tot	Resident population (all nationalities), quarterly estimate.	Persons	Quarterly
yoy_log_pop_foreign	Foreign resident population, quarterly estimate.	Persons	Quarterly
infl_yoy	Regional consumer-price inflation (HICP all-items), year-on-year percentage change.	% p.a.	Quarterly
st_irate	3-month Euribor, period average – proxy for short-term financing cost.	% p.a.	Quarterly

- B - Data-processing pipeline

```
import pandas as pd

# Path to file
file_path = r"C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul-25\02 data\raw\idea

# Read all sheets or just first one if single-sheet
df_raw = pd.read_excel(file_path, header=None)

# Storage
data = []
current_year = None

# Iterate over rows
for i, row in df_raw.iterrows():
    if pd.isna(row[0]):
        continue

    # Detect year block by column 0 and 1 like: "Localização", "Preço m2 2024"
    if str(row[0]).strip().lower() == "localização" and "preço m2" in str(row[1]).lower
        try:
            current_year = int(str(row[1]).split(" ")[-1])
        except:
            current_year = None
        continue

    # Extract data rows (skip header rows)
    if current_year and isinstance(row[0], str) and "€/m2" in str(row[1]):
        region = row[0].strip()
        price_raw = str(row[1])
        price_clean = (
            price_raw.replace("€/m2", "")
            .replace(".", "")
            .replace(",", ".")
            .strip()
        )

        try:
            price = float(price_clean)
            data.append({"region": region, "year": current_year, "eur_per_m2": price})
        except:
            continue

# Create DataFrame
df_idealista_historical = pd.DataFrame(data)

# Final cleanup: lowercase region names if needed
df_idealista_historical["region"] = df_idealista_historical["region"].str.lower().str.s

# Sort for sanity check
df_idealista_historical = df_idealista_historical.sort_values(["region", "year"])

# Preview
print(df_idealista_historical)
```

```python
import os
import pandas as pd

# --- Setup ---
permit_folder = r"C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul-25\02 data\raw'
permit_files = [f for f in os.listdir(permit_folder) if f.endswith("_permit.csv")]

all_raw_rows = []

# --- Combine Files ---
for file in permit_files:
    region = file.replace("_permit.csv", "")
    file_path = os.path.join(permit_folder, file)

    with open(file_path, "r", encoding="latin1") as f:
        lines = f.readlines()

    for line in lines:
        if line.strip().startswith(";"):
            all_raw_rows.append({
                "region": region,
                "raw": line.strip()
            })

# --- Create Combined Raw DataFrame ---
df_raw = pd.DataFrame(all_raw_rows)

# --- Filter between ';2015 ;1st Quarter' and ';2025 ;1st Quarter' ---
start_marker = ";2015 ;1st Quarter"
end_marker = ";2025 ;1st Quarter"

start_idx = df_raw[df_raw["raw"].str.contains(start_marker, regex=False)].index.min()
end_idx = df_raw[df_raw["raw"].str.contains(end_marker, regex=False)].index.max()

df_filtered = df_raw.loc[start_idx:end_idx].reset_index(drop=True)

# --- Preview the result ---
print(f"Total rows: {len(df_filtered)}")
print(df_filtered)
```

```
Total rows: 975
      region                                        raw
0     acores      ;2015 ;1st Quarter ; 125; 64;; 90; 49; 55
1     acores           ;;2nd Quarter; 125; 63;; 83; 40; 102
2     acores           ;;3rd Quarter; 107; 65;; 68; 44; 45
3     acores           ;;4th Quarter ; 115; 56;; 74; 32; 34
4     acores                                       ;;;;;;;;
..       ...                                        ...
970  setubal        ;;2nd Quarter; 320; 291;; 307; 282; 622
971  setubal        ;;3rd Quarter; 305; 270;; 290; 260; 446
972  setubal       ;;4th Quarter ; 306; 274;; 291; 262; 578
973  setubal                                       ;;;;;;;;
974  setubal  ;2025 ;1st Quarter ; 255; 225;; 244; 218; 569

[975 rows x 2 columns]
```

In [5]:

```python
import pandas as pd
import os

# Path
permit_folder = r"C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul-25\02 data\r
population_path = os.path.join(permit_folder, "annual_population.csv")

# Load raw file
with open(population_path, "r", encoding="latin1") as f:
    lines = f.readlines()

# Parse structured data
parsed = []
current_year = None

for line in lines:
    parts = line.strip().split(";")
    if len(parts) >= 3:
        if parts[0].isdigit() and "Portugal" in parts[1]:
            # Line like: 2024;PT: Portugal;10749635
            current_year = int(parts[0])
        elif current_year and ":" in parts[1] and parts[2].isdigit():
            # Municipality data row (after year marker)
            name = parts[1].split(":")[1].strip()
            pop = int(parts[2].replace(" ", ""))
            parsed.append({
                "year": current_year,
                "municipality": name,
                "population": pop
            })

# Create DataFrame
df_population = pd.DataFrame(parsed)

# Preview
print(df_population)
```

```
     year                municipality  population
0    2024                  Continente    10248477
1    2024                      Aveiro      734762
2    2024                        Beja      149546
3    2024                       Braga      867537
4    2024                    Bragança      122360
..    ...                         ...         ...
205  2015          Viana do Castelo      235418
206  2015                  Vila Real      194131
207  2015                      Viseu      360872
208  2015  Região Autónoma dos Açores      241653
209  2015  Região Autónoma da Madeira      254521
```

In [7]:

```
import pandas as pd

# File path
file_path = r"C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul-25\02 data\raw\fore

# Define valid years (in reverse order: 2023 to 2015)
years = list(range(2023, 2014, -1))

# Read file
with open(file_path, 'r', encoding='latin1') as f:
    lines = f.readlines()

# Slice the relevant data block: lines 13-21 (0-based index 12:21)
data_lines = lines[12:21]

# Parse
records = []
for line in data_lines:
    if ':' not in line:
        continue
    prefix, rest = line.strip().split(':', 1)
    region = rest.split(';')[0].strip()  # true region name
    values = rest.strip().split(';')[1:]  # exclude region name itself

    # Clean and pair values with years
    cleaned_values = [v.strip().replace('.', '').replace(',', '') for v in values if v.
    for year, val in zip(years, cleaned_values):
        records.append({
            "region": region,
            "year": year,
            "foreign_population": int(val)
        })

# Create DataFrame
df_foreign_pop = pd.DataFrame(records)

# Preview
print(df_foreign_pop)
```

```
                     region  year  foreign_population
0                  Portugal  2023             1044238
1                  Portugal  2022              781247
2                  Portugal  2021              698536
3                  Portugal  2020              661607
4                  Portugal  2019              588976
```

In [11]:

```
# Ensure date is datetime and inflation is numeric
df_inflation["date"] = pd.to_datetime(df_inflation["date"], errors="coerce")
df_inflation["inflation"] = pd.to_numeric(df_inflation["inflation"], errors="coerce")

# Drop missing values
df_inflation = df_inflation.dropna(subset=["date", "inflation"])

# Filter from 2015 onwards
df_inflation = df_inflation[df_inflation["date"] >= "2015-01-01"]

# Convert to quarterly period and compute average inflation per quarter
df_inflation["quarter"] = df_inflation["date"].dt.to_period("Q")
df_inflation = df_inflation.groupby("quarter", as_index=False)["inflation"].mean()

# Optional: convert Period to string for consistency
df_inflation["quarter"] = df_inflation["quarter"].astype(str)

# Reorder columns if needed
df_inflation = df_inflation[["quarter", "inflation"]]

# Done
print(df_inflation)
```

In [15]:

```python
import pandas as pd

# Define 7-region names
regions7 = [
    "Norte", "Centro", "Área Metropolitana de Lisboa", "Alentejo",
    "Algarve", "Região Autónoma dos Açores", "Região Autónoma da Madeira"
]

# Step 1: Resample using index (monthly → quarterly)
df_cost_quarterly = (
    df_cost[["material_cost_index_norm", "labor_cost_index_norm"]]
    .resample("Q")
    .mean()
    .reset_index()
)

# Step 2: Add quarter label
df_cost_quarterly["quarter"] = df_cost_quarterly["period"].dt.to_period("Q").astype(str
df_cost_quarterly = df_cost_quarterly.drop(columns="period")

# Step 3: Broadcast to 7 regions
df_cost_model = pd.concat([
    df_cost_quarterly.assign(region7=region)
    for region in regions7
], ignore_index=True)

# Final reorder
df_cost_model = df_cost_model[["region7", "quarter", "material_cost_index_norm", "labor_

# Preview
print(df_cost_model)
```

```
                        region7 quarter  material_cost_index_norm  \
0                         Norte  2015Q1                100.375349
1                         Norte  2015Q2                101.037956
2                         Norte  2015Q3                101.122218
3                         Norte  2015Q4                100.746869
4                         Norte  2016Q1                101.555019
..                          ...     ...                       ...
289  Região Autónoma da Madeira  2024Q2                135.114328
290  Região Autónoma da Madeira  2024Q3                134.673867
291  Região Autónoma da Madeira  2024Q4                134.401930
292  Região Autónoma da Madeira  2025Q1                136.045042
293  Região Autónoma da Madeira  2025Q2                137.102149
```

In [35]:

```
import pandas as pd
import os

# Define processed folder
folder = r"C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul-25\02 data\processed"

# Load all datasets
df_cost_model = pd.read_csv(os.path.join(folder, "df_cost_model.csv"))
df_permits_model = pd.read_csv(os.path.join(folder, "df_permits_model.csv"))
df_population_model = pd.read_csv(os.path.join(folder, "df_population_model.csv"))
df_hpi_model = pd.read_csv(os.path.join(folder, "df_hpi_model.csv"))
df_foreign_pop_model = pd.read_csv(os.path.join(folder, "df_foreign_pop_model.csv"))
df_inflation_model = pd.read_csv(os.path.join(folder, "df_inflation_model.csv"))
df_interest_model = pd.read_csv(os.path.join(folder, "df_interest_model.csv"))

# Rename period → quarter where needed
for df in [df_permits_model, df_population_model, df_hpi_model, df_foreign_pop_model]:
    if "period" in df.columns:
        df.rename(columns={"period": "quarter"}, inplace=True)

# Define a filter function to keep only 2015Q1 and onward
def filter_2015_onward(df):
    return df[df["quarter"] >= "2015Q1"]
```



```
# Apply filter
df_cost_model = filter_2015_onward(df_cost_model)
df_permits_model = filter_2015_onward(df_permits_model)
df_population_model = filter_2015_onward(df_population_model)
df_hpi_model = filter_2015_onward(df_hpi_model)
df_foreign_pop_model = filter_2015_onward(df_foreign_pop_model)
df_inflation_model = filter_2015_onward(df_inflation_model)
df_interest_model = filter_2015_onward(df_interest_model)

# Merge on region7 and quarter
df_model = df_permits_model.merge(df_cost_model, on=["region7", "quarter"], how="left")
df_model = df_model.merge(df_population_model, on=["region7", "quarter"], how="left")
df_model = df_model.merge(df_hpi_model, on=["region7", "quarter"], how="left")
df_model = df_model.merge(df_foreign_pop_model, on=["region7", "quarter"], how="left")
df_model = df_model.merge(df_inflation_model, on=["region7", "quarter"], how="left")
df_model = df_model.merge(df_interest_model, on=["region7", "quarter"], how="left")

# Save final model dataframe
output_path = os.path.join(folder, "df_model.csv")
df_model.to_csv(output_path, index=False)

print("✅ df_model created and saved to:", output_path)
```

```
✅ df_model created and saved to: C:\Users\tansu\iCloudDrive\ISEG\Master Final Work\Jul
-25\02 data\processed\df_model.csv
```

- C – Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          yoy_log_hpi_m2   R-squared:                       0.329
Model:                             OLS   Adj. R-squared:                  0.302
Method:                  Least Squares   F-statistic:                     1.612
Date:                 Sat, 19 Jul 2025   Prob (F-statistic):              0.286
Time:                         23:59:42   Log-Likelihood:                 444.50
No. Observations:                  257   AIC:                            -867.0
Df Residuals:                      246   BIC:                            -828.0
Df Model:                           10
Covariance Type:               cluster
===========================================================================================
======================
                                                   coef    std err          z      P>|z|
[0.025      0.975]
-------------------------------------------------------------------------------------------
----------------------
Intercept                                        0.0175      0.016      1.073      0.283
-0.014       0.049
C(region7)[T.Algarve]                            0.0516      0.001     63.030      0.000
0.050       0.053
C(region7)[T.Centro]                             0.0068      0.000     31.071      0.000
0.006       0.007
C(region7)[T.Norte]                              0.0249      0.001     30.092      0.000
0.023       0.027
C(region7)[T.Região Autónoma da Madeira]         0.0588      0.002     37.934      0.000
0.056       0.062
C(region7)[T.Região Autónoma dos Açores]         0.0147      0.001     11.084      0.000
0.012       0.017
C(region7)[T.Área Metropolitana de Lisboa]       0.0500      0.002     27.162      0.000
0.046       0.054
yoy_log_total_bldg                               0.0410      0.020      2.005      0.045
0.001       0.081
yoy_log_material_cost_index_norm                 0.4455      0.190      2.340      0.019
0.072       0.819
inflation                                       -0.0010      0.001     -1.042      0.297
-0.003       0.001
interest_rate                                    0.0090      0.007      1.373      0.170
-0.004       0.022
==============================================================================
Omnibus:                        0.109   Durbin-Watson:                   0.223
Prob(Omnibus):                  0.947   Jarque-Bera (JB):                0.045
Skew:                          -0.032   Prob(JB):                        0.978
Kurtosis:                       3.011   Cond. No.                         155.
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
yoy_log_total_bldg                 :   0.0410  **
yoy_log_material_cost_index_norm   :   0.4455  **
inflation                          :  -0.0010
interest_rate                      :   0.0090
C:\Users\tansu\anaconda3\lib\site-packages\statsmodels\base\model.py:1896: ValueWarning:
covariance of constraints does not have full rank. The number of constraints is 10, but
rank is 4
  warnings.warn('covariance of constraints does not have full '
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            yoy_log_hpi_m2   R-squared:                       0.396
Model:                               OLS   Adj. R-squared:                  0.366
Method:                    Least Squares   F-statistic:                     961.1
Date:                   Sun, 20 Jul 2025   Prob (F-statistic):           1.12e-08
Time:                           00:04:53   Log-Likelihood:                -725.61
No. Observations:                    257   AIC:                             147
Df Residuals:                        244   BIC:                             152
Df Model:                             12
Covariance Type:                 cluster
==============================================================================
```



```
---------------------
                                                  coef    std err          z      P>|z|
[0.025      0.975]
-----------------------------------------------------------------------------------------
---------------------
Intercept                                       6.9364      1.435      4.834      0.000
4.124       9.749
C(region7)[T.Algarve]                          -0.6182      1.494     -0.414      0.679
-3.547       2.310
C(region7)[T.Centro]                           -1.4293      0.687     -2.080      0.038
-2.776      -0.083
C(region7)[T.Norte]                             0.6796      0.966      0.704      0.482
-1.214       2.573
C(region7)[T.Região Autónoma da Madeira]        3.7987      0.635      5.987      0.000
2.555       5.042
C(region7)[T.Região Autónoma dos Açores]        0.0029      0.430      0.007      0.995
-0.840       0.846
C(region7)[T.Área Metropolitana de Lisboa]      1.2147      1.140      1.066      0.287
-1.019       3.449
yoy_log_total_bldg                              0.0076      0.014      0.545      0.585
-0.020       0.035
yoy_log_material_cost_index_norm                0.1626      0.160      1.017      0.309
-0.151       0.476
yoy_log_population                              4.9286      1.300      3.792      0.000
2.381       7.476
yoy_log_foreign_population                     -0.1177      0.103     -1.143      0.253
-0.319       0.084
inflation                                      -0.0022      0.188     -0.012      0.991
-0.371       0.367
interest_rate                                  -0.3955      0.475     -0.833      0.405
-1.327       0.536
==============================================================================
Omnibus:                           0.461   Durbin-Watson:                   0.202
Prob(Omnibus):                     0.794   Jarque-Bera (JB):                0.588
Skew:                             -0.081   Prob(JB):                        0.745
Kurtosis:                          2.830   Cond. No.                         176.
==============================================================================
```