Routledge
Taylor & Francis Group

Regional Studies, Regional Science

ISSN: 2168-1376 (Online) Journal homepage: www.tandfonline.com/journals/rsrs20

Spectral clustering of Berlin's housing spatial network to capture locational effects on pricing

Gianluca Cerruti, Sara Geremia & Alessio Sardo

To cite this article: Gianluca Cerruti, Sara Geremia & Alessio Sardo (2026) Spectral clustering of Berlin's housing spatial network to capture locational effects on pricing, Regional Studies, Regional Science, 13:1, 2618872, DOI: 10.1080/21681376.2026.2618872

To link to this article: https://doi.org/10.1080/21681376.2026.2618872

REGIONAL STUDIES, REGIONAL SCIENCE
2026, VOL. 13, NO. 1, 2618872
https://doi.org/10.1080/21681376.2026.2618872

OPEN ACCESS Check for updates

Spectral clustering of Berlin's housing spatial network to capture locational effects on pricing

Gianluca Cerruti [a*], Sara Geremia [b] and Alessio Sardo [c*]

[a]Department of Economics, University of Genova, Genova, Italy; [b]Department of Social and Political Sciences, University of Trieste, Trieste, Italy; [c]Department of Law, University of Genova, Genova, Italy

ABSTRACT
This study develops a hedonic pricing model to improve the estimation of locational effects by leveraging clustering algorithms. Specifically, we examine and compare the predictive accuracy of four models: a baseline specification without locational effects, a standard model with traditional administrative boundaries, one using constrained *K* means clustering, and another employing a constrained spectral clustering. Focusing on Berlin's housing market (2021–2022), our results show that spectral clustering improves predictive accuracy across standard metrics (root mean square error (RMSE), R squared (R^2), mean absolute error (MAE)). To the best of our knowledge, this is the first study to apply spectral clustering to estimate locational effects within a hedonic pricing framework. Beyond the methodological contribution, we also address Berlin's regulatory and institutional housing context and highlight how improved modelling of locational effects can enhance fairness in property taxation by providing more accurate assessments of locational value.

ARTICLE HISTORY
Received 17 July 2025
Accepted 12 January 2026

KEYWORDS
Spatial clustering; constrained clustering; spectral clustering; hedonic pricing model; housing market; locational effects; Berlin

JEL
C13; C63; 018; R31

1. Introduction

Understanding the determinants of housing prices is essential for both urban planning and policy development. As highlighted by Ryu et al. (2024), accurately measuring the value of a location is a key component of mass appraisals of residential properties, house price index estimation, property taxation and the evaluation of construction projects. To make informed decisions, investors, but also local governments, rely on accurate models reflecting locational value, thus improving the fairness of property taxation.

Over the years, economists, urban planners and researchers from various fields have developed a wide array of quantitative models to estimate the market value of housing. The theoretical foundation of these efforts was laid by Rosen (1974), whose seminal contribution formalised how product differentiation in competitive markets leads to equilibrium prices that reflect the value of individual product attributes. Building on this framework, a vast literature has applied hedonic pricing models (HPMs) to study housing markets, environmental amenities and various aspects of product quality.

A major strength of hedonic models is their ability to decompose a property's price into the contribution of its individual characteristics, offering both flexibility and interpretability. However, this simplicity often comes at the cost of predictive accuracy. Recognising these limitations, scholars have proposed various modifications to improve hedonic models, particularly with the aid of pre-processing techniques, spatial analysis with geographic information systems (GIS) and advanced machine learning algorithms. Wei et al. (2022) provided an extensive review of over 100 studies in this domain. Despite the key developments highlighted in the review, capturing locational effects accurately remains a challenging but essential task for improving housing price models, with limited contributions in the existing literature addressing this issue.

A promising approach in this regard was proposed by Ryu et al. (2024), who introduced a two-stage framework that incorporates fixed location effects via *K*-means clustering. It is important to note that

CONTACT Alessio Sardo alessio.sardo@unige.it
*Gianluca Cerruti and Alessio Sardo are now also at Cities – London School of Economics and Political Science (LSE), London, UK
Supplemental data for this article can be accessed online at https://doi.org/10.1080/21681376.2026.2618872

alternative clustering techniques may provide superior performance in capturing spatial variation. In this paper, we extend this approach by leveraging a rich dataset from ImmoScout24 – Germany's largest platform for residential and commercial property listings – focusing on the Berlin housing market in 2021 and 2022. Alongside detailed property characteristics, we incorporate geospatial information using two advanced clustering techniques to improve the modelling of locational effects: K-means and spectral clustering. Both methods are applied in their constrained forms to account for the heterogeneity in housing density across neighbourhoods. While both methods provide a more granular and flexible spatial representation compared to administrative boundaries, spectral clustering proves particularly effective in capturing the geographic relationships between neighbouring cells, further enhancing the spatial component of the hedonic model. In this approach, sparse areas are grouped together into larger clusters, while dense areas are allowed to form smaller clusters, including isolated individual cells.

It is also worth noting that the other models considered – although less accurate than the spectral clustering specification – still outperformed the baseline model without locational indicators. To the best of our knowledge, we are the first to apply spectral clustering to estimate locational effects within a hedonic pricing framework.

Improving the estimation of locational value is not only crucial for enhancing the accuracy of property valuation models but also has important implications for public policy. More accurate assessments of locational effects allow property taxation systems to better reflect the actual market value of dwellings, thereby improving taxation fairness. In turn, this reduces the risk that wealthier households end up benefiting from undervaluation, which would otherwise contribute to increasing horizontal inequity and exacerbating income and wealth inequalities within urban areas (see Sheffrin (2008) and Gilderbloom et al. (2012), among others).

The case of Berlin represents an especially relevant application, given the critical importance of location in shaping property values in metropolitan contexts. However, estimating these effects remains challenging due to Berlin's spatial heterogeneity of urban environments. To provide context, we also complement our empirical analysis with an overview of Berlin's institutional and regulatory framework, which is characterised by a complex mix of stratified regulation, harmonisation efforts and strong public sector control over housing markets.

The remainder of the paper is structured as follows. Section 2 provides a brief review of the relevant literature. Section 3 introduces Berlin's institutional and regulatory context. Section 4 presents the data and methodological framework, including the clustering techniques and hedonic models (Sections 4.3 and 4.4). Section 5 reports the empirical findings, comparing the performance of alternative approaches to modelling locational effects. Section 6 concludes by discussing the implications of our findings for urban policy and real estate valuation, with particular emphasis on the benefits of incorporating refined spatial indicators into property price models. Finally, Appendices A and B, in the online supplemental data, present, respectively, results for alternative clustering configurations and an additional application for Hamburg.

2. Literature review

2.1. Hedonic pricing models

Since Rosen's (1974) seminal contribution, the HPM has become the standard framework for estimating the implicit value of non-directly observable characteristics of goods, particularly in the housing market. By decomposing the price of a good into the value of its constituent attributes, this approach allows researchers to disentangle the contribution of structural housing features, neighbourhood characteristics and location-specific factors to property values, along with other determinants. The main strengths of hedonic models lie in their simplicity and interpretability, particularly through the estimation of the shadow prices of individual attributes. However, a persistent limitation of this framework is that many of the factors influencing housing prices – especially those related to location – are difficult to observe or quantify, making them challenging to incorporate into conventional regression models.

Over the past decades, a growing body of research has sought to address these limitations from multiple angles. Scholars have experimented with several enhancements to the basic model. It is also worth noting the important role of the spatial econometrics literature in the development of hedonic pricing models.

Seminal contributions, such as Dubin (1992), who examined spatial autocorrelation in hedonic house price models, and Anselin (1988) laid the conceptual foundations for modelling spatial dependence in property values. These frameworks were further developed and extended by LeSage and Pace (2009a, 2009b, 2014), providing robust methodological tools and applications across a wide range of economic and real estate contexts, as well as by numerous other researchers. More recent methodological advances include geostatistical and semiparametric spatial models for real estate pricing (e.g., Muto et al. (2023) and Schirripa Spagnolo et al. (2024)), which further enhance the modelling of spatial heterogeneity and autocorrelation in property markets, and represent a methodological backdrop for the subsequent extensions put forth by machine learning developments and the analysis of locational effects presented below.

2.2. Machine learning developments

A comprehensive overview of this strand of research is provided by Wei et al. (2022), who surveyed 124 studies on the application of big data technologies to improve real estate valuation using HPM. They identify over a dozen categories of big data employed in this context, stemming from three main sources: the internet, remote sensing and the Internet of Things. Among data collection techniques, web scraping emerges as particularly prominent. The integration of advanced pre-processing methods, spatial analysis through GIS and machine learning techniques has demonstrably improved estimation accuracy. Yet, the literature increasingly highlights the challenge posed by heterogeneous data quality, an issue that remains only partially addressed. Wei et al. (2022) emphasised the need for integrated, multidisciplinary approaches capable of combining different techniques to mitigate the shortcomings of individual models.

The integration of machine learning with hedonic modelling has been a particularly active area of research. Zaki et al. (2022), for example, investigated the potential of machine learning algorithms to enhance predictive performance, proposing a hybrid approach that integrates XGBoost with specific methods for outlier management. Their results indicate a marked improvement in predictive accuracy relative to standard approaches.

Similarly, Yoo et al. (2012) applied machine learning techniques to both variable selection and price prediction in housing markets. Using rule-based regression methods – specifically Cubist and Random Forest – they compare their performance against traditional linear regression (ordinary least squares (OLS)) on data from the Onondaga County housing market (New York State). Their findings show that Random Forest delivers the most accurate predictions, followed by Cubist, with OLS performing the worst. These results highlight the advantages of machine learning in dealing with complex, nonlinear relationships in real estate pricing.

More recent contributions have explored the potential of machine learning to utilise information sources that conventional hedonic models struggle to incorporate. Among others, Shen and Ross (2021) and Potrawa and Tetereva (2022) applied machine learning techniques to extract value from textual descriptions and images of properties, types of information typically excluded from traditional models but highly relevant.

Overall, the literature reviewed in this section highlights a trade-off between interpretability and predictive accuracy in hedonic pricing models. Although recent regression techniques for HPM have generally increased predictive accuracy, machine-learning approaches such as Random Forest, gradient boosting and neural networks tend to outperform traditional linear hedonic regressions in terms of predictive performance, this typically comes at the expense of parameter interpretability – as shown, among others, by Bertsimas et al. (2019). Specifically, the main results reported by Yoo et al. (2012, see Table 9 and Figure 5) are particularly illustrative. Machine-learning techniques, including both Cubist and Random Forest, produced lower root mean square error (RMSE), relative RMSE (rRMSE) and Akaike information criteria (AIC) values than the traditional OLS regression, indicating superior model fit. Ceteris paribus (using the same 100 m buffer and the same set of covariates), the Random Forest model achieved an R^2 of 0.97, the Cubist model 0.85 and the OLS regression 0.76, meaning that 76% of the variation in house transaction prices was explained by the linear model. Similar patterns were observed when using a 1 km buffer instead of 100 m. Results along these lines have also been reported by Čeh et al. (2018), Manasa et al. (2020), Potrawa and Tetereva (2022), Zaki et al. (2022) and Soegianto et al. (2024), among others.

Regarding our model, identifying an optimal balance between accuracy and interpretability lies beyond the scope of this paper. Our objective is not to identify the best-performing predictive model but to develop a new predictor of location that can be incorporated into a wide range of econometric or machine-learning frameworks. We adopt OLS as our empirical benchmark because it is the most widely used specification in the hedonic pricing literature and, thanks to its straightforward interpretability, provides a clear baseline against which the contribution of our proposed approach can be highlighted.

2.3. Capturing locational effects

While many contributions have significantly improved the predictive power of hedonic models, one area remains relatively underexplored: the modelling of locational effects. As anticipated in the introduction, traditional hedonic models typically capture location effects through linear terms or coarse spatial dummies, which often fail to account for the complex, nonlinear spatial dynamics characterising urban real estate markets.

The literature has provided extensive evidence of how locational factors shape housing prices, particularly through their interaction with local labour markets. Early contributions, such as Johnes and Hyclak (1999) linked housing prices with labour market conditions like wages, unemployment rates and labour force participation. Wheaton and Lewis (2002) and Osland and Thorsen (2008) showed how urban agglomeration and labour market accessibility shape wage variation. Moretti (2011, 2012, 2013) further developed this perspective, analysing how cities offering better employment opportunities tend to attract skilled labour, thereby increasing housing demand and, consequently, housing prices. This mechanism, in turn, amplifies spatial inequalities across urban areas.

Educational amenities, especially school quality, have been widely studied as influential locational effects in HPM (Feng & Lu, 2013; Haurin & Brasington, 1996; Kane et al., 2006). However, it is important to emphasise that the correlation between school quality and housing prices does not necessarily imply causality. Higher-income families are more likely to live in expensive neighbourhoods and, at the same time, to invest more in their children's education, making it difficult to disentangle the causal relationship between school quality and house prices.

Recent literature has also begun to explore the interplay between school choice mechanisms, residential decisions and housing market dynamics (see, among others, Reback (2005), Bayer et al. (2007) and Avery and Pathak (2021)). Beyond education, cultural amenities have increasingly been recognised as important drivers of housing prices by contributing to neighbourhood attractiveness (Borgoni et al., 2018; Moro et al., 2013; Sheppard, 2010). Environmental factors also represent an important category of locational effects frequently analysed through HPM, particularly in relation to environmental externalities. Studies (Brasington & Hite, 2005; Miłuch & Kopczewska, 2024; Nicholls, 2019) have explored how pollution and environmental disamenities impact housing prices, though data limitations often undermine precise estimations. Moreover, even advanced machine learning approaches often struggle to effectively disentangle locational effects from structural housing characteristics, largely due to their typically lower interpretability (see Ryu et al. (2024)). As a result, accurately capturing locational effects remains a challenging yet crucial task for enhancing the performance of housing price models.

Indeed, as previously discussed, even standard hedonic models often face inherent limitations in fully incorporating the wide range of relevant locational covariates, primarily due to constraints in data availability. This highlights the persistent need for methodological innovation to improve the treatment of spatial heterogeneity in property valuation models. Several studies have sought to address this issue by experimenting with alternative definitions of spatial units and neighbourhood structures. Kryvobokov (2013), for example, proposed redefining neighbourhood boundaries using Thiessen polygons constructed around individual observations, combined with a fuzzy clustering approach to be embedded within HPM. His findings showed mixed results – improvements were more evident when using simple OLS estimation – yet the study was one of the first to emphasise the advantages of moving beyond conventional administrative boundaries.

Expanding on this line of inquiry, Kwon et al. (2017) introduced an innovative methodology for delineating urban areas in Seoul by applying *K*-means clustering in conjunction with hedonic models. Their results demonstrated clear improvements in price prediction accuracy when compared to models based on traditional administrative divisions. Similarly, Calka (2019) advanced the field by combining clustering techniques (specifically, *K*-means clustering) with geostatistical methods such as kriging. This hybrid approach provided a more realistic and fine-grained representation of spatial price variation, particularly valuable in urban environments characterised by strong heterogeneity in housing markets.

Among the most promising contributions in recent years is the two-stage framework developed by Ryu et al. (2024). Their approach first partitions the housing market into clusters based on geographic proximity via *K*-means clustering, using these clusters as fixed locational effects in a subsequent hedonic regression. The benefits of this method are multiple: it requires only basic geographic coordinates (latitude and longitude), accommodates both nonlinearities and unobserved locational factors without imposing a restrictive functional form, and consistently yields better predictive accuracy compared to traditional hedonic specifications. Crucially, this approach offers a meaningful conceptual separation between structural housing characteristics and pure locational effects.

Beyond these approaches, spatially constrained clustering offers additional tools for improving the modelling of locational effects. The Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm partitions space by constructing a minimum spanning tree based on spatial contiguity and attribute similarity, and then pruning edges to obtain geographically contiguous and internally homogeneous clusters (Assunção et al., 2006). In other words, the spanning tree encodes the adjacency relationships between spatial units, and this network based spatial constraint represents a key advantage over techniques such as *K*-means: it produces clusters that more closely reflect real neighbourhood structures and are therefore better suited to capturing localised price variation.

A recent extension, SKATER-reg (Anselin & Amaral, 2024), incorporates regression-based criteria into the clustering process, producing spatial regimes that minimise within-cluster hedonic prediction error. This makes SKATER-reg especially useful for HPM applications in which spatial heterogeneity is unlikely to align with administrative boundaries or standard functional forms.

However, a crucial limitation for our context is that both SKATER and SKATER-reg rely on detailed spatial adjacency information to construct the underlying network. In parallel to these algorithmic approaches, a substantial literature has developed around model-based clustering methods for spatially dependent data, often relying on probabilistic frameworks such as Markov random fields or conditional autoregressive structures (Alfò et al., 2008, 2009). While model-based spatial clustering offers important theoretical advantages, these methods typically require detailed spatial adjacency matrices, impose distributional assumptions, and can be computationally demanding, especially for large datasets. In many real estate applications – including the present study – geographic data are only available in aggregated form, preventing the direct use of these methods. This necessitates the development of an alternative clustering strategy that explicitly incorporates the available aggregation structure when constructing the spatial network, a challenge that is common across empirical housing market research.

Building on these observations, the objective of the present study – which will be detailed in the following sections – is to explore alternative clustering techniques that may offer superior performance in capturing spatial variation in housing prices, thereby addressing some of the unresolved challenges in the empirical modelling of real estate markets.

3. Institutional framework

Berlin's housing market trajectory since reunification presents a paradigmatic case of institutional transformation, fiscal retrenchment and the reconfiguration of housing as an investment asset. The post-1990 transition integrated East Berlin into the Federal Republic's legal framework, triggering a wave of restitution and privatisation, including the sell-off of over 220,000 public housing units, representing more than half of the municipal stock – a process catalysed by the city-state's fiscal crisis and €60 billion public debt (Holm, 2013; Uffer, 2011; Kitzmann, 2017). These asset transfers led to significant ownership concentration, with private landlords and companies managing over 60% of the rental stock and institutional investors holding 150,000 units (Holm, 2013).

It is worth noting that trends in the rental housing market may affect house prices by shaping both investor incentives and household decisions. Rising rents increase the profitability of property investments, stimulating demand for housing purchases and pushing prices upward. Conversely, declining rents reduce the relative attractiveness of ownership, potentially exerting downward pressure on sale prices. The interplay between rental yields and property values thus represents a key transmission channel between the two markets. Between 1990 and 2005, average rents remained stable, owing to legacy subsidies, post reunification high construction volumes, and rent control mechanisms. However, the termination of construction subsidies and the decline in new housing starts, coupled with rising population and household formation, resulted in price tensions. From 2003 onwards, existing contract rents increased by 23%, with new contracts rising more than 100% in certain districts by 2010 (Holm, 2013). The stock of social housing fell from 370,000 units in 1993 to under 150,000 in 2012 (Holm, 2013).

The entry of private equity capital, facilitated by institutional liberalisation, accelerated these dynamics (Heeg, 2004). Firms such as *Terra Firma* and *Deutsche Wohnen* acquired large portfolios and applied strategies based on rent gap exploitation and cost externalisation, turning rental housing into internationally tradable financial assets (Fields and Uffer, 2016). These strategies targeted neighbourhoods such as Friedrichshain, Kreuzberg and Neukölln, which became sites of successive waves of gentrification and price appreciation. Short-term rental platforms amplified these dynamics. Duso et al. (2020) showed, using ZIP-code level panel regressions, that a 10% increase in Airbnb density correlates with a 1.3–2.4% increase in median asking rents, controlling for neighbourhood fixed effects and time trends. The policy response – most notably the *Gesetz über das Verbot der Zweckentfremdung von Wohnraum (ZwVbG)* of 29 November 2013 – has had a limited but evolving impact. The 2023 decision of the Higher Administrative Court of Berlin-Brandenburg (OVG 5 B 5/22) clarified that long-term leases in 'serviced apartments' do not exempt a unit from housing-use requirements, reinforcing the regulatory tools available to limit speculative misuses of the rental stock.

Simultaneously, the *BerlinStrategie – Urban Development Concept 2030* formalises a spatially selective development strategy aimed at densification and infrastructure-led growth in transformation zones such as City West, Adlershof and the *Urban Tech Republic in Tegel* (Senatsverwaltung für Stadtentwicklung und Umwelt, 2014). These zones are prioritised for capital investment and are projected to attract higher-income, mobile professionals – thus reinforcing housing price polarisation across the metropolitan area.

From the perspective of public building, planning and taxation law, spatial heterogeneity in Berlin is not merely a market outcome but is structurally embedded in a multi-level regulatory architecture. Public construction law is divided between federal planning law (*Bauplanungsrecht*, primarily the *Baugesetzbuch* (BauGB) and the *Baunutzungsverordnung* (BauNVO)) and state building regulation (*Bauordnungsrecht*), codified in the *Bauordnung für Berlin* (BauOBln) and the *Ausführungsgesetz zum Baugesetzbuch* (AGBauGB) (Siegel and Waldhoff, 2023). At the city-wide scale, the *Flächennutzungsplan* and the residual 1958/1960 *Baunutzungsplan* define broad land-use and density patterns, while an expanding set of binding *Bebauungspläne* specifies detailed development rights for selected areas and progressively replaces the legacy *Baunutzungsplan*. At the district level, the twelve Districts (*Bezirke*) prepare and implement *Bebauungspläne* and deploy federal planning-law instruments such as *Sanierungsgebiete* and *soziale Erhaltungsgebiete* (*Milieuschutzgebiete* under § 172 BauGB), each applying to precisely delimited territories and imposing specific constraints on conversion, modernisation and land transactions. Spatial differentiation is further reinforced on the fiscal side by Germany's reformed property tax (*Grundsteuer*), which Berlin applies under the federal model: tax liabilities are calculated as the product of a property-specific tax base (*Grundsteuerwert*), a federal coefficient (*Steuermesszahl*) and a municipal rate (*Hebesatz*), with the *Grundsteuerwert* itself depending on zonal standard land values (*Bodenrichtwerte*) determined for small *Bodenrichtwertzonen* by the *Gutachterausschuss* pursuant to §§ 192 ff. BauGB and published in BORIS Berlin. As a result, formally similar dwellings in different *Kieze* face markedly different regulatory constraints and effective tax bases purely because of their location. This layered and spatially selective framework generates precisely the within-city variability in supply elasticities, rent and price dynamics, and capitalisation of amenities that our empirical strategy seeks to capture by modelling locational effects through fine-grained clusters rather than relying solely on conventional administrative boundaries.

4. Data and methodological framework

4.1. Data

The analysis is restricted to the years 2021 and 2022, which represent the most recent period available in the dataset and allow us to analyse contemporary market conditions. In line with Arlia (2024), we exclude 2020 to avoid distortions in market dynamics associated with the COVID-19 pandemic. Focusing on this two-year window minimises disruptions such as listing suspensions and irregular transaction patterns, resulting in a more consistent and homogeneous dataset. We also refrain from extending the temporal window further back, as doing so would reduce temporal uniformity and introduce heterogeneity across market regimes; instead, we prioritise a recent and internally consistent sample.

We use data from ImmoScout24 (Schaff and Thiel, 2023), the largest real estate advertisement platform in Germany. The dataset provides detailed information on property characteristics – such as living area, available amenities and construction year – together with spatial referencing at the 1 km^2 grid level, which represents the finest spatial resolution available. Data covers both rentals and sales, for both houses and flats. Since the markets for flats and houses have structural differences and therefore need to be analysed separately, this study focuses exclusively on flats, which offer greater homogeneity, improving the accuracy of hedonic estimates. Furthermore, we focus on sales prices rather than rents for the Berlin market of 2021–2022. Although a substantial body of research finds sales prices generally more volatile than rents (Gallin, 2008; Hilber and Mense, 2021; Miles, 2023, among others), our data suggest that, in Berlin during 2021–2022, sales prices exhibited relatively smoother dynamics. This likely reflects both the post-pandemic

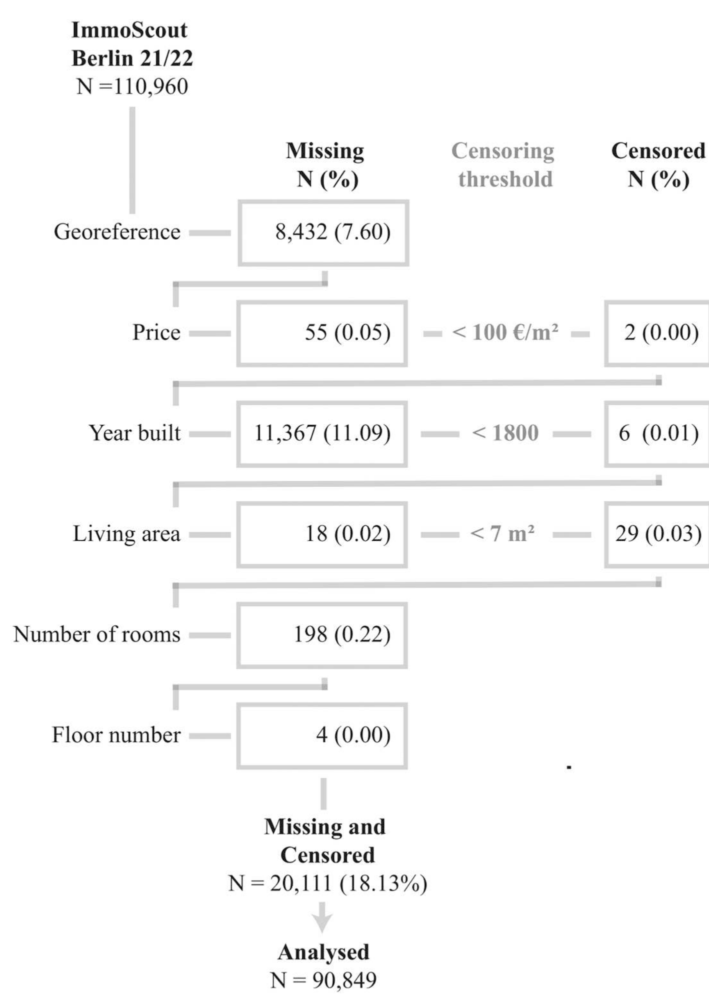

Figure 1. Flow chart.Source: Authors' elaboration based on Immoscout24 data.

acceleration in rent inflation in the euro area (Arioli et al., 2023) and institutional features of the Berlin rental market, which has historically been characterised by rent controls and contract rigidities, and in these years was experiencing the gradual unwinding of policies such as the Mietendeckel, historically moderating short-run adjustments in rents. Consequently, sales prices provide a suitable measure for capturing structural pricing patterns in Berlin over this period.

The sample is restricted to those units for which complete information on the main observed characteristics is available and meet pre-defined criteria. After applying these filters, the final sample consists of 90,849 observations, distributed over 611 grid cells of 1 km^2 (Figure 1).

To refine the dataset, several variables are recategorised. The detailed heating system types are grouped into broader categories by combining similar heating types, reflecting modern or eco-friendly, not eco-friendly and unspecified heating.

The number of rooms variable is recategorised into three groups: small (up to two rooms), medium (more than two but up to three rooms) and large (more than three rooms).

The missing information in the parking variable has been re-coded as 'No' and a new variable 'Parking Coverage' has been created, measuring the percentage of available parking spaces relative to the total number of flats in the same grid.

To incorporate neighbourhood quality indicators into the analysis, we utilised the RWIGEO-GRID dataset. This dataset offers socio-economic data for Germany, structured at the same 1 km^2 resolution, enabling consistent merging with the ImmoScout24 data. It is useful to include socio-spatial variables in the model to account for key determinants of property values and to capture observed spatial heterogeneity across neighbourhoods. Specifically, household density and commercial units reflect land-use intensity and the mix of residential and commercial activity, which influence local amenities and externalities relevant for prices. Likewise, car density proxies mobility patterns and transport dependency, foreign household share serves as an indicator of socio-demographic composition, and children per household signals the presence of families, a feature typically associated with school quality, safety and overall neighbourhood liveability. The unemployment rate conveys local labour-market conditions and economic vitality, directly affecting neighbourhood attractiveness, while housing type reflects the scale and typology of the housing stock, influencing both supply characteristics and buyer preferences. While we acknowledge that buyers do not explicitly take these variables into account, they provide information on underlying neighbourhood attributes that indirectly affect housing prices. As such, they help capture both observable and latent spatial effects and reduce the risk of omitted-variable bias. District level information was integrated using the Berlin districts shapefile (Figure 2). All spatial analyses are conducted at the 1 km^2 grid level, each grid

Table 1. Overview of variables, definitions and data sources used in the analysis.

Variable	Definition	Source property characteristics (ImmoScout24)
Property characteristics (ImmoScout24)		
Price	Purchasing price in EUR	ImmoScout24
Living area	Living space in square metres	ImmoScout24
Construction year	Year the flat was built	ImmoScout24
Elevator	Presence of elevator	ImmoScout24
Number of rooms	Recategorised in: $\leq$2, >2 and $\leq$3, >3 rooms	ImmoScout24
Balcony or garden	Presence of balcony or garden	ImmoScout24
Heating Type	Recategorised in: eco-friendly, not eco-friendly, other	ImmoScout24
Parking	Missing values recoded as 'No'	ImmoScout24
Parking Coverage	Share of flats with parking within the same 1 km^2 grid cell	Constructed variable
Neighbourhood characteristics (RWI-GEO-GRID)		
Households	Number of private households	RWI-GEO-GRID
Commercial Buildings	Number of commercial buildings	RWI-GEO-GRID
Car Density	Cars per household	RWI-GEO-GRID
Foreign Households	Share of households with foreign head of household	RWI-GEO-GRID
Children per Household	Number of children per household	RWI-GEO-GRID
Unemployment Rate	Share of unemployed residents	RWI-GEO-GRID
Housing Type	Recategorised in: 3–5 Family Houses, 6–9 Family Houses, 0–19 Households, >20 Households	RWI-GEO-GRID
Spatial reference and auxiliary data		
Grid Cell ID	Spatial unit at the 1 km^2 resolution (finest available level)	ImmoScout24 & RWI-GEO-GRID
District ID	Berlin administrative district	Berlin Shapefile

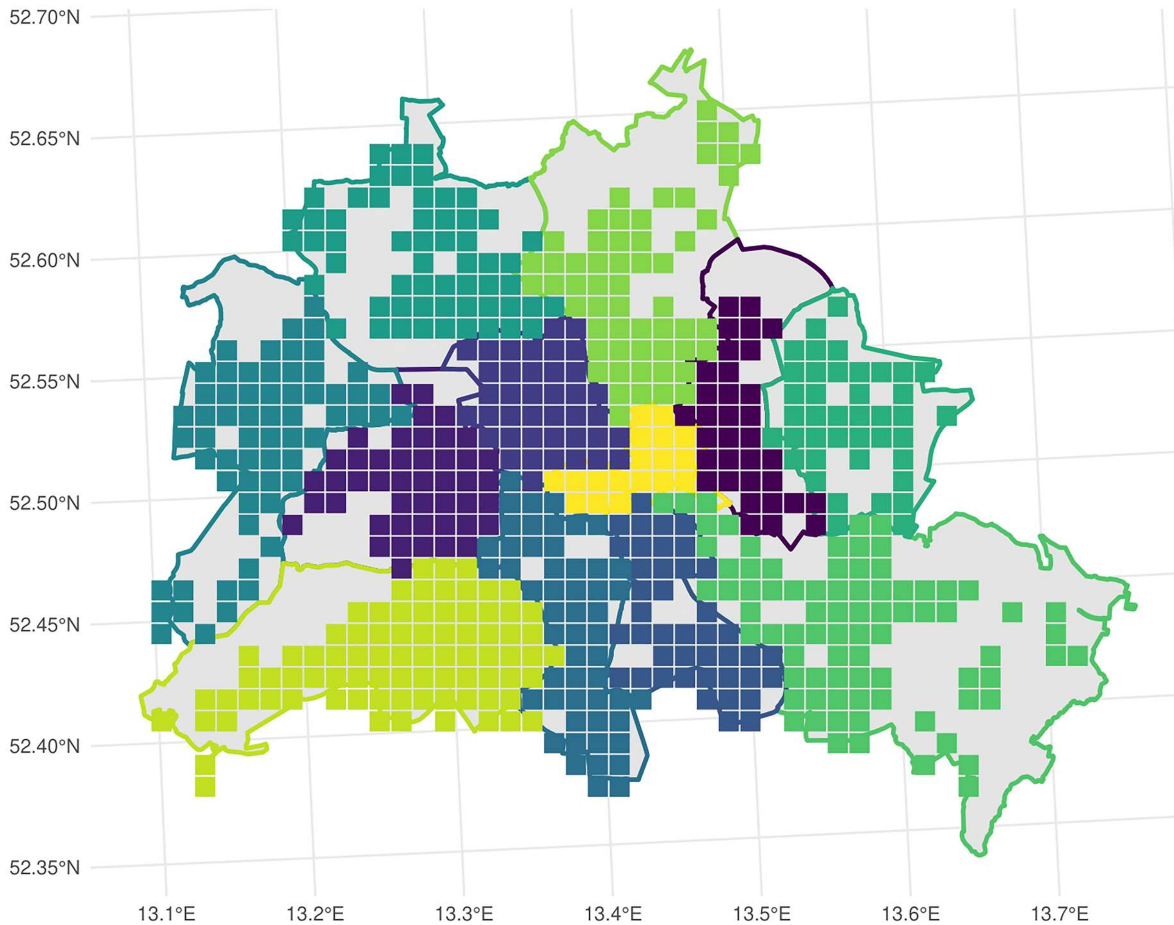

Figure 2. Berlin districts map and 1 km2 grid level map. Source: Authors' elaboration based on RWI-GEO-GRID data and Berlin shapefiles.

corresponding to a square on the map. To support the methodological discussion and provide a unified overview of the dataset, Table 1 presents all variables employed in the analysis, along with their definitions and corresponding data sources.

4.2. Descriptive statistics

Information on various residential properties attributes is summarised in Table 2 in terms of central tendency and distribution for numerical variables, and frequency distribution for categorical variables. The

Table 2. Summary of characteristics for flats listed for sale in Berlin, 2021–2022 (N = 90,849).

		Mean	SD	Median
Price		550,508.32	549,768.64	385,000.00
Construction Year		1953	46	1956
Living area		82.76	46.55	70.00
	Level		N	%
Elevator	No		55,244	60.8
	Yes		35,605	39.2
N Rooms	≤2		44,561	49.0
	>2 and ≤3		27,340	30.1
	>3		18,948	20.9
Balcony or garden	No		24,242	26.7
	Yes		66,607	73.3
Heating type	Modern or eco-friendly		22,957	25.3
	Not eco-friendly		54,881	60.4
	Other		13,005	14.3

average offer posted selling price of properties is €550,483.32, with a substantial variability (standard deviation of €549,764.64) and median price of €385,000.00, suggesting a positive skew in the distribution (i.e., more properties priced below the mean). The median construction year is 1956, indicating that many properties were built in the mid-twentieth century. The mean living area is 82.76 m^2, with considerable variation (standard deviation of 46.55 m^2). The majority of properties (60.8%) do not have an elevator. Approximately half of the properties (49.0%) have two or fewer rooms, which also comprises half rooms (a room with a size between 6 m^2 and 10 m^2). Properties with more than two but less than or equal to three rooms constitute 30.1%, while 20.9% of properties have more than three rooms. A large proportion of properties (73.3%) have either a balcony or a garden, and 25.3% have modern or eco-friendly heating types.

Neighbourhood-level characteristics from the RWI-GEO-GRID dataset are summarised in Table 3. The dataset reveals significant variation in the number of private households per grid cell, with a mean of 3129.90 and a median of 2175.00, reflecting a high standard deviation of 2769.40. Similarly, the number of commercial buildings varies considerably, with an average of 415.05 and a median of 226.00.

The proportion of foreign households averages 16.0%. The children per household ratio remains relatively consistent, averaging 0.25, with little variation across the grid cells. The unemployment rate displays moderate variability, averaging 7.5% with a standard deviation of 3.6%. The parking coverage variable indicates that, on average, 22.0% of flats within a 1 km^2 area have parking available. On average, car density is 0.64 cars per household.

In terms of housing type distribution, the majority of grid cells are characterised by larger residential buildings. Specifically, half of the cells predominantly feature buildings with more than 10 households.

4.3. Clustering

The first step in this work is to subdivide the location plan by taking into account the spatial distribution of flats for sale. The distribution of these properties is typically irregular, influenced by city topography, transport infrastructure and other socio-economic variables. A simplistic subdivision approach using the 1 km^2 grid is unsuitable. Although it provides a uniform structure, it does not take into account variations in the density of housing samples within the city. Cells with few flat listings lack representative characteristics.

Using administrative districts as a framework for capturing location effects also proves inadequate. Districts are not designed for analytical precision, which leads to grids that are not adequately segmented. As a result, they fail to reflect granular variations in neighbourhood characteristics, which are essential for constructing accurate and robust price models. One solution could be to divide the 1 km^2 cells into disjoint sets based on the similarity of geographic location.

Ryu et al. (2024) proposed treating latitude and longitude coordinates as numerical features and using *K*-means clustering, which allows control over the level of spatial approximation as it involves choosing the number of clusters. In our study, we apply a minimum cluster size-constrained version of the *K*-means clustering (Ganganath et al., 2014). Higher-density regions result in a greater number of clusters, while the constraint ensures that clusters with very few observations are avoided. This guarantees consistent estimation quality across areas with varying densities.

Table 3. Summary of characteristics for 1 km^2 Berlin raster cells in 2017 (N = 611).

		Mean	SD	Median
Parking Coverage		0.22	0.20	0.16
Household		3129.90	2769.40	2175.00
Commercial		415.05	557.20	226.00
Car density		0.64	0.19	0.65
Foreign Household		15.98	8.01	13.96
Children per household		0.25	0.05	0.25
Unemployment rate		7.46	3.64	7.72
	Level		N	%
House type	1–2 Family Houses		197	32.2
	3–5 Family Houses		14	2.3
	6–9 Family Houses		96	15.7
	10–19 Households		143	23.4
	>20 Households		161	26.4

An alternative approach constructs a spatial network based on geographical distances (Assunção et al., 2006) and applies constrained spectral clustering (von Luxburg, 2007). We introduce a novel method for spatial network construction designed specifically for contexts in which only aggregated geographical information is available. In our proposed framework, each cell in the 1 km^2 grid is connected by an edge to its adjacent cells – defined as those whose coordinates differ by only one unit – thereby establishing a simple and consistent adjacency structure.

To account for the heterogeneity inherent in aggregated data, we assign edge weights inversely proportional to the total number of observations in the two connected cells. This weighting scheme makes sparsely populated cells more likely to cluster together, while denser cells are able to form more refined, standalone groups. In this way, our approach explicitly leverages the aggregation structure of the data and adjusts for differences in population density, allowing the constructed spatial network to remain informative even when fine-grained geographic detail is unavailable.

The number of clusters is set to 303, implying an average size of approximately 300 flats per cluster, with a minimum cluster size of 20. This lower bound is imposed to ensure statistical stability in the subsequent hedonic regressions and to prevent the formation of sparse clusters, which would undermine the reliability of coefficient estimates. The choice of K is guided by the empirical distribution of listings and reflects a balance between two competing objectives: achieving sufficient within-cluster homogeneity in locational effects while retaining enough observations per cluster to support stable inference.

Although one motivation for clustering is dimension reduction, the degree of reduction is necessarily constrained in settings with highly uneven spatial data density. In our case, specifying a small K or a loose minimum-size constraint would have produced overly large clusters that mask spatial heterogeneity – precisely the variation the method aims to uncover. The selected configuration, therefore, represents a practical compromise between granularity and statistical adequacy.

However, to assess robustness to the clustering specification, we also estimated models using two alternative configurations corresponding to approximately 200 and 400 flats per cluster. The results – available in the Appendix in the online supplemental data – confirm that our main findings remain qualitatively unchanged under our alternative spatial subdivision. However, incorporating locational attributes derived from spectral clustering with $K = 303$ yields improved computational efficiency by operating on a much smaller number of nodes (611 cells versus 90,849 apartments) and quantitatively superior predictive performance.

4.4. The hedonic pricing models

In the second stage, we estimate a set of HPMs that differ in how they capture locational effects. Specifically, we compare three model specifications based on alternative locational indicators, along with a baseline model that omits locational controls. The general model specification takes the following form:

$$\ln P_i = \beta_0 + \sum \beta_s X_{s,i} + \sum \beta_n X_{n,i} + \sum \beta_l X_{l,i} + \epsilon_i$$

where P_i is the price of flat i (the dependent variable), and X represents the independent variables, which are grouped into structural (X_s), neighbourhood (X_n) and locational (X_l) attributes. Structural attributes include the logarithm of the living area, the number of rooms, the availability of an elevator, the presence of a balcony or garden, the type of heating system and the year of construction. Neighbourhood attributes comprise parking coverage, car density, the number of commercial buildings, the number of households and the housing type, as well as the number of foreign households, children per household and the unemployment rate within each grid cell. The locational attribute is a factor with levels representing administrative or cluster assignments. The first hedonic model incorporates administrative districts as location indicators. The second employs clusters generated through constrained K-means with $K = 303$, while the third model, representing our proposal, utilises spectral clustering with $K = 303$ as the locational indicator. Spectral clustering is particularly advantageous due to its capacity to effectively capture spatial dependencies by considering both geographic proximity and variations in housing density, thereby providing a more accurate spatial representation within the model.

Additionally, as anticipated, we extend the comparison to include a baseline model that excludes locational information entirely, providing a reference point to assess the impact of incorporating spatial attributes on predictive performance.

To evaluate the model's prediction accuracy, we used three standard metrics: root mean squared error (RMSE), which measures the average error between observed and predicted values, with lower RMSE indicating a better model fit; R-squared (R^2), which represents the proportion of variance in the dependent variable explained by the independent variables, with higher R^2 values indicating improved explanatory power; and mean absolute error (MAE), which assesses the average magnitude of prediction errors, regardless of their direction.

Beyond assessing predictive performance, we investigate which variables have the greatest impact on predicting flat sale prices. For this purpose, we compute standardised feature importance values within the best-performing hedonic model. The importance of each variable X_q is calculated as the product of its estimated coefficient βq and the standard deviation of X_q; the value is rescaled to a 0–100 range for ease of interpretation.

5. Results

5.1. Clustering

Figure 3 illustrates the spatial distribution of flats for sale across Berlin, with each square representing a cell in the 1 km^2 grid. Darker colours indicate cells with a higher concentration of listings. This grid division results in a highly segmented location plan. The map highlights the highly uneven spatial distribution of housing supply, with dense cells located primarily in central and well-connected areas, while peripheral

Figure 3. Flats for sale density in 1 km2 cells in Berlin in 2021-2022. Source: Authors' elaboration based on Immoscout24, RWI-GEO-GRID data and Berlin shapefiles.

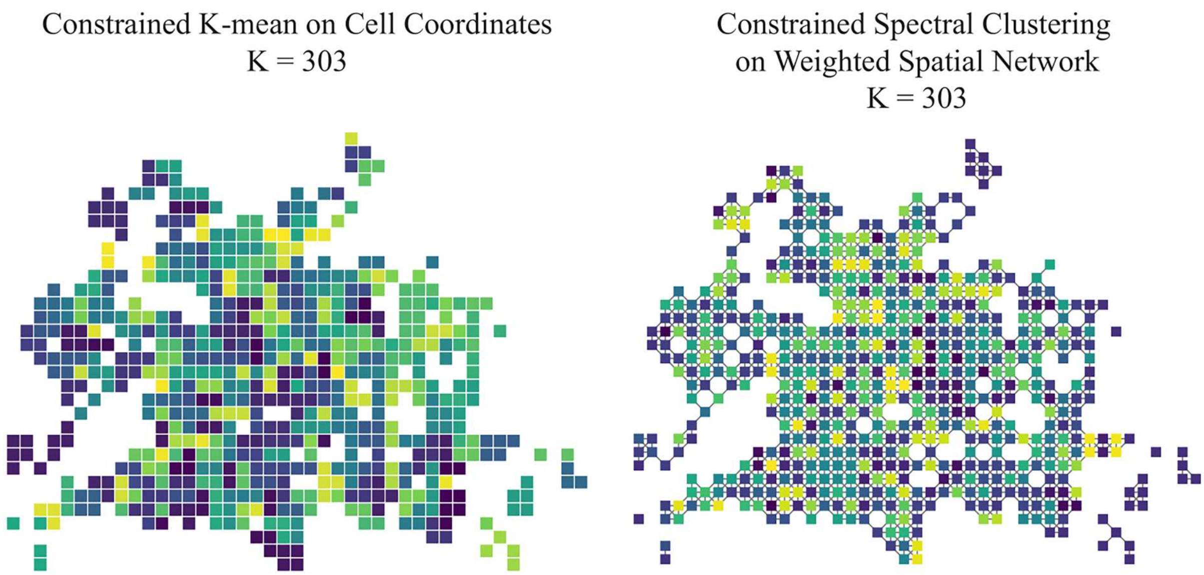

Figure 4. Spatial clustering results. Source: Authors' elaboration based on Immoscout24, RWI-GEO-GRID data and Berlin shapefiles.

regions exhibit significantly lower densities. Employing this partitioning of the city in the model would lead to inconsistent estimates due to insufficient data.

To address this issue, we apply two spatial clustering approaches that aggregate neighbouring cells based on both proximity and listing density, ensuring sufficient observations within each spatial unit for stable estimation. The results of the clustering procedures are presented in Figure 4, where coloured blocks represent different clusters. When incorporating a minimum cluster size constraint, the resulting number of clusters in constrained K-means is $K = 259$ for an initial K of 303, while the spectral clustering results in 249 constrained clusters. The slight reduction in the number of clusters relative to the initial K reflects the merging of low-density cells to meet the minimum observation threshold, ensuring that each cluster contains a sufficient number of listings for reliable analysis. Notably, both clustering methods produce spatial units that better align with the underlying distribution of flats for sale compared to administrative districts or the raw grid. Spectral clustering, in particular, demonstrates an enhanced ability to respect geographic contiguity and local density variations, producing spatial clusters that more accurately capture neighbourhood effects relevant for price modelling.

5.2. The hedonic pricing models

The four estimated models differ exclusively in how they incorporate locational attributes, ensuring that any observed improvement in predictive accuracy is directly attributable to the model's ability to capture the complexity of neighbourhood effects. Any differences underscore the crucial role of accurately modelling spatial influences in the predictive performance of HPMs.

Table 4 presents the predictive performance of the models using three metrics (RMSE, R^2 and MAE), showing the impact of the different locational representations on predictive accuracy. Our proposed

Table 4. Predictive performance of hedonic models using different locational attributes, along with percentage improvements relative to the baseline model without locational controls.

	Performance metrics			Improvement vs. baseline (%)		
Locational attribute	RMSE	R^2	MAE	ΔRMSE	ΔR^2	ΔMAE
None (baseline)	0.31	0.78	0.24	–	–	–
Administrative districts	0.27	0.83	0.21	+12.9%	+6.4%	+12.5%
K–Means clustering	0.29	0.81	0.22	+6.5%	+3.8%	+8.3%
Spectral clustering	0.25	0.85	0.19	+19.4%	+9.0%	+20.8%

specification, which employs constrained spectral clustering as the locational indicator, demonstrates the strongest performance among all alternatives. It achieves the lowest RMSE (0.25), indicating the smallest average prediction error, and the highest R^2 value (0.85), meaning that approximately 85% of the variance in sale prices is explained by the model. The MAE of 0.19 further reflects minimal average deviation between predicted and observed prices.

To facilitate comparison across models, Table 4 also reports percentage changes relative to the baseline model without locational controls. Improvements in RMSE and MAE are computed as percentage reductions in error:

$$\Delta\text{RMSE}_k = \frac{\text{RMSE}_{\text{baseline}} - \text{RMSE}_k}{\text{RMSE}_{\text{baseline}}} \times 100, \Delta\text{MAE}_k = \frac{\text{MAE}_{\text{baseline}} - \text{MAE}_k}{\text{MAE}_{\text{baseline}}} \times 100,$$

while the percentage change in explanatory power is computed as:

$$\Delta R_k^2 = \frac{R_k^2 - R_{baseline}^2}{R_{baseline}^2} \times 100$$

Under these measures, constrained spectral clustering yields the largest gains: approximately a 19% improve in RMSE, a 21% improve in MAE and a 9% improve in R^2 relative to the baseline. Hence, the model without any locational controls performs substantially worse, confirming that omitting spatial information leads to notable predictive loss. Incorporating administrative districts improves performance (RMSE = 0.27, R^2 = 0.83), outperforming the K-means clustering approach, but still less effectively than the spectral clustering specification.

These findings underscore the advantages of constrained spectral clustering in flexibly capturing spatial structure and heterogeneity relevant for property valuation – advantages that neither administrative boundaries nor density-agnostic clustering methods fully exploit. Importantly, the results remain consistent when evaluated on an external test set, confirming the robustness of the proposed approach.

In terms of variable significance, nearly all structural and neighbourhood attributes included in the model are statistically significant at the 1% level ($p < 0.001$), except for the 'Children per Household' variable, which shows weaker significance ($p < 0.005$). Table 5 reports the estimated coefficients for the model employing spectral clustering, along with standard errors and significance levels.[1]

Consistent with theoretical expectations, larger living space, more rooms, the presence of an elevator, a balcony or garden, and modern heating systems are all positively associated with higher sale prices. Neighbourhood characteristics such as higher parking coverage and lower unemployment rates also correspond to higher property values. Interestingly, greater car density is associated with a negative price effect, suggesting that traffic congestion or limited accessibility detracts from perceived property value.

Table 5. Estimated coefficients, standard errors and significance levels for the hedonic model using locational attributes derived from constrained spectral clustering.

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	5.48	6.06×10^{-2}	<0.001
log(Living Area)	1.07	3.01×10^{-3}	<0.001
N Rooms > 2 & ≤ 3	1.24×10^{-2}	2.45×10^{-3}	<0.001
N Rooms > 3	5.25×10^{-2}	3.56×10^{-3}	<0.001
Elevator	1.20×10^{-1}	2.22×10^{-3}	<0.001
Balcony or Garden	4.03×10^{-2}	2.12×10^{-3}	<0.001
Not Eco-friendly Heating	-1.03×10^{-1}	2.25×10^{-3}	<0.001
Not Specified Heating Type	-9.83×10^{-2}	2.99×10^{-3}	<0.001
Construction Year	1.84×10^{-3}	2.48×10^{-5}	<0.001
Parking Coverage	1.29×10^{-1}	8.36×10^{-3}	<0.001
Households	-1.18×10^{-5}	7.23×10^{-7}	<0.001
Commercial Buildings	8.63×10^{-5}	3.24×10^{-6}	<0.001
Car Density	-8.53×10^{-1}	2.38×10^{-2}	<0.001
Foreign Households	-5.59×10^{-3}	2.76×10^{-4}	<0.001
Children per Household	1.47×10^{-1}	5.20×10^{-2}	0.005
Unemployment Rate	-2.47×10^{-2}	8.63×10^{-4}	<0.001
3–5 Family Houses	-1.89×10^{-1}	1.73×10^{-2}	<0.001
6–9 Family Houses	-1.63×10^{-1}	7.52×10^{-3}	<0.001
10–19 Households	-7.89×10^{-2}	6.91×10^{-3}	<0.001
>20 Households	−1.03 × 10-1	8.01×10^{-3}	<0.001

Table 6. Standardised variable-importance values for all predictors in the hedonic model with spectral clustering.

Variable importance	
Living Area	100.00
Construction Year	20.88
Elevator	15.21
Not Eco-friendly Heating	12.87
Car Density	10.09
Not specified Heating Type	9.24
Unemployment Rate	8.04
Commercial Buildings	7.49
6–9 Family Houses	6.08
Foreign Households	5.70
Balcon or Garden	5.33
Households	4.58
Parking Coverage	4.32
N Rooms > 3	4.14
> 20 Households	3.61
10–19 Households	3.20
3–5 Family Houses	3.06
N Rooms > 2 & <= 3	1.41
Children per Household	0.78

Table 6 reports the standardised variable-importance values, which quantify each predictor's relative contribution to the model's predictive performance. The importance scores are rescaled so that the most influential variable – the logarithm of living area – takes a value of 100, with all other variables expressed proportionally. Because the scores are computed from standardised predictors, a variable with a small estimated coefficient may still receive a high importance ranking if it exhibits substantial variability in the data. This explains why construction year, despite having a numerically small estimated coefficient, ranks among the most influential predictors: its wide dispersion across observations leads to a substantial contribution to predictive accuracy.

The ranking indicates that structural attributes, particularly living area, construction year and presence of an elevator, play the dominant role in explaining price variation. Among neighbourhood characteristics, car density, heating system type and unemployment rate emerge as significant contributors to price prediction, underscoring the role of both individual house characteristics and local socio-economic context in shaping property values.

6. Conclusion

This study focuses on the locational effects in HPMs and compares the predictive accuracy of different model specifications. To this end, we use a rich dataset from ImmoScout24 – Germany's largest platform for residential and commercial property listings – focusing on the Berlin housing market in 2021 and 2022. We estimate four HPMs: a traditional specification without locational effects, a standard model including locational controls based on administrative boundaries, one leveraging constrained *K*-means clustering, and another employing constrained spectral clustering.

The key difference across models lies in how they incorporate locational attributes, ensuring that any improvement in predictive accuracy can be directly attributed to the model's ability to capture the complexity of neighbourhood effects.

Unlike the two more conventional specifications, the clustering-based models are specifically designed to account for the non-uniform distribution of listings and the heterogeneity in housing density across neighbourhoods. While both clustering approaches offer a more flexible and granular spatial representation than administrative boundaries, spectral clustering proves particularly effective in capturing the geographic relationships between adjacent spatial units. In addition, it implicitly incorporates information on cell density, further enhancing the model's ability to account for the spatial distribution of flats for sale.

To the best of our knowledge, this is the first study integrating a spectral clustering algorithm within a standard hedonic pricing framework to estimate locational effects, an approach that improves predictive performance and enhances the accuracy of valuation estimates.

We assess the predictive performance of the models using three widely adopted metrics: RMSE, R^2 and MAE. According to the results, the model incorporating spectral clustering consistently outperforms the alternatives across all three metrics. It is worth noting that although the other specifications perform worse than the spectral clustering model, they still outperform the baseline hedonic model without any locational controls. Moreover, the model using standard K-means clustering performs similarly to the conventional specification based on administrative boundaries.

In addition to its superior predictive accuracy, the spectral clustering model (as well as the K-means clustering model) offers another key advantage: it requires fewer data inputs than conventional specifications. As shown in the paper, the clustering approaches rely solely on georeferenced information (latitude and longitude), whereas traditional models require access to a much broader range of contextual data.

Berlin represents a particularly relevant case study given the critical role of location in shaping housing prices in metropolitan contexts. To better contextualise our empirical analysis, we also explore Berlin's institutional and regulatory framework, which is characterised by a complex layering of regulatory regimes, harmonisation efforts and strong public sector involvement in the housing market.

All in all, improving the estimation of locational value is not only crucial for advancing property valuation techniques, but it also carries significant policy implications. More accurate assessments of locational effects allow property taxation systems to better align assessed values with actual market prices, thereby improving fairness in taxation. In turn, this reduces the likelihood that wealthier households benefit from undervaluation, a phenomenon that would otherwise exacerbate horizontal inequities and contribute to wealth and income disparities in urban areas (see Sheffrin (2008) and Gilderbloom et al. (2012), among others).

Additionally, this approach allows for a clearer separation between locational and structural effects within the hedonic model, a feature that holds particular value for urban studies researchers. However, while the two effects remain distinct, clustering-based approaches – both K-means and spectral – do not provide further insights into the specific composition of locational effects.

Naturally, this study has limitations. Although we test the robustness of our approach by considering alternative clustering configurations and by replicating the analysis on Hamburg (see Appendices A and B in the online supplemental data for further details), the results are based on only two cities and a limited time period (2021–2022). Future research should extend the analysis to additional locations with different housing market characteristics, longer time spans and rental data as well as transaction prices. Applying spectral clustering to estimate locational factors in HPMs across diverse metropolitan housing markets would help assess the generalisability of the model and test its improved predictive accuracy across different contexts and modelling strategies.

Note

1. We have also computed the location indicators for the other estimated models – including the standard hedonic model without locational attributes, the model with administrative location controls and the constrained K-means model. These results are available upon request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Funded by the European Union (ERC STG HABITAT – How European Big Cities & Legal Systems Trigger Urban Inequality: An Inquiry into Law and Economics – Grant agreement ID: 101076616). The views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

ORCID

Gianluca Cerruti http://orcid.org/0000-0002-8755-0050
Sara Geremia http://orcid.org/0000-0001-7780-040X
Alessio Sardo http://orcid.org/0000-0003-1015-2641

Data availability statement

The data employed in this study are drawn from the proprietary RWI-GEO-GRID dataset and Immoscout24. Access to these data is subject to contractual and licensing restrictions, and the authors do not hold the rights to disseminate them. As a result, the datasets cannot be made publicly available.

References

Alfó, M., Nieddu, L., & Vicari, D. (2008). A finite mixture model for image segmentation. *Statistics and Computing*, *18*(2), 137–150. https://doi.org/10.1007/s11222-007-9044-9

Alfó, M., Nieddu, L., & Vicari, D. (2009). Finite mixture models for mapping spatially dependent disease counts. Biometrical journal. *Biometrische Zeitschrift*, *51*(1), 84–97. https://doi.org/10.1002/bimj.200810494

Anselin, L. (1988). *Spatial econometrics: Methods and models*, volume 4. Springer Science & Business Media.

Anselin, L., & Amaral, P. (2024). Endogenous spatial regimes. *Journal of Geographical Systems*, *26*(2), 209–234. https://doi.org/10.1007/S10109-023-00411-2

Arioli, R., Bobeica, E., Roma, M., & Soudan, M. (2023). Rent inflation in the euro area. *Economic Bulletin Boxes*, *7*.

Arlia, D. (2024). *Labor market shocks across heterogeneous housing markets*. December 2024, mimeo.

Assunção, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, *20*, 797–811. https://doi.org/10.1080/13658810600665111. https://www.academia.edu/25106099/Efficient_regionalization_techniques_for_socio_economic_geographical_units_using_minimum_spanning_trees

Avery, C., & Pathak, P. A. (2021). The distributional consequences of public school choice. *American Economic Review*, *111*(1), 129–152.

Bayer, P., Ferreira, F., & McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, *115*(4), 588–638.

Bertsimas, D., Delarue, A., Jaillet, P., & Martin, S. (2019). The price of interpretability. *arXiv preprint arXiv:1907.03419*, 1–19.

Borgoni, R., Michelangeli, A., & Pontarollo, N. (2018). The value of culture to urban housing markets. *Regional Studies*, *52*(12), 1672–1683.

Brasington, D. M., & Hite, D. (2005). Demand for environmental quality: A spatial hedonic analysis. *Regional Science and Urban Economics*, *35*(1), 57–82.

Calka, B. (2019). Estimating residential property values on the basis of clustering and geostatistics. *Geosciences*, *9*(3), 143.

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.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, *22*(3), 433–452.

Duso, T., Michelsen, C., Schaefer, M., & Tran, K. D. (2020). Airbnb and rents: Evidence from berlin. Technical report, DIW Discussion Papers.

Feng, H., & Lu, M. (2013). School quality and housing prices: Empirical evidence from a natural experiment in Shanghai, China. *Journal of Housing Economics*, *22*(4), 291–307.

Fields, D., & Uffer, S. (2016). The financialisation of rental housing: A comparative analysis of New York city and Berlin. *Urban Studies*, *53*(7), 1486–1502.

Gallin, J. (2008). The long-run relationship between house prices and rents. *Real Estate Economics*, *36*(4), 635–658.

Ganganath, N., Cheng, C. T., & Tse, C. K. (2014). Data clustering with cluster size constraints using a modified K-means algorithm. In *2014 International conference on cyber-enabled distributed computing and knowledge discovery* (pp. 158–161). Shanghai, China: IEEE. https://doi.org/10.1109/CyberC.2014.36

Gilderbloom, J. I., Hanka, M. J., & Ambrosius, J. D. (2012). Without bias? Government policy that creates fair and equitable property tax assessments. *The American Review of Public Administration*, *42*(5), 591–605.

Haurin, D. R., & Brasington, D. (1996). School quality and real house prices: Inter-and intrametropolitan effects. *Journal of Housing Economics*, *5*(4), 351–368.

Heeg, S. (2004). Mobiler immobilienmarkt? finanzmarkt und immobilienökonomie. *Zeitschrift für Wirtschaftsgeographie*, *48*(1), 124–137.

Hilber, C. A. L., & Mense, A. (2021). Why have house prices risen so much more than rents in superstar cities? CEP Discussion Papers dp1743, Centre for Economic Performance, LSE, Jan 2021. https://ideas.repec.org/p/cep/cepdps/dp1743.html

Holm, A. (2013). Berlin's gentrification mainstream. In M. Bernt, B. Grell, & A. Holm (Eds.), *The Berlin reader. A compendium of urban change and activism* (pp. 171–188). The Berlin Reader.

Johnes, G., & Hyclak, T. (1999). House prices and regional labor markets. *The Annals of Regional Science*, *33*, 33–49.

Kane, T. J., Riegg, S. K., & Staiger, D. O. (2006). School quality, neighborhoods, and housing prices. *American Law and Economics Review*, *8*(2), 183–212.

Kitzmann, R. (2017). Private versus state-owned housing in berlin: Changing provision of low-income households. *Cities*, *61*, 1–8.
Kryvobokov, M. (2013). Hedonic price model: Defining neighbourhoods with Thiessen polygons. *International Journal of Housing Markets and Analysis*, *6*(1), 79–97.
Kwon, S., Kim, S., Tak, O., & Jeong, H. (2017). A study on the clustering method of row and multiplex housing in Seoul using K-means clustering algorithm and hedonic model. *Journal of Intelligence and Information Systems*, *23*(3), 95–118.
LeSage, J., & Pace, R. K. (2009a). *Introduction to spatial econometrics*. Chapman and Hall/CRC.
LeSage, J. P., & Kelley Pace, R. (2009b). Spatial econometric models. In M. M. Fischer & A. Getis (Eds.), *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 355–376). Springer.
LeSage, J. P., & Kelley Pace, R. (2014). Interpreting spatial econometric models. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of regional science* (pp. 1535–1552). Springer.
Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine learning based predicting house prices using regression techniques. In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)* (pp. 624–630). IEEE.
Miles, W. (2023). Is there a stationary home price-rent relationship in us housing markets? *International Journal of Business & Economics (IJBE)*, *9*(2), 113–131. https://doi.org/10.58885/ijbe.v09i2.113.wm. https://ielas.org/ijbe/index.php/ijbe/article/view/142
Miłuch, O., & Kopczewska, K. (2024). Fresh air in the city: The impact of air pollution on the pricing of real estate. *Environmental Science and Pollution Research*, *31*(5), 7604–7627.
Moretti, E. (2011). Local labor markets. In D. Card & O. Ashenfelter (Eds.), *Handbook of labor economics* (Vol. 4, pp. 1237–1313). Elsevier.
Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.
Moretti, E. (2013). Real wage inequality. *American Economic Journal: Applied Economics*, *5*(1), 65–103.
Moro, M., Mayor, K., Lyons, S., & Tol, R. S. (2013). Does the housing market reflect cultural heritage? A case study of greater Dublin. *Environment and Planning A*, *45*(12), 2884–2903.
Muto, S., Sugasawa, S., & Suzuki, M. (2023). Hedonic real estate price estimation with the spatiotemporal geostatistical model. *Journal of Spatial Econometrics*, *4*(1), 10.
Nicholls, S. (2019). Impacts of environmental disturbances on housing prices: A review of the hedonic pricing literature. *Journal of Environmental Management*, *246*, 1–10.
Osland, L., & Thorsen, I. (2008). Effects on housing prices of urban attraction and labormarket accessibility. *Environment and Planning A*, *40*(10), 2490–2509.
Potrawa, T., & Tetereva, A. (2022). How much is the view from the window worth? machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, *144*, 50–65.
Reback, R. (2005). House prices and the provision of local public services: Capitalization under school choice programs. *Journal of Urban Economics*, *57*(2), 275–301.
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34–55.
Ryu, D., Hong, J., & Jo, H. (2024). Capturing locational effects: application of the K-means clustering algorithm. *The Annals of Regional Science*, *73*(1), 265–289. https://doi.org/10.1007/s00168-024-01263-4
Schaff, S., & Thiel, P. (2023). FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED v9)-Advertisements on the Internet Platform lmmobilienScout24. Technical report. www.rwi-essen.de
Schirripa Spagnolo, F., Borgoni, R., Carcagnì, A., Michelangeli, A., & Salvati, N. (2024). A spatial semiparametric m-quantile regression for hedonic price modelling. *AStA Advances in Statistical Analysis*, *108*(1), 159–183.
Senatsverwaltung für Stadtentwicklung und Umwelt. Berlinstrategie – urban development concept 2030. (2014). Technical report, Berlin: Senate department for urban development and the environment.
Sheffrin, S. M. (2008). *Fairness and market value property taxation*. Georgia State Univ., Andrew Young School of Policy Studies.
Shen, L., & Ross, S. (2021). Information value of property description: A machine learning approach. *Journal of Urban Economics*, *121*, 103299.
Sheppard, S. (2010). Measuring the impact of culture using hedonic analysis. *Center for Creative Community Development*, *28*, 1–28.
Siegel, T., & Waldhoff, C. (2023). Öffentliches recht in berlin: Verfassungs-und organisationsrecht, allgemeines verwaltungsrecht mit verwaltungsprozessrecht, polizei- und ordnungsrecht mit versammlungsrecht, öffentliches baurecht: eine prüfungsorien- tierte darstellung. *(No Title)*.
Soegianto, L. M., Hinandra, A. T., Suri, P. A., & Fajar, M. (2024). Comparison of model performance on housing business using linear regression, random forest regressor, svr, and neural network. *Procedia Computer Science*, *245*, 1139–1145.
Uffer, S. (2011). *The uneven development of Berlin's housing provision* [Doctoral dissertation]. London School of Economics and Political Science.
von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land*, *11*(3), 334.
Wheaton, W. C., & Lewis, M. J. (2002). Urban wages and labor market agglomeration. *Journal of Urban Economics*, *51*(3), 542–562.
Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293–306.
Zaki, J., Nayyar, A., Dalal, S., & Ali, Z. H. (2022). House price prediction using hedonic pricing model and machine learning techniques. *Concurrency and Computation: Practice and Experience*, *34*(27), e7342.