

Automated Housing Price Valuation and Spatial Data

Paulo Batista(✉) and João Lourenço Marques

Universidade de Aveiro – GOVCOPP, Campus universitário de Santiago, Aveiro, Portugal
pauloricardolb@ua.pt

Abstract. The demand for automated, reliable and understandable housing price valuation mechanisms is increasing. Most efforts have been made to improve model accuracy and prediction power through the well-established standard econometric models based on regression techniques. However, the modelling of the spatial attributes of housing through mass appraisal tools has been given less attention. Incorporating spatial modelling approaches through econometrics frameworks opens new opportunities for improving automated valuation tools.

This work presents an exploratory analysis of different approaches to incorporating spatial data into AVM tools, taking advantage of the potential of spatial (big) data, stored on different sources – census data, open street maps and public administration data.

Improvement of the standard housing price models embedded in a Portuguese housing appraisal decision system (held by PrimeYield SA) will be presented. Different strategies to incorporate spatial data from public sources are analysed, taking the Sintra municipality and PrimeYield data on this territory as a case study. The focus is the mitigation of the well-known pitfalls of spatial models, such as spatial heterogeneity and spatial dependence.

The results show the potential added-value of collecting and (pre)processing a different set of territorial variables – socioeconomic, accessibility, and land use – to improve the explanation power, parsimony and understanding of housing price models. Geographic weight regression models can be a balanced compromise to achieve those objectives which will be investigated.

Keywords: AVM · The housing market · Spatial econometric models

1 Introduction

In Portugal, as in most European countries, the free market is the primary gatekeeper for accessing housing. At the time of the 2011 Census, Portugal had 73% of families as owners of their residential dwellings, with 31% of Portuguese families ensuring their residential property through mortgage mechanisms. The characteristics of housing set it apart from more traditional assets concepts in economic theory [1], mainly due to its i) heterogeneity and singularity, ii) immobility and iii) durability.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 366–381, 2021.
https://doi.org/10.1007/978-3-030-86973-1_26

Some distinct elements of housing prices related to the territorial features of its location are well-known [2–4]. However, in some way, these elements are challenging to fully measure in housing markets. Their role as market drivers prevails, codified for the multiple market agents, including mediators such as house value appraisal agents [5].

New financial regulations, new tax policy requirements and an increasing demand from investors for more accurate price valuations have resulted in the adoption of more sophisticated approaches to appraisals in housing valuation support systems [6, 7]. As data on house transactions is growing and the demand for mass appraisals of housing assets is rising, the adoption of machine learning models has been gaining a central role despite questions remaining related to the role of space in these models.

As housing data and information on spatial features are expanding, efforts to adequately process and combine these different datasets need to be investigated. Appraisals agents should pay attention to the notion of space in housing market analysis as a spatially fixed good: the concept remains of paramount debate in housing economics, urban and regional sciences [8–10] and it is recognized that an inaccurate understanding of its role in housing market mechanisms can lead to incomplete identification of the housing price drivers [11]. Despite this, the segmentation in housing markets, specifically the emergence of spatial submarkets and their spatial interlinkages, is a well-known issue. Empirical observations based on exploratory statistical measures, such as the Moran index (Moran I) indicator of (global) spatial dependence [12] and the local indicators of spatial association [13], confirm the extension of this challenge.

In short, the specification of reliable housing market price models is essential to ensure the (partial) automation of valuation processes usually adopted by appraisal agents. More information needs to be provided, particularly related to the role of spatial features in housing prices. This data should also be incorporated into the usual decision support models. This paper intends to contribute to this emerging debate – see, for example, the chapter "AVM Methodological Challenges: Dealing with the Spatial Issue" in the book *Advances in Automated Valuation Modelling* [6]. Moreover, the general objective here is motivated by the efforts of PrimeYield S.A. (PY) – an official real estate appraisal operator in Portugal – to implement a more accurate (semi-)automated valuation model in its operations. Specifically, the intention of this work is to i) explore different ways to enhance housing price models with spatial data with different types and resolutions – such as spatial data specified as spatial points or as spatial polygons; ii) analyse the added-value of enhancing housing price models with indicators which measure territorial features; and iii) examine a different set of standard specification techniques of the standard housing price models to embed spatial data.

The paper is organized as follows. The following Sect. 2 presents the theoretical background on the standard econometric model framework to analyse housing price and its drivers – the hedonic approach – and the significant light which this framework sheds on dealing with spatial pitfalls – particularly, the understanding of spatial dependence and spatial heterogeneity. Proposed solutions in the literature will be highlighted. The Sect. 3 provides the case study data and the methods suggested in this exploratory work to

incorporate data about spatial features, taking advantage of new open data sources. The Sect. 4 presents the significant findings and the validation of the methods proposed, with a brief discussion on the relevance of incorporating the new data for better modelling parsimony. The paper finishes with the major highlights of these work contributions to the debate and a brief comment on the next necessary steps.

2 Theoretical Background

2.1 Housing Valuation, Housing Hedonic Price Models and Spatial Modelling Challenges

The hedonic housing pricing model (HHPM) [14] is a well-known modelling framework and is widely used in decision support systems for housing appraisal processes. Its technical simplicity – as models can be deployed through the standard OLS regression approach – and its anchorage in rational economic theory leads to HHPM being chosen to ensure the reliable and understandable role of the housing price drivers.

Housing is defined as a heterogeneous good for which a complex set of attributes should be selected to describe it, namely, intrinsic (physical) and spatial (neighbourhood) characteristics. The estimations of HHPM can be obtained through the traditional regression model [15], which assumes a reduced-form econometric model, such that:

$$\boldsymbol{P} = \alpha + \beta H + \varepsilon \tag{1}$$

where P is the vector of prices (or prices by square meter) for each one of the n dwellings; and β is a vector of hedonic (or shadow) prices to be estimated, describing the value of each one the k dwelling attributes H. This is a reduced form of the model where the reliability of estimations is linked to the theoretical assumption of a competitive market in equilibrium. Finally, ε is the stochastic model error.

Grounded on the categorization proposed by Stull [16] and the open debate on what features to include to describe spatial (territorial) features (see Galster [3]), the matrix H that quantifies the attributes of a dwelling can be decomposed into a set of four subcategories: F, E, L and S, plus time (T) – in order to fix time effects (such as inflation or other macro-economic time-dependent phenomena); F denotes structural characteristics of the dwelling; E, L, S[1] are environmental and neighbourhood characteristics, the location within the territorial system (or the housing market delimitation considered) and other spatial characteristics (access to utilities and public services, such as transport nodes, working place poles, schools, etc.).

2.2 Challenges Related to the Spatial Features of Residential Dwellings

Well-known social and economic phenomena supported the rise of spatial econometrics [17, 18], which combined the knowledge produced in economics, geography and other spatial sciences. These joint efforts have resulted in new light being shed on the nature of spatial phenomena, the challenges of modelling them and the tools concerned with

[1] In this work L will be used as a reference to all E, L and S types of spatial features.

fixing the models' spatial pitfalls. Two significant spatial challenges to producing reliable models are identified: the rise of heterogeneity and spatial dependence.

Spatial Dependence

Spatial dependence is a well-known phenomenon in a wide range of empirical studies in different fields. It can be observed on different spatial scales of analysis or other spatial (geographic) features – polygons or points as references to territorial attributes embedded in the modelling process.

The complexity behind the concept of territoriality is related to locality, urbanity, socioeconomic characteristics and other features. This has resulted in substantial uncertainties over model specifications. Spatial econometrics has developed several spatial interaction models where theoretical assumptions mainly guide the choices through economic mechanisms [19, 20]; however these models are focused on estimate average (global) spatial interaction effects rather than obtain point estimates (the spatial effects in each dwelling unit).

The spatial dependence usually requires both the spatial unit and its W to be known a priori [20]. Following a business-as-usual approach, W is usually defined as the neighbourhoods' relations through a Euclidian geometrical reference frame, using a specific Euclidian distance threshold or the topological relations between geographical units (polygons). Landry & Chakraborty's work [21] adopted a function of (Euclidian) distance between spatial units, allowing for a more geographically coherent definition of neighbourhoods when the geographic units (polygons) are not regular (which is usual, for example, in administrative unit settings). An interesting lesson from the attempts to define W is the tendency to decrease spatial autocorrelation with increasing distance (see Getis and Aldstadt [22]). This can be understood as an expression of Tobler's law, according to which the closest things are more related than distant things [23].

The above observation can be used to introduce another modelling alternative: geographically weighted regression (GWR) [24, 25]. GWR focuses on obtaining spatially located parameter estimates through a relaxation of the theoretical assumptions usually imposed through W. The primary mechanism of GWR relies on the assumption that contextual, spatial factors may modify the strength and direction of the relationship between a dependent variable and its predictors. Estimations for each geographic data point are then obtained locally using a kernel function centred on that point and adapted so that neighbouring data points (in Euclidian space) are considered as weights based on a distance decay function. In short, the regression framework is adapted to allow spatial variation of the regression coefficients across space; different (kernel) functions can be used. The GWR regression model can be written as:

$$Y = \alpha(u, v) + \sum_{s=1}^{S} X_s \beta_s(u, v) + \varepsilon \tag{2}$$

where Y is the target variable (in this work P – the housing price) at location (u, v), X is the set of explanatory variables (the H characteristics of each dwelling), and $\beta_s(u, v)$ are the parameters for the regression coefficient β, obtained through a weighting scheme w_{ij} applied to take into account (u, v). The weighting scheme is based on the kernel function, such as a Gaussian kernel (as adopted later in this work); specifically, this function

incorporates a distance decay mechanism which allocates more weight to dwellings closer to a regression point than dwellings farther away, as follows:

$$w_{ij} = \exp\left(-\left(d_{ij}/b\right)^2\right) \tag{3}$$

Here, d_{ij} is the (Euclidean) distance[2] between dwelling i and the neighbourhood dwelling j, and b is the bandwidth, the distance where searching for dwellings is considered used in the weight estimation mechanism.

One critical assumption of GWR is that all dwelling transactions occur simultaneously, or, at the very least, that the time of transaction is not a crucial factor to consider. In standard HHPM, time non-stationarity can be addressed simply by including time fixed effects. Recently developments in the GWR framework lead to the adaptation of the kernel function to take into account local effects in both space and time dimensions. Fotheringham et al. [26] propose the GTWR model and demonstrate the increasing accuracy of that approach compared to standard models in time non-stationarity settings.

Spatial Heterogeneity

In another direction, the emergence of spatial dependence can be pinpointed to the heterogeneous nature of spatial phenomena. The distinction between both phenomena is not total. For example, the concept of substitutability (in economics) supports this close connection between both spatial phenomena: the work of Bourassa et al. [27] points that substitutability in housing markets is mainly related to location and neighbourhood attributes. Despite the induced spatial delimitation of such a concept, empirical approaches remain challenging [28].

As described for spatial dependence, a standard approach usually assumes the territorial units' boundaries a priori, and that they are reasonably homogenous. A typical straight path is the use of administrative units or other known spatial partitions. Alternatives may be to previously produce spatial clusters following the knowledge developed in geography, urban studies or regional science; moreover, places can be identified by local communities or can be defined as the zones adopted by housing market agents. In their PhD theses [8, 9], the authors of this work argued that these different approaches call for a conception of space beyond the classic, geometric and dimensional notions of space usually embedded in spatial sciences, HHPM in particular.

The diversity of solutions and the absence of an unequivocal consensus have led to adopting different solutions, usually constrained by data availability. Here two approaches will be followed: a) the use of pre-existing territorial areas (political-administrative boundaries and identifiable "neighbourhoods"); b) the use of a set of measured spatial variables (such as, for example, indicators retrieved from census data or accessibility indexes) as proxies to fix the effects of neighbourhood characteristics.

[2] Although other distance measures can be used.

3 Data and Methods

3.1 Processing and Combining Data from Different Sources

As argued before, HHHP analysis should incorporate spatial information on order to be reliable. Housing market databases usually include some spatial data: the zone/neighbourhood/administrative unit assigned (a spatial polygon that contains the dwelling) or the specific housing address (which will define a particular spatial point through geocoding tools[3]). Each type of spatial feature encodes different information details, which should be decoded. Also, the data type conditions the information precision that it will be possible to retrieve.

This work uses three different data sources: i) data accumulated by PY, which describes physical housing attributes, housing price and its geographic location (geographic coordinates); the data is referenced to PY's appraisal service; ii) territorial data will be collected from three different open data sources, namely, the Portuguese statistical authority, the government spatial planning department (for land use data) and, the SapoMapas geoservices (http://api.mapas.sapo.pt/) to retrieve a collection of points of interest (POIs) and distance matrix between the different spatial points.

Decoding spatial information usually requires processing data produced by different sources. This data typically presents significant modelling challenges, namely: geographically, different polygons may be topologically inconsistent; thematically, other datasets may have diverse attributes or coding typification; and methodologically, as a different type of geometries are used to represent spatial data (points, lines, polygons), it can result in imprecisions or inconsistencies.

In the following two sections, the details of the data available on these different sources and the most important pre-processing steps will be presented.

Dwelling Data

For this study, PY provided a dataset with 625 housing transaction records (based on data for the Sintra municipality) and comprising a period of 7 years (between 2008 and 2015). The available target variable (price) is derived from the PY appraisal auditions – which adopted certified appraisal processes such as the RICS [29] and TEGOVA [30] guidelines. Only a small number of variables are available from the original dataset after data cleaning steps (such as dropping variables with missing values). The dataset provides each dwelling's geographical location (address or x, y coordinates provided on WGS84 coordinates system), which was matched with the external geographic datasets (administrative units, locality, census tracts and the regular geographic grid used to produce the accessibility index). Summary statistics are provided in Table 2.

Territorial Data

Combining the dwellings' precise location with spatial data on the additional data sources will provide an enrichment of the dataset with territorial information.

This information can be divided into three types: a) the geographic boundaries of the smaller administrative units (the parishes), b) the geographic boundaries of familiar

[3] Sometimes that spatial point is given through precise geographic latitude and longitude coordinates – this is the case in the database used here.

distinct urban places (localities) and c) a set of indicators that can be derived from i) processing census data and ii) the distribution of POIs across the territory and the accessibility to them via the road network.

The set of indicators described in (c) are obtained from a pre-processing step where retrieved open data is subject to a variable reduction modelling approach (principal component factor analysis – PCA-FA) to i) obtain summary (reduced) data (described by the scores of FA) and ii) ensure a better match with statistical assumptions of the HHPM and the OLS in particular (such as avoiding collinearity). The PCA-FA was applied separately to i) the socioeconomic, housing stock characteristics[4] and the land-use coverage[5] (retrieved from the INE and DGOTDU data sources, for each census tract polygons[6]) and ii) the accessibility index obtained by the data processing of the POIs' geographic location and distance matrix through the road network obtained for the regular square grid where each square has sides of 600 m; the accessibility index base at each census tract is an average of the accessibility index on the original grid. Moreover, the accessibility index is calculated for each POI category described in Table 1.

The summary statistics for spatial data and PY data is described in Table 2; the column notes include a short description of the meaning of positive values associated with the component scores resulting from the PCA-FA pre-processing.

Table 1. Collected geographic data – themes

Census tract data		POIs data		
Thematic categories	N. of classes	Thematic categories	N. of classes	N. of POIs
Type of building	2	Workplace poles (1)	1	21
Number of floors in building	2	Transportation (2)	4	38
Dwelling size	2	Education (3)	6	110
Building age	9	Health (5)	4	30
Dwelling type of occupation	2	Food stores and basic services (6)	4	223
Population age	4	Leisure (7)	3	449
Population education	5			
Population work place location (outside the municipality)	1			
Population work sector (tertiary only)	1			
Land use (urban category only)	3			

[4] This data is referenced to the public Census 2011 dataset, available at https://bit.ly/3ssuqNj.

[5] This data is referenced to the 2010 land use coverage – COS2010, available at https://bit.ly/3tSS99M.

[6] The statistical subsection is a georeferenced polygon which is closely similar to the urbanistic concept of a city block – more information on: https://smi.ine.pt/Conceito/Detalhes/1926.

Table 2. Data summary statistics

	Avg	Std. Dev	Min	Max	Notes
Price (€)	120,215	123,534	14,000	988,930	
Area (m2)	110.82	76.21	22.00	915.77	
Age (years)	26.42	11.00	4.00	78.00	
Backyard/land area (ratio)	0.26	1.02	0.00	14.24	
Type (categorical) (n. of records by class)	Flats (1) = 557, Single house (0) = 68				
Scores PCA-FA – A1	−0.85	0.87	−2.61	1.70	Low density land occupation
Scores PCA-FA – A2	0.42	0.71	−1.64	2.32	Prevalence of low qualified pop
Scores PCA-FA – A3	−0.01	0.94	−4.04	1.93	Prevalence of employed pop. with basic classifications and working in tertiary sector
Scores PCA-FA – A4	−0.17	0.84	−2.45	4.49	High urban density associated with old buildings
Scores PCA-FA – A5	−0.31	1.25	−2.43	2.74	Prevalence of old, non-active pop. and buildings built on the 60s and 70s
Scores PCA-FA – A6	−0.06	1.01	−3.39	4.93	Prevalence of younger pop. and buildings built in the 00s
Scores PCA-FA (accessibility) – B1	0.76	0.96	−1.69	3.98	Close to train stations, basic schools, ATM & postal services and leisure facilities
Scores PCA-FA (accessibility) – B2	0.30	1.03	−1.52	3.37	Close to main highway nodes and to the townhall (Sintra city centre)
Scores PCA-FA (accessibility) – B3	0.19	0.93	−2.51	2.01	Close to principal itineraries nodes
Parishes (categorical) (statistics about the n. of records)	48.08	40.96	4.00	172.00	13 parishes of 20 Sintra municipality parishes
Locality (categorical) (statistics about the n. of records)	20.16	18.37	2.00	54.00	31 localities of 163 Sintra municipality localities

3.2 Case Study and Spatial Data

Sintra is one of the most densely populated Portuguese municipalities. It is part of the Lisbon Metropolitan Area, and large parts of the territory are dominated by suburban settlements – mainly across the Lisbon–Sintra train line. As the spatial variables described in Table 2 can anticipate, this is a territory characterized by compelling territorial patterns. Here it is possible to find small neighbourhoods where discrepant spatial characteristics can be seen at different territorial dimensions – social, economic, land use, historical landmarks and geomorphological, among others. This makes this case study interesting for this work. It clearly emphasizes the modelling challenges as spatial data can be essential elements to understand housing price drivers.

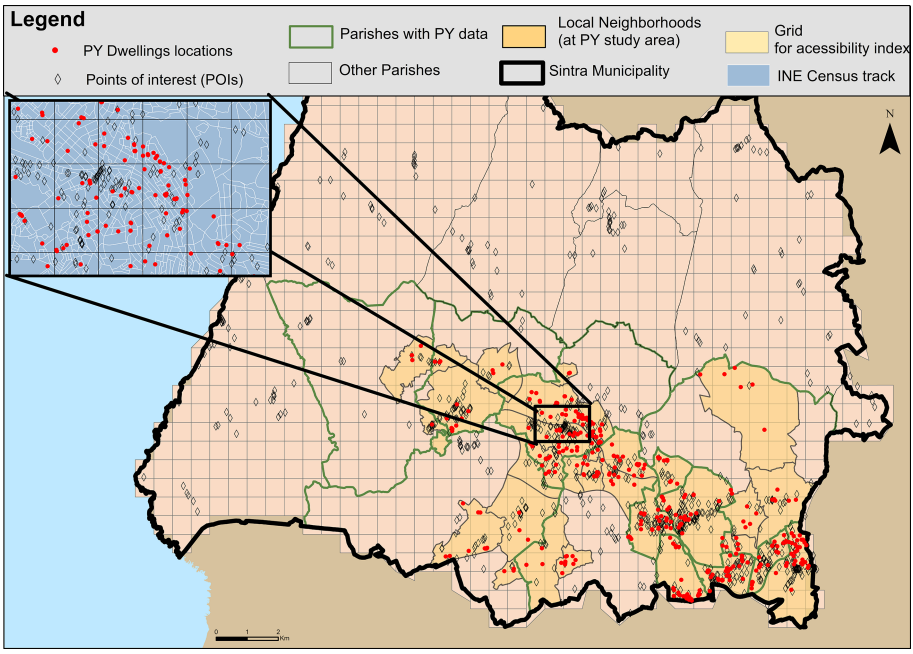


Fig. 1. PY Dataset and spatial data of Sintra municipality. The maps show the multiple spatial georeferentiation and resolution available. (Color figure online)

However, it is essential to highlight that PY data is not fully representative of all the Sintra municipality territory. PY data is concentrated in the high-density locations, namely the parishes crossed by the train line and most relevant road networks (highways), plus the city centre parish (where the train line ends). Rural areas and historical places are absent from the PY data. However, as can be anticipated from Table 2, the physical attributes of the dwellings present a high dispersion, showing that dwellings on the data revealed the great diversity of the Sintra municipality. The diversity reinforces the idea of a territory where high levels of heterogeneity can be observed. In fact, in a small area, it is possible to find housing complexes with characteristics close to slums side by side with housing complexes of very high quality, occupied by some of the wealthiest families living in Portugal. Figure 1 shows Sintra municipality and some spatial data features, which were retrieved to be combined with the PY dataset.

Observing the PY data (red dots), it can be identified that the spatial data distribution is not uniform. In fact, it is concentrated in well-known, highly populated places with a suburban nature.

3.3 Estimation Framework

Basic Specification of HHPM

Despite the diversity of available model specifications, one important point is to correctly and completely identify the relevant explanatory variables. As argued before, this

is a particular challenge and remains naturally uncertain, given the enormous spatial challenges related to HHPM. The spatial data enhancement presented before will be adopted here as an answer to achieve this model assumption.

Another important point in model specification is the nature of the functional relation between dependent (Y, the price or, in this study, the price per square metre) and independent variables (X). Further studies have shown that the housing market implies a non-linear pricing structure [31]. As standard regression models assume linearity, then a variable transformation will be performed. The Box-Cox transformation toolbox is the usual approach; in housing market models a strictly log-log or semi-log specification is commonly used – an option which will be followed here.

Specification of Dimension F and T

The available intrinsic attributes of the PY dataset are reduced, but they correspond to the usual main physical dwelling characteristics. It will be assumed that the potential effects of missing attributes will be negligible for the objectives of this work. Thus, the specifications follow including variable transformations such:

$$H = \beta_1 lnArea_{Dwel.} + \beta_2 lnAge_{Dwel.} + \beta_3 D_{DwellingType} + \beta_4 Ratio_{DwArea/LandArea} \quad (4)$$

where $lnArea_{Dwel.}$ is the dwelling area (m^2) transformed into a natural logarithm, $Age_{Dwel.}$, is the dwelling age transformed into a natural logarithm, $D_{DwellingType}$ is a dummy variable identifying a single-house (1) or a flat (0) and $Ratio_{DwArea/LandArea}$ is the ratio between dwelling area and the open or backyard area. β_i are the hedonic prices of each variable to be estimated.

Time stationarity of PY data will be achieved through a standard time fixed effect specification, as follows:

$$H = \sum_{i=2007}^{2014} \beta_i D_{Ti}$$

where D_i is the dummy variable identifying the year when the dwelling price was stored in the database. It comprises dummies for the years between 2008 and 2014. The 2015 year dummy is dropped to avoid the dummy trap [32].

Specification of Dimension S

The focus of this work is to present an exploratory analysis of the spatial modelling challenges related to the two usual spatial model pitfalls: a) the statistical difficulties of dealing with the heterogeneity and correlation of spatial data, and b) the different detail (precision) related with the type of spatial representation (points or polygons), the kind of georeferenced detail (such as census tracks, administrative boundaries or regular spatial grids) and the compatibility between different spatial data sources (for example, how to merge diverse spatial representation or spatial data resolutions).

Thus, in addition to an initial (benchmark) model (without spatial data – M0), five models were investigated (see Table 3 to a general overview). In detail, each model comprises:

- **M1:** Spatial features are considered as captured by the assignment of each dwelling to its corresponding parish. In this model, spatial effects are captured through the

classical approach of fixed effects: dummy variables for each parish (minus one, to avoid the dummy trap) will be accomplished to the model.

- **M2:** This model is similar to M1 but the territorial unit considered is the locality[7] (neighbourhood) information provided by INE. With this model, it is possible to compare different spatial resolutions (parishes and neighbourhoods) for similar spatial data types (polygons).
- **M3:** The third model specification intends to analyse the effect of use quantitative variables measuring the spatial features of the dwelling's surroundings rather than an explicit assignment of a dwelling to a territorial unit (which is exogenously defined).
- **M4:** This model will introduce the geographic weight regression (GWR) framework – namely in its recently time-space variant (GTWR) as referenced before – to include time non-stationarity presented in PY data. In this (and next model, M5), the time effects dummy are dropped. The GWR approach is selected because its focus is on producing reliable point estimates, which relax the need for a priori and strong assumptions on the territorial partitions to considered. Moreover, GWR has gained relevance in appraisal decision support systems.
- **M5:** Finally, the intention of the fifth model is to explore if GTWR will benefit from spatial data enhancement, namely, including explanatory variables whose parameters estimations will be obtained by a similar spatial (geographic) weighted scheme .

Table 3. General overview of the spatial modelling specification strategies adopted

Models	Spatial components specifications
M2; M3;	$H = \sum_{i=1}^{S} \beta_i D_{Si}$ Where D_{Si} is a dummy for each parish (M2) or each locality (M3)
M4; M5;	$\sum_{s=1}^{S} X_s \beta_s(u, v)$ Where X_s are each intrinsic attribute (M4) or each intrinsic attribute plus each one of the PCA-FA territorial indicators (M5)

Note that in GWR the Gaussian kernels with fixed bandwidths are adopted, chosen following a cross-validation (CV) process for each of the models M4 and M5, respectively. All data processing and modelling steps are performed in R language; for GTWR modelling, the GWmodel R library was selected [33].

4 Results

4.1 Model Validation and Comparison

Model comparison and basic model validation is performed through two types of model performance assessment approach: a) the standard classical statistical indicators – adjusted R2 and AIC – and b) the standard indicators advocated by guidelines on property valuation performance [34] – the coefficient of dispersion (COD) and price-related differential (PRD) (see Table 4).

[7] See INE for further description https://smi.ine.pt/Conceito/Detalhes/2990.

Table 4. Selected model evaluation indicators

Statistical measures	Real estate appraisal measures
$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}$ $AdjR^2 = 1 - \left(\frac{n-1}{n-k}\right)\left(1 - R^2\right)$	$COD = \frac{100}{n}\frac{\sum_{i=1}^{n}\left\lvert \frac{\hat{y}_i}{y_i} - Median\left(\frac{\hat{y}_i}{y_i}\right)\right\rvert}{Median\left(\frac{\hat{y}_i}{y_i}\right)}$
$AIC = -2\log(L) + 2K$	$PRD = \frac{Mean\left(\frac{\hat{y}_i}{y_i}\right)}{\sum_i^n \hat{y}_i / \sum_i^n y_i}$

The statistical approach follows the well-established model fit evaluation measures: the adjusted coefficient of determinations. R^2 provides a straightforward interpretation: the amount of variation accounted for in the fitted model. A pitfall of this measure is that R^2 always increases with the model size; the adjusted R^2 tries to limit this effect by adding a penalization to the coefficient estimation based on the number of variables. The Akaike information criteria (AIC) are usually assumed to be more robust to the effect of increasing model size, although the value calculated is less meaningful for non-statistical experts. Both measures are well-known in model evaluation and model comparison settings.

As one focus of this work is on the appraiser's needs for reliable models to help the valuation process, it is valuable to introduce two evaluation indicators advocated by real estate appraisal guidelines, which ensures model estimates will follow the uniformity and equity guidelines of the valuation process. COD is a dispersion coefficient focused on the uniformity of the set of evaluations performed, and PRD is concentrated in vertical equity across the set of valuations. In IAAO guidelines, the acceptability threshold for single-family homes is set to COD value between 5 and 15. PRD threshold is set between 0.98 and 1.03, with values above (below) this range providing evidence of regressivity (progressivity).

4.2 Modelling Results and Discussion

Table 5 shows the results obtained, where M4 and M5 can be identified as producing the best performance in all of the validation and comparison indicators. This performance is easily explained by the GWR/GTWR model's specificities reported in the theoretical review section: the GWR/GTWR model allows a local adjustment of the model which results in greater flexibility concerning the spatial specification. The presence of non-uniform phenomena of heterogeneity and spatial dependence in space naturally hinders the capacity for specifications that limit the scope of these effects a priori, to adjust correctly to local peculiarities. The difference in performance between models M4 and M5 and model M3 also highlights the role of this greater flexibility. By restricting the estimates to the observation (spatial) points, the M3 model is more vulnerable to the spatial pitfalls effects, particularly spatial correlation; this can explain the lower comparative performance.

Table 5. Results

	M0 (F)	M1 (F + parish)	M2 (F + places)	M3 (F + L)	M4 (F GTWR)	M5 (F + L GTWR)
N. of var	4 + 7	4 + 7 + 12	4 + 7 + 30	4 + 7 + 6 + 3	4	4 + 6 + 3
Adj. R2	0.840	0.867	0.888	0.869	0.881	0.878
AIC	−73.879	−179.493	−266.022	−190.548	−322.783	−318.029
COD	13.485	12.454	11.252	12.803	12.281	12.188
PRD	1.058	1.049	1.038	1.049	1.038	1.036

As concerns GWR/GTWR, it is essential to highlight some possible general drawbacks. First, it needs to be borne in mind that as GWR/GTWR models focus on local estimations, they are more susceptible to data quality and spatial data representativeness; these issues can induce various model pitfalls – such as model overfitting or the small area estimation traps [35].

Another critical point in GWR/GTWR is the need to include high-resolution spatial data. One of the interesting features of the model is that it takes advantage of the geographic coordinates to implement a weighting scheme in the process of parameter estimation. This is usual with data that is not fully compatible with the best resolution (as is the case with M5, where additional spatial variables are defined not for each dwelling's exact point but the centroid of the tract census).

Although the M5 model presents a lower level of performance than the M4 model on the statistical indicators, note that it offers slightly superior performance in terms of the remaining two appraisal performance indicators. Given the minor differences between the performance of M4 and M5, a simple possible explanation is that the statistical measures selected tend to penalize models with more variables. Moreover, the M4 model may present a tendency towards overfitting the data sample.

Finally, one of the weaknesses of the GWR is its lower capacity to provide the evaluating agent with an immediate perception of the contribution of tangible spatial characteristics to the price estimation (although, in the M5 model, when evaluating the contribution of spatial variables, this weakness is mitigated).

Models M1 and M2 should not be forgotten in this discussion. They reveal the determining role that coherent territorial partitions can play as explanatory drivers of dwelling prices. Moreover, spatial data resolution is sometimes limited (for example, see the real estate listing portals in Portugal, such as Casa Sapo or Imovirtual), which means these model specifications remain an adequate model approach' Moreover, this type of specification also has the advantage of adapting more effectively to the tacit knowledge of the expert in the local housing market, for which these spatial partitions encode a remarkable amount of information (about their characteristics).

The results obtained underline the need to develop a more sophisticated validation scheme. It should be noted that schemes such as bootstrap or cross-validation have not yet been implemented but are under development: they have specific challenges, such

as the need to ensure sampling consistency (for example, in models of type M1 and M2, it should be necessary to ensure that sampling will provide the spatial stratification associated the records in each spatial unit). Moreover, the information revealed by COD and PRD indicators points to values outside acceptable thresholds. This needs further investigation.

5 Conclusions

For an applied case study, this work shows the contributions of spatial data enhancement to the efforts to deploy reliable housing price models to be embedded on housing appraisal support systems – such as AVMs. The results show that spatial data can be incorporated through different specifications which support complementary notions on the role of space as a driver of housing prices. Further, the other model specifications can offer practitioners the flexibility to adapt their models to the spatial extent, spatial data types and spatial compatibility of the different datasets. Data analysts will be aware of the strengths and weaknesses of each model specification.

From a general point of view, this work highlights the complexity of space in the housing market model. It identifies its known (spatial) pitfalls on empirical applications – heterogeneity and spatial dependence. Moreover, it reveals the challenges that emerge both in econometric terms (including the model's theoretical assumptions) and the necessary interpretation of model results – as required by the appraisal activities' codes.

Although considerable uncertainty prevails about the interchange between spatial heterogeneity and spatial dependence, this work reinforces flexible but remaining simple modelling approaches regarding these theoretical and technical debates and offers a reasonable and reliable solution to the real estate appraisal agent. Thus, comprehensible and parsimonious appraisal support models are a crucial feature advocated by international guidelines. Despite the increasing use of machine learning model approaches, this work points to the need to avoid hidden spatial challenges by increasing black-box adoptions, despite the better performance in predicting house prices.

The housing market has undergone profound transformations in societies; demographic dynamics, social modifications of the population, lifestyle changes and families' preferences have led to new housing demands and requirements. Thus, it is expected that improving housing appraisal support systems will help its role in supporting market agents to produce more informed investment decisions. Moreover, a transparent appraise support model on its spatial features will help market agents and policymakers (or spatial planners) better understand the potential interplays between changes in the territorial systems and housing market behaviour.

Funding Information:. This work has been financially supported by the project "Drivers of urban transformation – DRIVIT-UP (POCI-01–0145-FEDER-031905) funded by FCT - Fundação para a Ciência e a Tecnologia and co-funded by FEDER through COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI).

References

1. Bourne, L.S.: The Geography of Housing. Edward Arnold, London (1981)
2. Can, A.: The measurement of neighborhood dynamics in urban house prices. Econ. Geogr. **66**(3), 254–272 (1990). https://doi.org/10.2307/143400
3. Galster, G.: On the nature of neighbourhood. Urban Stud. **38**(12), 2111–2124 (2001). https://doi.org/10.1080/00420980120087072
4. Palm, R.: Spatial segmentation of the urban housing market. J. Econ. Geogr. **54**, 210–221 (1978)
5. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., French, N.: Real estate appraisal: a review of valuation methods. J. Property Investment Financ. **21**(4), 383–401 (2003). https://doi.org/10.1108/14635780310483656
6. d'Amato, M., Kauko, T.: Appraisal methods and the non-agency mortgage crisis. In: d'Amato, M., Kauko, T. (eds.) Advances in Automated Valuation Modeling, pp. 23–32. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49746-4_2
7. Wang, D., Li, V.J.: Mass appraisal models of real estate in the 21st century: a systematic literature review. Sustainability **11**(24), 7006 (2019). https://doi.org/10.3390/su11247006
8. Marques, J.: The Notion of Space in Urban Housing Markets. Universidade de Aveiro (2012)
9. Batista, P.: The interaction structure of e-territorial systems: territory, housing market and spatial econometrics/A estrutura de interação de um sistema e-territorial. Território, mercado de habitação e econometria espacial. Universidade de Aveiro (2019)
10. Marques, J.L., Batista, P., Castro, E.A.: Espaço e território no contexto do desenvolvimento regional. In: Marques, J.L., Carballo-Cruz, F. (eds.) 30 anos de ciência regional em perspetiva, Almedina, Coimbra, Portugal, pp. 11–46 (2021)
11. Marques, J.L., Batista, P., Castro, E.A., Bhattacharjee, A.: Housing consumption. In: Chkoniya, V., Madsen, A.O., Bukhrashvili, P. (eds.) Anthropological Approaches to Understanding Consumption Patterns and Consumer Behavior, pp. 265–285. IGI Global (2020)
12. Moran, P.A.P.: A test for the serial independence of residuals. Biometrika **37**(1/2), 178–181 (1950). https://doi.org/10.2307/2332162
13. Anselin, L.: Local indicators of spatial association-LISA. Geogr. Anal. **27**(2), 93–115 (2010). https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
14. Rosen, S.: Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. **82**(1), 35–55 (1974)
15. Malpezzi, S.: Hedonic pricing models: a selective and applied review. In: Gibb, K., O'Sullivan, A. (eds.) Housing Economics and Public Policy: Essays in Honour of Duncan Maclennan, pp. 67–89. Blackwell Science, Oxford (UK) (2003)
16. Stull, W.J.: Community environment, zoning, and the market value of single-family homes. J. Law Econ. **18**(2), 535–557 (1975)
17. Paelinck, J.H.P.: Spatial econometrics: a personal overview. Stud. Ekon. **152**, 106–118 (2013)
18. Pace, R.K., LeSage, J.: Spatial econometrics and real estate. J. Real Estate Financ. Econ. **29**(2), 147–148 (2004)
19. Corrado, L., Fingleton, B.: Where is the economics in spatial econometrics. J. Reg. Sci. **52**(2), 210–239 (2012)
20. Elhorst, J.P.: Spatial Econometrics: From Cross-Sectional Data to Spatial Panels. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-40340-8
21. Landry, S.M., Chakraborty, J.: Street trees and equity: evaluating the spatial distribution of an urban amenity. Environ. Plan. A Econ. Sp. **41**(11), 2651–2670 (2009). https://doi.org/10.1068/a41236
22. Getis, A., Aldstadt, J.: Constructing the spatial weights matrix using a local statistic. Geogr. Anal. **36**(2), 90–104 (2004). https://doi.org/10.1111/j.1538-4632.2004.tb01127.x

23. Anselin, L., Li, X.: Tobler's law in a multivariate world. Geogr. Anal. **52**(4), 494–510 (2020). https://doi.org/10.1111/gean.12237
24. Brunsdon, C., Fotheringham, A.S., Charlton, M.E.: Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. **28**(4), 281–298 (2010). https://doi.org/10.1111/j.1538-4632.1996.tb00936.x
25. Cleveland, W.S., Devlin, S.J.: Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. **83**(403), 596–610 (1988). https://doi.org/10.1080/01621459.1988.10478639
26. Fotheringham, A.S., Crespo, R., Yao, J.: Geographical and temporal weighted regression (GTWR). Geogr. Anal. **47**(4), 431–452 (2015). https://doi.org/10.1111/gean.12071
27. Bourassa, S., Hamelink, F., Hoesli, M.: Defining housing markets. J. Hous. Econ. **8**, 160–183 (1999)
28. Meen, G.: Modelling Spatial Housing Markets: Theory, Analysis, and Policy, vol. 2. Springer, Boston (2001). https://doi.org/10.1007/978-1-4615-1673-6
29. RICS Valuation – Global Standards (2019)
30. European Valuation Standards 2020 (2020)
31. Sheppard, S.: Hedonic analysis of housing markets. In: Cheshire, P.C., Mills, E.S. (eds.) Handbook of Regional and Urban Economics, vol. 3, pp. 1595–1635. Elsevier, Amesterdam, The Netherlands (1999)
32. Wooldridge, J.: Introductory Econometrics: A Modern Approach, 4th edn. Cengage Learning, Mason, USA (2008)
33. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P.: GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. **63**(17), 1–50 (2013). http://arxiv.org/abs/1306.0413. Accessed 28 Mar 2021
34. Standard on Automated Valuation Models (AVMs) (2018)
35. Openshaw, S.: Ecological fallacies and the analysis of areal census data. Environ. Plan. A **16**(1), 17–31 (1984). https://doi.org/10.1068/a160017