J Geogr Syst (2018) 20:27–55
https://doi.org/10.1007/s10109-017-0257-y

ORIGINAL ARTICLE

Housing price prediction: parametric versus semi-parametric spatial hedonic models

José-María Montero[1] · **Román Mínguez**[2] · **Gema Fernández-Avilés**[1]

Received: 25 July 2016 / Accepted: 13 July 2017 / Published online: 3 August 2017

Abstract House price prediction is a hot topic in the economic literature. House price prediction has traditionally been approached using a-spatial linear (or intrinsically linear) hedonic models. It has been shown, however, that spatial effects are inherent in house pricing. This article considers parametric and semi-parametric spatial hedonic model variants that account for spatial autocorrelation, spatial heterogeneity and (smooth and nonparametrically specified) nonlinearities using penalized splines methodology. The models are represented as a mixed model that allow for the estimation of the smoothing parameters along with the other parameters of the model. To assess the out-of-sample performance of the models, the paper uses a database containing the price and characteristics of 10,512 homes in Madrid, Spain (Q1 2010). The results obtained suggest that the nonlinear models accounting for spatial heterogeneity and flexible nonlinear relationships between some of the individual or areal characteristics of the houses and their prices are the best strategies for house price prediction.

Keywords Housing prices · Semi-parametric spatial hedonic models · Generalized additive models · Penalized splines · Mixed models

The original version of this article has been revised: In Table 1, there is a misprint in equations under models SDEM and SEM. Now, Table 1 has been corrected.

Electronic supplementary material The online version of this article (doi:10.1007/s10109-017-0257-y) contains supplementary material, which is available to authorized users.

✉ José-María Montero
jose.mlorenzo@uclm.es

[1] Department of Statistics, Faculty of Law and Social Sciences, University of Castilla-La Mancha, 45071 Toledo, Spain

[2] Department of Statistics, Faculty of Social Sciences, University of Castilla-La Mancha, 16071 Cuenca, Spain

JEL Classification C21 · R10 · R31

1 Introduction

There is a great deal of literature focusing on housing price prediction (and on identifying the factors that have the greatest effect on these prices). To this end, a number of models have been used, ranging from traditional hedonic regressions (Cebula 2009) to more sophisticated strategies such as generalized additive models (GAMs) with integrated model selection using penalized regression splines (Wood and Augustin 2002), geographically weighted regression (GWR) with a non-Euclidean distance metric (Lu et al. 2014), mixed GWR models (Helbich et al. 2013b), spatial econometric models (Fernández-Avilés et al. 2012), alternatives to those standard spatial econometric models (von Graevenitz and Panduro 2015) and geostatistical methods such as kriging and cokriging (Kuntza and Helbichab 2014), to name just a few.

Traditional hedonic house price models, however, are either linear or, in most cases, intrinsically linear (they include squared terms, natural logarithms and so on) and do not account for the spatial nature of property data. Both spatial autocorrelation and spatial heterogeneity are inherent in property data (Mark and Goldberg 1988; Fletcher et al. 2000) and/or property attributes (Straszheim 1974). Hence, they should be accounted for in hedonic house price model specifications, along with potential nonlinear relationships between some of the covariates and the response variable. These nonlinear relationships can be specified not only as per conventional practice in standard nonlinear house price models (which are intrinsically linear), whereby the parametric form of such relationships is suggested by the researcher, but also by relaxing this parametric form assumption and adopting a nonparametric form.[1] This way, the researcher does not have to suggest the type of relationship (function) existing between the variables, but rather it is the model that defines this (smooth) function, a feature which represents a particular advantage when dealing with massive databases. Hereafter, when we refer to "nonlinear" or "nonlinearities" we are indicating that there is a nonlinear relationship between variables of the last type (smooth and nonparametric), regardless of whether or not the model includes squared terms, logarithms, etc., as is the case in many hedonic house price models. In addition, according to the literature, housing prices are nonstationary[2] due to, among other factors, similarities between neighboring houses and the similar behavior of homebuyers (Anselin 1988), as well as the influence of human communication and market demands (Jahanshiri et al. 2011).

Jahanshiri et al. (2011) point out that, for the reasons mentioned above, the last few years have seen the emergence of three successful lines of research that take into account spatial effects in the real estate field: spatial econometrics,

[1] Nonlinearity in the parameters also results in a nonlinear model, but this is not typical in hedonic house price models and it goes beyond the scope of this article.

[2] That is, housing prices cannot be explained by a simple "global" model. In this vein, Brunsdon et al. (1996) argue that: (a) relationships can vary significantly over space and a "global" estimate of the relationships may obscure interesting geographical phenomena; and (b) variation over space can be sufficiently complex that it invalidates simple trend-fitting exercises.

geographical local modeling and geostatistics. Spatial econometric research contributed the spatial autoregressive (SAR) and spatial error (SEM) models popularized by Anselin (1988), which allow for spatial autocorrelation of housing prices, property attributes and errors, as well as of pair-wise combinations of housing prices, property attributes and errors. Geographic research yielded local models of GWR and moving window regression (MWR), two special types of regression that divide the area under study into local sections devised to neutralize the heterogeneity of autocorrelations (Fotheringham et al. 2002). The basic difference between GWR and MWR is that in MWR the weights given to the observations in the same window are equal, while GWR uses distance decay functions. From the area of geostatistical research, different types of kriging and cokriging (the multivariate version of kriging) models emerged (Cressie 1993; Montero et al. 2015). Unlike regression methods, kriging models primarily deal with the spatial dependencies in property data and aim to predict the price of an unsold property using the spatial correlations in the prices of sold properties, which are estimated by variography (see Montero et al. 2015). Kriging differs from SAR models in the calculation of the weights: kriging weights are based on the semivariogram (or covariogram) function which models the spatial dependencies in the price of properties, whereas SAR weights are based on the spatial contiguity between the observed properties. The combination of the hedonic pricing model with kriging, known as regression kriging, has also been used in the housing price prediction field [some of the best-known examples are Dubin (1998) and Anselin et al. (2004)].

Our research is grounded in the field of spatial econometrics and contributes to the literature on housing price prediction by assessing (via cross-validation) the out-of-sample prediction performance of 14 competing models: the conventional a-spatial hedonic model (HM) and 13 additional model specifications accounting for spatial autocorrelation and/or spatial heterogeneity and/or nonlinear relationships between the response and some of the predictor variables.

More specifically, we first include a spatial drift (SD), via penalized splines (PS)—which we denote with the acronym PSSD—in HM but, more interestingly, we also do so in the spatial lag model (SLM), the spatial Durbin model (SDM) and the traditional a-spatial GAM, giving rise to the PSSD-SLM, PSSD-SDM and the PSSD-GAM. The SLM and SDM autoregressive structures are also combined with a GAM. In addition, we consider two more specifications that account for spatial autocorrelation and spatial heterogeneity as well as for nonlinear relationships (nonparametrically specified via smooth functions). We do so by including a GAM term in the PSSD-SLM and the PSSD-SDM, which results in the PSSD-GAM-SLM and the PSSD-GAM-SDM (which allows for global spillovers), respectively.

The semi-parametric counterparts of SLM, SEM and SDM, augmented with a smooth spatial drift, had been already estimated in Basile et al. (2014), but their focus was on in-sample prediction rather than out-of-sample forecasting performance. Among other significant differences from Basile et al. (2014), are the fact that the spatial weight matrix is not sparse but includes the neighbors in a 1.5-km radius (fewer than three neighbors per observation, on average, in Basile et al. 2014), and that the estimation of the models belonging to the PSSD family has been

performed with our own new codes based on the SAP library (Rodriguez-Alvarez et al. 2015), which provides analytical solutions for the penalty parameters associated with the spatial drift. In addition, these new codes significantly reduce the computation time for model estimation, which is an advantage when dealing with massive databases.

It is worth noting the difficulty involved in the estimation of housing price specifications including a spatial drift via penalized splines, a spatial autoregressive model and nonlinear relationships (nonparametrically specified via smooth functions) between certain house attributes and housing prices. To tackle this challenge, we contribute to the literature by extending the approach developed for the first time in Montero et al. (2012), and Basile et al. (2014), based on the representation of the hedonic housing price specifications as mixed models (though it may not be intuitive, models including PS terms admit representation as a mixed model) to complex hedonic specifications accounting for spatial autocorrelation, spatial heterogeneity and nonlinearity. This approach allows the simultaneous estimation of the spatial drift, the smoothing parameters, the spatial autocorrelation coefficients, the impacts of the covariates that enter into the model parametrically and the smooth nonparametrically specified functions of the predictor variables that are assumed to have a flexible nonlinear relationship with the price of houses.

The estimation of the 14 competing models is carried out using data from Madrid—one of the European cities whose economic performance is most affected by the housing sector—just after the burst of the real estate bubble. We use a massive proprietary database consisting of the price and 19 other characteristics of 10,512 owner-occupied single-family homes. As far as we know, this is the largest database ever used to analyze the Madrid housing market. It is of note that the estimation procedure we contribute allows for model estimation with massive databases, which is a remarkable difference with previous research using the same approach.

The remainder of the paper is organized as follows. Section 2 outlines the description of the competing models used in the article as well as the estimation method for models belonging to the PSSD or PSSD-GAM families. Section 3 is devoted to a case study in which the variables, data and the spatial weight matrix are briefly described and then the results regarding the predictive capabilities of the competing spatial models are detailed and discussed. Section 4 concludes the paper and suggests possible future research directions.

2 The hedonic house price model and spatial model variants

In this section, we briefly outline the 14 hedonic house price models estimated in this article in order to assess their predictive capabilities. The competing models include the conventional a-spatial hedonic house price model, which has been considered as the base model, 12 spatial variants and a GAM specification which has been included for the sake of comparison. Table 1 lists these models—which have been divided into two groups: parametric and semi-parametric—and provides information on their specification (first and second columns), highlighting whether

Table 1 Competing hedonic house price models: Parametric and semi-parametric model specifications

Model	Specification	Spatial lag			Spatial drift	Nonparametric functions of covariates
		Response	Covariates	Error		
Parametric models						
(i) Conventional a-spatial hedonic model						
HM	$\mathbf{y} = \alpha\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$					
(iia) SAR models						
SLM	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	x				
SDM	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	x	x			
SDEM	$\mathbf{y} = \alpha\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$		x	x		
(iib) SEM model						
SEM	$\mathbf{y} = \alpha\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = l\,\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$			x		
Semi-parametric models						
(iii) PSSD-HM and PSSD-SAR models						
PSSD-HM	$\mathbf{y} = f(s_1, s_2) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$				x	
PSSD-SLM	$\mathbf{y} = f(s_1, s_2) + \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	x			x	
PSSD-SDM	$\mathbf{y} = f(s_1, s_2) + \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	x	x		x	
(iv) GAM model						
GAM	$\mathbf{y} = \sum_{r=1}^{l} f_{1,r}\left(x_r^+\right) + \alpha\mathbf{i}_n + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$					x
(v) GAM-SAR models						
GAM-SLM	$\mathbf{y} = \sum_{r=1}^{l} f_{1,r}\left(x_r^+\right) + \alpha\mathbf{i}_n + \rho\mathbf{W}\mathbf{y} + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	x				x
GAM-SDM	$\mathbf{y} = \sum_{r=1}^{l} f_{1,r}\left(x_r^+\right) + \sum_{r=1}^{l} f_{2,r}\left(\mathbf{W}x_r^+\right) + \alpha\mathbf{i}_n + \rho\mathbf{W}\mathbf{y} + \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	x	x			x

Table 1 continued

Model	Specification	Spatial lag			Spatial drift	Nonparametric functions of covariates
		Response	Covariates	Error		
(vi) PSSD-GAM model						
PSSD-GAM	$\mathbf{y} = f(s_1, s_2) + \sum_{r=1}^{l} f_{1,r}(x_r^+) + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$				x	x
(vii) PSSD-GAM-SAR models						
PSSD-GAM-SLM	$\mathbf{y} = f(s_1, s_2) + \sum_{r=1}^{l} f_{1,r}(x_r^+) + \rho\mathbf{W}\mathbf{y} + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	x			x	x
PSSD-GAM-SDM	$\mathbf{y} = f(s_1, s_2) + \sum_{r=1}^{l} f_{1,r}(x_r^+) + \sum_{r=1}^{l} f_{2,r}(\mathbf{W}x_r^+) + \rho\mathbf{W}\mathbf{y} + \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	x	x		x	x

they include (1) a spatial lag (and if so where) or not (third, fourth and fifth columns), (2) a nonparametric spatial drift (sixth column) and (3) smooth nonparametric functions of covariates (last column).

2.1 The hedonic price model as a base model

As stated above, the traditional parametrically specified HM is taken as a starting point. In HM [Table 1, (i)] $\mathbf{y}$ represents an $n \times 1$ vector of housing transaction prices, $\mathbf{X}$ is an $n \times k$ matrix containing the observations of the individual and areal characteristics associated with each dwelling (in our case $n = 10{,}512$ and $k = 21$), $\mathbf{i}_n$ is a $n \times 1$ unit vector for the intercept,[3] α is the intercept parameter, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression parameters and $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of *iid* disturbances with $N(\mathbf{0}, \sigma^2\mathbf{I})$ distribution.

The theoretical basis of hedonic price modeling can be found in Lancaster (1966), who argues that it is not the good itself that creates utility, but rather its individual characteristics. Following Helbich et al. (2013b), as housing characteristics are non-separable and traded in bundles, housing (and real estate in general) is usually treated as a heterogeneous good. Houses are valued for their utility-bearing attributes with implicit prices, which can be considered as the component's specific prices (McDonald 1997). Thus, according to Helbich et al. (2013b), a household implicitly chooses a set of different goods and services by selecting a specific object, and households aim to maximize their utility depending on their own social and economic characteristics.

2.2 Parametric spatial econometric model variants

The group of parametric specifications includes, together with the traditional HM, the conventional spatial autoregressive models that come under the denomination of global SAR models [Table 1, (iia)]: SLM and SDM (see Anselin 1988), though we also estimate the spatial Durbin error model (SDEM, see LeSage and Pace 2009, a useful model to account for the local spillovers present in many economic research problems). The SEM closes this group of parametric hedonic specifications [Table 1, (iib)]. We exclude the spatial autoregressive model with autoregressive disturbances (SARAR) from the list of global SAR models to be estimated due to identification problems in practice (Elhorst 2014).

In this group of models, $\mathbf{W}$ is a $n \times n$ row-stochastic matrix of spatial weights so that $\mathbf{Wy}$ is the vector of the average $\mathbf{y}$ over each house's neighbors (the vector of the mean prices of neighboring houses), $\mathbf{WX}$ is the matrix of spatially lagged covariates and $\mathbf{Wu}$ is the spatial lag of the error vector $\mathbf{u}$. Finally, ρ, $\boldsymbol{\theta}$ and λ are spatial parameters that weight the corresponding spatial lag. It is of note that the unity column has been removed from $\mathbf{X}$ (and included explicitly in an additive term) to avoid having singular

[3] We have removed the unitary column from $\mathbf{X}$ and included it explicitly in the model, in an additive term, in order to follow the same structure as in the parametric spatial variants of the model.

[4] For the sake of consistency, we proceed in the case of SLM as with the other parametric models.

matrices in the estimation process, since two of the three parametric spatial econometric model variants considered include spatially lagged covariates.[4]

Recent literature coming from spatial econometric research on property prices using SAR and SEM models includes Brasington and Hite (2005), Anselin and Le Gallo (2006), Small and Steimetz (2006), Neill et al. (2007), Anselin and Lozano-Gracia (2008), Bourassa et al. (2010), Osland (2010), Montero et al. (2011, 2012), Fernández-Avilés et al. (2012), Mínguez et al. (2013) and Basile et al. (2014).

2.3 Semi-parametric spatial hedonic models

The semi-parametric hedonic house price specifications considered in this study have been grouped in five categories [see Table 1, groups (iii) to (vii)]. The third group of models includes the PSSD-HM and PSSD-SAR model specifications. The first of these models (PSSD-HM) arises from the inclusion of a geo-additive nonparametric spatial drift $f(s_1, s_2)$ in HM to account for spatial heterogeneity, with s_1 and s_2 representing the geographic longitude and latitude of housing location, respectively.

The PSSD-SLM and PSSD-SDM specifications take both spatial autocorrelation and spatial heterogeneity into account by combining a nonparametric spatial drift with a standard SLM or SDM. It is of note that, with the family of PSSD models, the unity column removed from $\mathbf{X}$ does not appear explicitly in the specification of the models because it is included in the PSSD term.

There is also abundant literature on the use of parametric and nonparametric local models to account for spatial heterogeneity when it comes to modeling housing prices, since modeling spatial heterogeneity can be at least as important as modeling spatial dependencies. In the real estate field, it has been proven that hedonic prices can vary across space (Bourassa et al. 1999; Goodman and Thibodeau 2003; Bischoff and Maennig 2011; Helbich et al. 2013a). Local modeling is not the only way of modeling spatial heterogeneity, but it is the most widely used strategy. The most popular of the local models is GWR, and it has been widely applied in the real estate field. Pavlov (2000), Yu (2006), Bitter et al. (2007) and, recently, Manganelli et al. (2014) are some illustrative examples of GWR and GWR-type methodology applied to housing markets.

In addition, from the geostatistical perspective, there is a vast literature that addresses the spatial effects (usually spatial autocorrelation, but also spatial heterogeneity) in housing price prediction, yielding spatial interpolation surfaces, for which the nonparametric spatial drift $f(s_1, s_2)$ could be considered as an alternative. Chica-Olmo (1995), Gámez et al. (2000) and Montero and Larraz (2010, 2011) are examples of the literature in which kriging (the geostatistical procedure used for prediction) is applied to housing price prediction. A recent article by Cellmer (2014) outlines the possibilities and limitations of geostatistical methods in real estate market analysis. Cokriging, which also uses the information provided by one or more auxiliary variables, such as other property prices correlated with the price of the main real estate variable, has recently been used to predict housing location price (Chica-Olmo 2007), housing prices (Chica-Olmo et al. 2013), and to overcome the critical obstacle researchers have to face when only limited

data are available, which is the case when predicting commercial property prices (Montero et al. 2009).

Group (iv) in Table 1 relates to the pure GAM model specification. As stated at the beginning of the section, pure GAM is not a spatial specification, but is used as a base model for comparison purposes. GAM includes smooth functions of the covariates that are assumed to have a nonlinear, nonparametrically specified smooth relationship with the response variable (Wood 2006a, b). These covariates are denoted by the superscript '+', while $f_{1,r}\left(x_r^+\right), r = 1, \ldots, l$, represents smooth nonparametrically specified functions of said covariates. $\mathbf{X}^*$ is the matrix of the rest of the covariates (those that have a parametrically specified relationship with the response variable or that are of a qualitative nature). GAM specifications for house price prediction can also be found in real estate literature; widely cited articles include Pace (1993, 1998) and Mason and Quigley (2007).

In group (v) the SLM and SDM autoregressive structures are combined with the GAM model specification (instead of with a smooth spatial drift, as in group (iii)). This way, the two resulting specifications, GAM-SLM and GAM-SDM, account not only for spatial autocorrelation but also for smooth functional relationships (nonparametrically specified) between the covariates not included in $\mathbf{X}^*$ and the response variable (in the case of GAM-SLM), $f_{1,r}\left(x_r^+\right), r = 1, \ldots, l$, and also (in the case of GAM-SDM) between such covariates, spatially lagged, and the response, $f_{2,r}\left(\mathbf{W}x_r^+\right), r = 1, \ldots, l$.

Group (vi) in Table 1 extends the GAM specification by including a PSSD term, so that not only smooth nonparametrically specified nonlinearity is considered but also spatial heterogeneity (as is the case with the PSSD-GAM). Finally, the models of group (v), GAM-SLM and GAM-SDM, are also extended with a PSSD term, resulting in two new model specifications, PSSD-GAM-SLM and PSSD-GAM-SDM [group (vii)], which jointly consider spatial autocorrelation, spatial heterogeneity and nonlinearity.

The estimation of the semi-parametric models is carried out using penalized splines methodology (Eilers and Marx 1996). We present below the estimation algorithm for the most general semi-parametric specification in Table 1, the PSSD-GAM-SDM, since the other competing semi-parametric models are nested within this specification.

First, the spatial drift is expressed as a smooth function of s_1 and s_2 by using B-splines, that is $f(s_1, s_2) = (\mathbf{B}_{s_2} \square \mathbf{B}_{s_1})\boldsymbol{\delta}_s$, where $\mathbf{B}_{s_i}$, $i = 1, 2$ are the matrices containing the B-spline bases for each spatial coordinate, '$\square$' represents the row-wise Kronecker product or box product and $\boldsymbol{\delta}_s$ is the vector of coefficients associated with the box product of $\mathbf{B}_{s_2}$ and $\mathbf{B}_{s_1}$.

Second, as with the spatial drift, the additive GAM terms $f_{1,r}(x_r^+)$ and $f_{2,r}(\mathbf{W}x_r^+)$ are expressed in terms of B-spline bases as $f_{1,r}(x_r^+) = \mathbf{B}_r^+\boldsymbol{\delta}_r$ and $f_{2,r}(\mathbf{W}x_r^+) = \mathbf{W}\mathbf{B}_r^+\boldsymbol{\delta}_r^W$, respectively. $\mathbf{B}_r^+$ are the matrices containing the B-spline bases for each x_r^+ covariate, and $\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_r^W$ are the vectors of coefficients associated with the bases matrix and the spatially lagged bases matrix, respectively.

According to the representation of both the spatial drift and the functions of the covariates that are assumed to have a nonlinear nonparametrically specified smooth

relationship with the response in terms of B-splines, the PSSD-GAM-SDM specification can be rewritten as follows:

$$\mathbf{Ay} = (\mathbf{B}_{s_2}\square\mathbf{B}_{s_1})\boldsymbol{\delta}_s + \sum_{r=1}^{l}\mathbf{B}_r^+\boldsymbol{\delta}_r + \sum_{r=1}^{l}\mathbf{WB}_r^+\boldsymbol{\delta}_r^W + \mathbf{X}^*\boldsymbol{\beta} + \mathbf{WX}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}$. This allows for interactions between the spatial coordinates in the spatial drift, but not for interactions between covariates.

Third, in order to overcome the typical overfitting problem caused by the large number of columns in the B-spline bases matrices (especially in the case of the spatial drift, where interaction terms are involved), the vectors of coefficients $\boldsymbol{\delta}_s$, $\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_r^W$ $(r = 1,\ldots,l)$ are penalized using quadratic penalty matrices $\mathbf{P}_s$, $\mathbf{P}_r$ and $\mathbf{P}_r^W$ in the form $\boldsymbol{\delta}_s^T\mathbf{P}_s\boldsymbol{\delta}_s$, $\boldsymbol{\delta}_r^T\mathbf{P}_r\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_r^{WT}\mathbf{P}_r^W\boldsymbol{\delta}_r^W$, respectively, with:

$$\mathbf{P}_s = \kappa_{s_1}\mathbf{I}_{c_{s_2}}\otimes\mathbf{D}_{s_1}^T\mathbf{D}_{s_1} + \kappa_{s_2}\mathbf{D}_{s_2}^T\mathbf{D}_{s_2}\otimes\mathbf{I}_{c_{s_1}} \tag{2}$$

$$\mathbf{P}_r = \kappa_r\mathbf{D}_r^T\mathbf{D}_r,\quad r = 1,\ldots,l \tag{3}$$

$$\mathbf{P}_r^W = \kappa_r^W\mathbf{D}_r^T\mathbf{D}_r,\quad r = 1,\ldots,l \tag{4}$$

where $\mathbf{D}_{s_i}(i = 1,2)$ and $\mathbf{D}_r$ are discrete second order difference matrices, $c_{s_i}(i = 1,2)$ are the number of columns in matrices $\mathbf{B}_{s_i}(i = 1,2)$ and κ_{s_1}, κ_{s_2}, κ_r and κ_r^W are smoothing coefficients which govern the degree of penalization.

Fourth, in order to estimate the smoothing parameters together with the parameters of PSSD-GAM-SDM as expressed in Eq. (1) (otherwise these smoothing parameters should be determined a priori), Eq. (1) is represented as a mixed model (Lee and Durbán 2011):

$$\mathbf{Ay} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\omega} + \boldsymbol{\varepsilon}\quad \boldsymbol{\omega}\sim\mathrm{N}(\mathbf{0},\mathbf{G})\quad \boldsymbol{\varepsilon}\sim\mathrm{N}(\mathbf{0},\sigma^2\mathbf{I}) \tag{5}$$

where $\boldsymbol{\gamma}$ is a fixed effects parameter vector and $\boldsymbol{\omega}$ a random effects parameter vector with covariance matrix $\mathbf{G}$. The matrices involved in the mixed model (5) are obtained from Eqs. (1)–(4) as follows:

$$\mathbf{X} = [\mathbf{X}_{s_2}\square\mathbf{X}_{s_1}|\mathbf{X}^+|\mathbf{WX}^+|\mathbf{X}^*|\mathbf{WX}^*] \tag{6}$$

$$\mathbf{X}_{s_i} = [\mathbf{i}_n|\mathbf{s}_i],\mathbf{s}_i = \mathbf{s}_1,\mathbf{s}_2 \tag{7}$$

$$\mathbf{X}^+ = \left[\mathbf{x}_1^+|\mathbf{x}_2^+|\ldots|\mathbf{x}_l^+\right];\ \mathbf{X}^* = \left[\mathbf{x}_1^*|\mathbf{x}_2^*|\ldots|\mathbf{x}_k^*\right] \tag{8}$$

$$\mathbf{Z} = \left[\mathbf{Z}_{s_2}\square\mathbf{X}_{s_1}|\mathbf{X}_{s_2}\square\mathbf{Z}_{s_1}|\mathbf{Z}_{s_2}\square\mathbf{Z}_{s_1}|\mathbf{Z}_1^+|\mathbf{Z}_2^+\ldots|\mathbf{Z}_l^+|\mathbf{WZ}_1^+|\mathbf{WZ}_2^+|\ldots|\mathbf{WZ}_l^+\right] \tag{9}$$

$$\mathbf{Z}_{s_i} = \mathbf{B}_{s_i}\widetilde{\mathbf{U}}_{s_i},\ \mathbf{s}_i = \mathbf{s}_1,\mathbf{s}_2 \tag{10}$$

$$\mathbf{Z}_r^+ = \mathbf{B}_r^+\widetilde{\mathbf{U}}_r,\ r = 1,2,\ldots,l \tag{11}$$

$$\mathbf{G}=\sigma^2\begin{pmatrix}\kappa_{s_2}\widetilde{\boldsymbol{\Sigma}}_{s_2}\otimes\mathbf{I}_{q_{s_1}} & & & & & & & & \\ & \kappa_{s_1}\mathbf{I}_{q_{s_2}}\otimes\widetilde{\boldsymbol{\Sigma}}_{s_1} & & & & & & & \\ & & \kappa_{s_1}\mathbf{I}_{c_{s_2}-q_{s_2}}\otimes\widetilde{\boldsymbol{\Sigma}}_{s_1}+\kappa_{s_2}\widetilde{\boldsymbol{\Sigma}}_{s_2}\otimes\mathbf{I}_{c_{s_1}-q_{s_1}} & & & & & & \\ & & & \kappa_1\widetilde{\boldsymbol{\Sigma}}_1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & \kappa_l\widetilde{\boldsymbol{\Sigma}}_l & & & \\ & & & & & & \kappa_1^W\widetilde{\boldsymbol{\Sigma}}_1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & \kappa_l^W\widetilde{\boldsymbol{\Sigma}}_l\end{pmatrix}^{-1} \tag{12}$$

where $\widetilde{\boldsymbol{\Sigma}}_{s_i}(i=1,2)$ are diagonal matrices whose diagonal entries are the nonzero eigenvalues of the singular value decomposition of the penalty matrix. That is:

$$\mathbf{D}_{s_i}^T\mathbf{D}_{s_i}=(\mathbf{U}_{s_i,N}|\widetilde{\mathbf{U}}_{s_i})\begin{pmatrix}\mathbf{0}_{q_{s_i}} & \\ & \widetilde{\boldsymbol{\Sigma}}_{s_i}\end{pmatrix}\begin{pmatrix}\mathbf{U}_{s_i,N}^T\\ \widetilde{\mathbf{U}}_{s_i}^T\end{pmatrix}\quad \text{for } i=1,2, \tag{13}$$

with $\mathbf{U}_{s_i,N}$ and $\widetilde{\mathbf{U}}_{s_i}$ being the eigenvector matrices corresponding to the null space and the nonzero eigenvalues, respectively, of $\widetilde{\boldsymbol{\Sigma}}_{s_i}$, and $q_{s_i}(i=1,2)$ the order of the penalty. Similarly, as the same decomposition is done for $\mathbf{D}_r^T\mathbf{D}_r$ $(r=1,\ldots,l)$, $\widetilde{\mathbf{U}}_r$ is the eigenvector matrix corresponding to the nonzero eigenvalues of $\widetilde{\boldsymbol{\Sigma}}_r$, and q_r the order of the penalty.

Finally, once the PSSD-GAM-SDM specification has been represented as a mixed model, all the parameters of the model, that is, ρ, σ^2, $\boldsymbol{\gamma}$ and the smoothing parameters included in $\mathbf{G}$, can be estimated by restricted maximum likelihood (REML) [see Montero et al. (2012) and Basile et al. (2014), for details]. The restricted log-likelihood function of the above set of parameters is as follows:

$$\begin{aligned}\log L_R(\rho,\sigma^2,\kappa_{s_1},\kappa_{s_2}\kappa_r,\kappa_r^W)=&-\frac{1}{2}\left\{\log|\mathbf{V}|+\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|\right.\\ &\left.+(\mathbf{A}\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\gamma}})^T\mathbf{V}^{-1}(\mathbf{A}\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\gamma}})\right\}+\log|\mathbf{A}|,\end{aligned} \tag{14}$$

with $\mathbf{V}=\mathbf{Z}\mathbf{G}\mathbf{Z}^T+\sigma^2\mathbf{I}_n$ and $\hat{\boldsymbol{\gamma}}=(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{A}\mathbf{y}$.

The REML estimation of $\kappa_{s_1},\kappa_{s_2},\kappa_1,\ldots,\kappa_l,\kappa_1^W,\ldots,\kappa_l^W$ is carried out by applying the separation of anisotropic penalties (SAP) algorithm (Rodriguez-Alvarez et al. 2015), which provides a closed solution. Then, the score equation for ρ (conditional on the estimates of $\kappa_{s_1},\kappa_{s_2},\kappa_1,\ldots,\kappa_l,\kappa_1^W,\ldots,\kappa_l^W$) is solved numerically. This process is then iterated until it converges. Finally, we use the conventional unbiased estimator for the estimation of σ^2. As is well known, under the assumption of a correctly specified model, the REML estimators of the above parameters are unbiased and asymptotically efficient (Harville 1977).

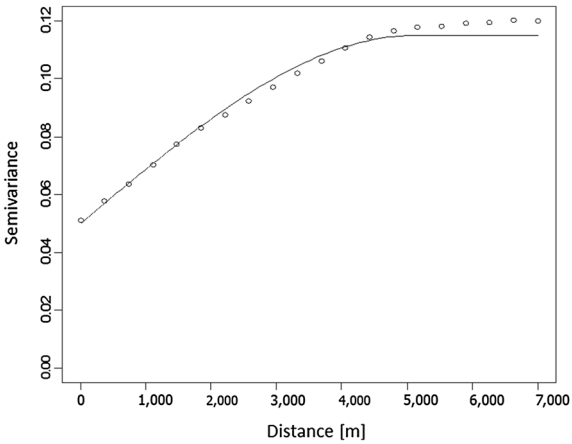

Fig. 1 Empirical (method-of-moments) and permissible semivariograms for log housing

3 Empirical application: predicting housing prices in the city of Madrid

3.1 Variables, data and the spatial weight matrix

The database used in this article was created by the authors from the sales that took place in the city of Madrid (Spain) in the first quarter of 2010. It includes the price and characteristics of 10,512 owner-occupied single-family homes sold in the city of Madrid during that period, which accounts for 90% of all real estate sales. The database is shown in the electronic supplementary material.

Table 2 presents the list of variables included in the competing models, which mirrors (and extends) the set typically used in the literature. Some of them have been codified as categorical to allow for more flexibility (nonlinearity) in the specification of the model, while others (age and area, since most of the nonlinearities that matter are in these variables) have been squared and normalized, and used as control variables in models of groups (i), (ii) and (iii) in Table 1. As can be seen in Table 2, this list includes the basic attributes of the dwellings, facilities, quality variables and areal characteristics (including a subjective environmental variable, the residents' perception of air and odor pollution problems, RP).

Moran's I and Geary's C tests for spatial autocorrelation reject the null of no spatial correlation in housing prices (in logs) ($I = 3.731\text{e}^{-01}$, p-value $< 2.2\text{e}^{-16}$; $C = 6.188\text{e}^{-01}$, p-value $< 2.2\text{e}^{-16}$). As for the range and structure of the spatial autocorrelation, the empirical semivariogram[5] shown in Fig. 1 is far from a pure nugget structure and can be perfectly fitted by a spherical isotropic model (the directional empirical semivariograms do not indicate the existence of significant anisotropies) with nugget effect equal to 0.05, partial variance equal to 0.065 and range 5108.1 m. As is well known, a spherical semivariogram is typical of phenomena exhibiting a certain degree of irregularity (see Montero et al. 2015).

Therefore, both the geostatistical and spatial econometric approaches call for the inclusion of the spatial factor in any specification that aims to predict housing prices

[5] The empirical semivariogram is used to capture the structure and range of spatial autocorrelation.

Table 2 Response variable and covariates

Variable	Description	Source
Response variable		
Log (Price, euro/m^2)	Napierian logarithm of house price (euro/m^2)	http://www.idealista.com/ A company owned by Apax Fund that provides information about both sale and rental properties in Spain.
Individual covariates		
Good condition	Indicator variable for good condition	http://www.idealista.com/ A company owned by Apax Fund that provides information about both sale and rental properties in Spain
Built-up area	Number of square meters of built-up area	
Type		
Top-floor apartment	Indicator variable for top-floor apartments	
Apartment	Indicator variable for apartments	
Studio–apartment	Indicator variable for studios	
House	Indicator variable for houses	
Age	Age of the house/apartment	
Floor		
Ground level	Indicator variable for ground level	
1st floor	Indicator variable for 1st floor	
2nd–3rd floor	Indicator variable for 2nd and 3rd floors	
4th–5th floor	Indicator variable for 4th and 5th floors	
6th floor or higher	Indicator variable for 6th floor or higher	
Baths	Number of bathrooms	
Garage	Indicator variable for parking space	
Elevator	Indicator variable for elevator	
Air conditioning	Indicator variable for central air conditioning	
Swimming pool	Indicator variable for swimming pool	

Table 2 continued

Variable	Description	Source
Areal covariates		
Environmental conditions perceived	Subjective air pollution indicator	http://www.ine.es/ The Spanish Statistics Office
Shopping area	Indicator for houses in the shopping area	http://www.madrid.org/desvan/Inicio.icm?enlace=almudena
Historical quarter	Indicator for houses in the historical quarter	A database of the Autonomous Community of Madrid whose objective is to collect and provide statistical information on the municipalities forming part of the region
Retired (%)	Percentage of retired people in the district	
Children (%)	Percentage of children under 14 years old in the district	
Immigrants (%)	Percentage of immigrants in the district	
Crime (%)	Crime rate in the district	

in the city of Madrid. Note that, though it goes beyond the scope of this article, this finding indicates that the covariates exert not only the traditional direct impacts on the response variable but can also potentially exert indirect impacts arising from the existing spatial autocorrelation (see LeSage and Pace 2009, for details).

In light of the above results, in the spatial weight matrices included in the competing models neighboring houses are those within a 1.5-km radius of one another, a distance at which spatial correlation is still high and computations are not a problem.

3.2 Estimation results

The assessment of the prediction power of the competing models included in this research has been carried out by cross-validation, and more specifically by a leave more than 2000 observations out procedure. The database has been partitioned into five sets, each containing more than 2000 observations. One-by-one, one of these sets has been left out and the prices of the houses it contains have been predicted using models estimated with the information provided by the other four sets. The root mean squared prediction error (RMSPE) has been used to measure the accuracy of these predictions (see Table 3). As stated in Sect. 3.1, when **W** matrices are involved in the specification of the model, two houses are considered neighboring if they are less than 1.5 km away from one another. Regarding the question of which houses should be considered as "neighboring," during the cross-validation process the definition of neighboring is not limited to the left-out sets; houses in the training sets can also be considered as neighbors of houses in the left-out sets.

The statistical software used in the estimation process of the 14 competing models was R (R Core Team 2015). The estimation of the HM and the spatial autoregressive

Table 3 Root mean squared prediction error from cross-validation (leave-*n*-out)

	Left-out group				
	Group 1	Group 2	Group 3	Group 4	Group 5
No. observations out	2050	2159	2072	2165	2066
HM	0.2402	0.2319	0.2309	0.2435	0.2394
SLM	0.2398	0.2324	0.2317	0.2427	0.2410
SEM	0.2177	0.2121	0.2100	0.2231	0.2216
SDM	0.2261	0.2207	0.2195	0.2300	0.2274
SDEM	0.2186	0.2133	0.2110	0.2238	0.2214
PSSD-HM	0.2154	0.2107	0.2088	0.2200	0.2186
PSSD-SLM	0.2154	0.2107	0.2088	0.2200	0.2186
PSSD-SDM	0.2149	0.2107	0.2093	0.2207	0.2186
GAM	0.2184	0.2095	0.2105	0.2177	0.2161
GAM-SLM	0.2186	0.2093	0.2109	0.2173	0.2163
GAM-SDM	0.2179	0.2064	0.2166	0.2152	0.2121
PSSD-GAM	0.2086	0.2007	0.2054	0.2086	0.2083
PSSD-GAM-SLM	0.2086	0.2007	0.2052	0.2085	0.2083
PSSD-GAM-SDM	0.2102	0.2015	0.2020	0.2095	0.2086

models included in group (ii) of Table 1 was carried out with the `spdep` library (Bivand 2013). The estimation of the rest of the models was performed with our own codes, based on the `SAP` library (Rodriguez-Alvarez et al. 2015). The codes are provided in the electronic supplementary material.

Although the focus of this article is on the out-of-sample prediction performance, the total impacts on the response variable of the continuous predictor variables that are parametrically related with it, as well as, in the case of dichotomous or polytomous variables, the difference (with respect to the reference category) in impacts on the log of house prices, are listed in Tables 4 and 5. The impacts estimated for the models corresponding to groups (i), (ii) and (iii) are listed in Table 4, and those estimated for the models included in groups (iv), (v), (vi) and (vii) are shown in Table 5. The codes for the calculation of these impacts are in the electronic supplementary material.

Figures 2 and 3 display the effects functions of the covariates that have a nonlinear relationship with the price of houses represented by a smooth nonparametric function for GAM (the simplest model containing this type of covariate) and PSSD-GAM-SDM (the most general specification), respectively. The codes for the calculation of the effects functions are in the electronic supplementary material. The results listed in Tables 4 and 5 and Figs. 3 and 4 have been obtained using the whole sample.

Nearby all the variables are significant, and the impacts and effects functions are in line with the expected results. Although the values of these impacts and effects functions are not directly comparable (impacts are constant derivatives and effects functions do not result from a derivative procedure), in general, the individual characteristics that most influence house prices are, as expected, the age and especially how built-up the area is. Among the areal characteristics, the variables

Table 4 Estimates of total impacts for models of groups (i), (ii) and (iii), see Table 1

	HM	SLM	SEM	SDM	SDEM	PSSD-HM	PSSD-SLM	PSSD-SDM
Good condition	0.0499***	0.0562***	0.0513***	−0.4034***	0.1173	0.0531***	0.0527***	0.3119***
Built-up area	−2.5023***	−2.9043***	−3.1662***	17.5683***	10.4195***	−3.2739***	−3.2473***	1.3018
Squared built-up area	2.6077***	3.0360***	3.4214***	−21.9780***	−13.9482***	3.4423***	3.4166***	−2.2990
Type (apartment)	−0.0893***	−0.0980***	−0.0726***	−0.2760***	0.3034**	−0.0677***	−0.0672***	0.4251***
Type (studio–apartment)	0.0365**	0.0360*	−0.0038	0.5531**	1.0749***	−0.0013	−0.0003	0.6583**
Type (house)	−0.0153	−0.0039	0.0421**	−1.1290***	−0.0682	0.0514***	0.0517***	0.3385*
Age	−0.5243***	−0.5952***	−0.6073***	−2.8871***	−1.3751*	−0.6474***	−0.6405***	−0.7621
Squared age	0.8555***	0.9578***	0.9092***	5.8972***	3.2095**	0.9666***	0.9576***	1.0020
Floor (1st floor)	0.0388***	0.0481***	0.0490***	−0.3542***	−0.4409**	0.0485***	0.0490***	−0.1970
Floor (2nd–3rd floor)	0.0514***	0.0626***	0.0647***	−0.3027***	−0.2813	0.0656***	0.0650***	−0.0281
Floor (4th–5th floor)	0.0534***	0.0659***	0.0701***	−0.5039***	−0.3867*	0.0706***	0.0705***	−0.0521
Floor (6th floor or higher)	0.0625***	0.0793***	0.0864***	−0.4784***	−0.2509	0.0881***	0.0884***	0.1551
Bath	0.0640***	0.0720***	0.0564***	−0.1609***	−0.0823	0.0565***	0.0560***	0.0239
Garage	0.0510***	0.0587***	0.0593***	−0.0072	0.0518	0.0633***	0.0628***	0.0530
Elevator	0.0885***	0.0955***	0.0670***	0.2542***	0.2530***	0.0646***	0.0640***	0.1157*
Air conditioning	0.0571***	0.0626***	0.0561***	0.1628***	0.1054	0.0544***	0.0537***	0.0754
Environmental conditions	−0.2092***	−0.2148***	−0.0399	−0.5527***	−0.3468**	−0.0175	−0.0185	0.0380
Swimming pool	0.0543***	0.0590***	0.0640***	−0.0916*	−0.0363	0.0653***	0.0651***	0.0784
Shopping area	0.1517***	0.1536***	0.1623***	−0.0049	0.0335	0.1500***	0.1475***	−0.1457**
Historical quarter	0.1359***	0.1427***	0.0859***	0.1668***	0.1703***	0.0997***	0.1002***	0.2876***
Retired (%)	0.0006	0.0014	0.0095***	−0.0316***	−0.0015	0.0085***	0.0084***	0.0070
Children (%)	−0.0308***	−0.0302***	−0.009***	−0.0555***	−0.0185*	−0.0070***	−0.0069***	0.0305***
Immigrants (%)	−0.0051***	−0.0050***	−0.0041***	−0.0074***	−0.0032	−0.0045***	−0.0044***	−0.0069

Table 4 continued

	HM	SLM	SEM	SDM	SDEM	PSSD-HM	PSSD-SLM	PSSD-SDM
Crime (%)	−0.1256***	−0.1276***	−0.0750**	0.1638***	0.2260*	−0.0557	−0.0546	0.3357***

Top-floor apartment and ground level are the reference categories for "Type" and "Floor," respectively

*, **, *** denote significance at the 10, 5 and 1% level, respectively

Table 5 Estimates of total impacts for models of groups (iv), (v), (vi) and (vii), see Table 1

	GAM	GAM-SLM	GAM-SDM	PSSD-GAM	PSSD-GAM-SLM	PSSD-GAM-SDM
Good condition	0.0479***	0.0504***	0.4170**	0.0511***	0.5100***	0.3338***
Type (apartment)	−0.0656***	−0.0691***	0.2700	−0.0601***	−0.0600***	0.2973**
Type (studio–apartment)	−0.0863***	−0.0915***	0.5820	−0.0900***	−0.0900***	0.6855***
Type (house)	−0.0238	−0.0213	−0.2130	0.0100	0.0091	0.3023*
Floor (1st floor)	0.0778***	0.08280***	1.1050***	0.0731***	0.0730***	−0.0157
Floor (2nd–3rd floor)	0.0937***	0.0998***	1.4910***	0.0900***	0.0900***	0.0602
Floor (4th–5th floor)	0.1026***	0.1096***	1.4800***	0.0970***	0.0960***	0.0856
Floor (6th floor or higher)	0.1229***	0.1318***	1.4710***	0.1210***	0.1191***	0.2782
Bath	0.0539***	0.0574***	0.2980***	0.0550***	0.0550***	0.0927**
Garage	0.0637***	0.0681***	0.3050**	0.0680***	0.0671***	0.1176*
Elevator	0.0893***	0.0939***	0.0620	0.0791***	0.0780***	0.0655
Air conditioning	0.0472***	0.0491***	0.0930	0.0480***	0.0470***	0.0388
Swimming pool	0.0538***	0.0561***	0.2790**	0.0650***	0.0650***	0.0441
Shopping area	−0.1883	−0.1858	−1.0777	−0.2000	−0.1900	−0.3236**
Historical quarter	0.5301*	0.5336	1.6390***	0.4940	0.4800	0.5446***

Top-floor apartment and ground level are the reference categories for “Type” and “Floor,” respectively

*, **, *** denote significance at the 10, 5 and 1% level, respectively

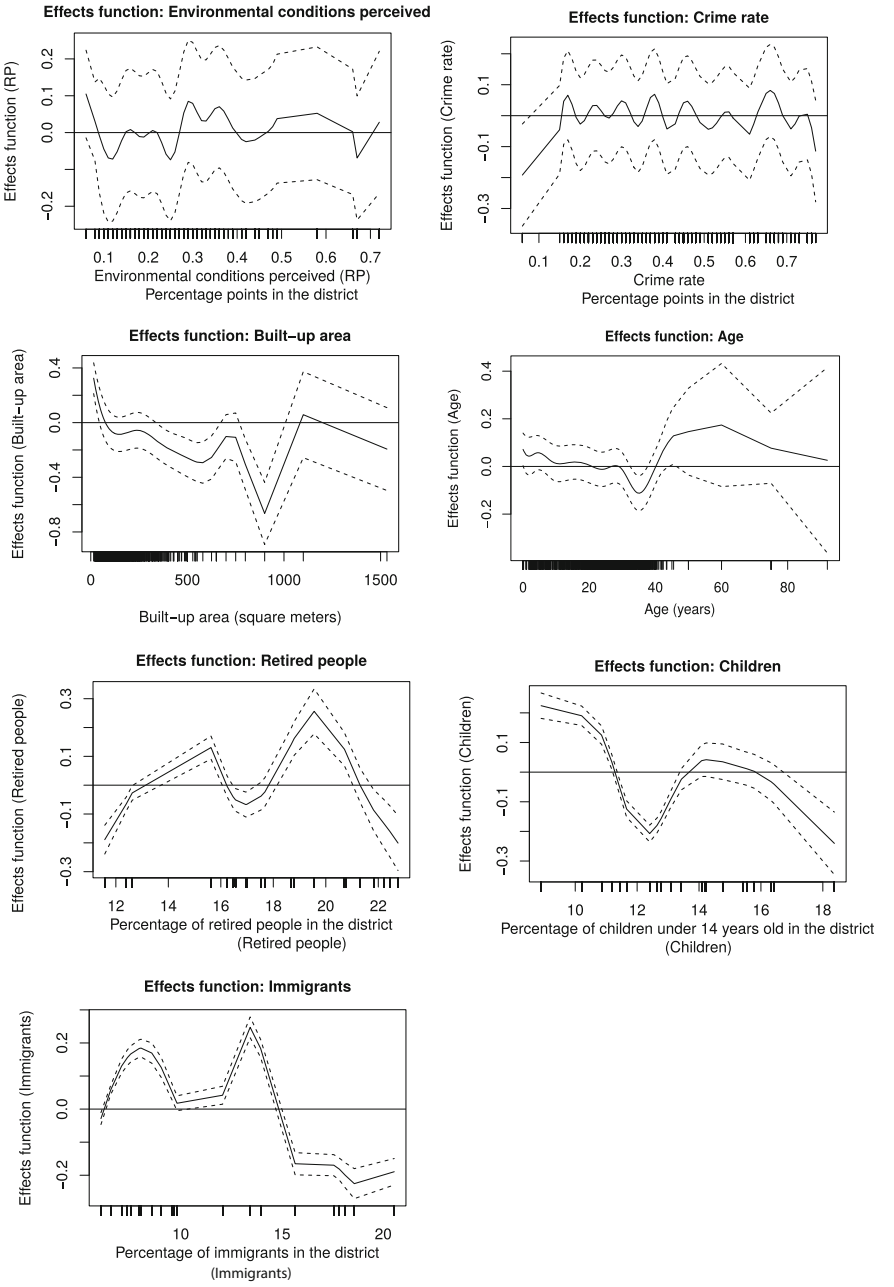

Fig. 2 Effects functions (and 0.95 confidence bands) of the covariates that have a nonlinear relationship with the price of houses. GAM

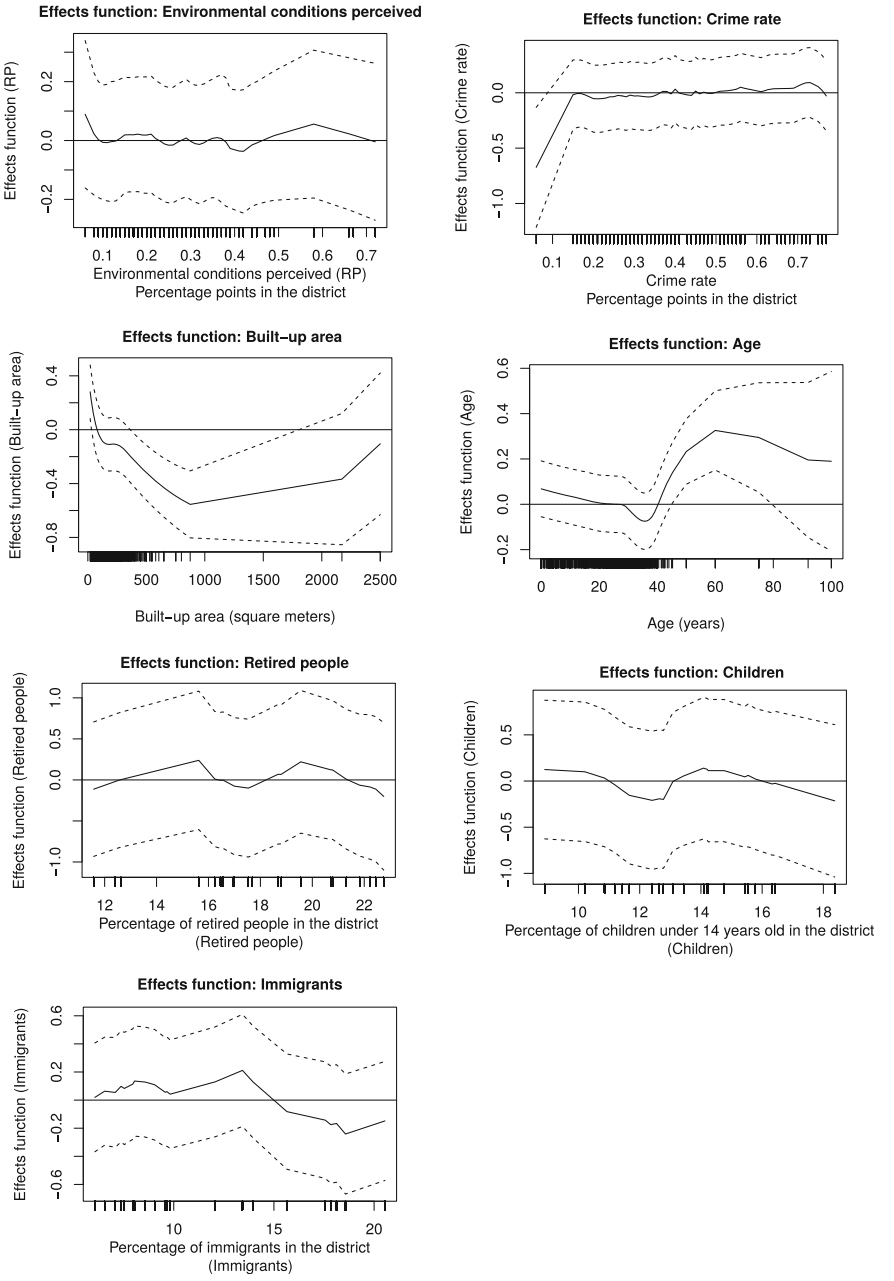

Fig. 3 Effects functions (and 0.95 confidence bands) of the covariates that have a nonlinear relationship with the price of houses. PSSD-GAM-SDM

that have the greatest influence are being located in the historical district, the percentage of children under 14 years old and, in the non-GAM-type model specifications, the environmental conditions (RP). Obviously, there are some differences between the models.

Turning now to the central objective of the paper, out-of-sample prediction, a number of findings can be drawn from Table 3 (the codes for replicating the results shown in Table 3 are in the electronic supplementary material). Within the group of spatial parametric models, the SEM produces the best predictions, reducing the RMSPE of the HM by 8.5% (on average). This is not a surprising result, given that the estimate of λ (with the complete database) is 0.920. The predictions provided by the SDEM are similar to those of the SEM, with this model reducing the RMSPE of the traditional a-spatial hedonic specification by 8.2% (also on average). In the case of the SDM specification, the equivalent average reduction is 5.2%. Finally, the performance of the SLM is similar to that of the HM, since the value of ρ is small (0.11, when estimated with the whole sample).

Including a spatial drift (more specifically a PSSD) in the a-spatial hedonic model reduces the RMSPE by 8.7–10.3% (depending on the left-out group). When the spatial drift is included in the SAR models, the RMSPE is reduced by 8.7–10.3% for the PSSD-SLM and by 8.7–10.5% for the PSSD-SDM. As can be seen, the RMSPE values of these three PSSD specifications are very similar. Therefore, as expected, when a spatial drift is included in the model its predictive performance increases, the reason being that the drift accounts for the potential spatial heterogeneity not captured by the traditional a-spatial and spatial (autoregressive) specifications. It can also be concluded from Table 5 that the inclusion of lagged (dependent and/or independent) variables once a drift has been included in the HM does not result in any prediction gains.

When some of the covariates enter into the model as smooth nonparametric functions of their values, the predictive capability of the models improves with respect to the conventional HM, and SLM and SDM specifications. However, the predictive capability is very similar to that of a PSSD combined with one of these two spatial autoregressive specifications, as well as to that of SEM and SDEM.

If the potential smooth nonparametrically specified relationship between some of the covariates and the price of houses (the response variable) is taken into account, the predictive power of the model increases. The GAM reduces the RMSPE of the HM by 8.8–10.7%, and that of the SLM and the SDM by 9.2–10.4 and 3.4–5.4%, respectively.

The extension of the GAM with any one of the autoregressive models considered in group (v) of Table 1 does not improve the prediction performance, with the RMSPE remaining practically the same as for GAM. However, if the GAM is extended with a PSSD, this produces a further improvement in prediction performance: the RMSPE of the PSSD-GAM is around 13% less than that of the HM. Finally, in accordance with these results, if a PSSD-GAM specification is extended with an SLM or SDM structure, no gains in prediction are evidenced.

From the above findings, it can be concluded that accounting for nonlinearity, via smooth nonparametric functions of some of the covariates, matters for prediction, even when a spatial drift in the form of penalized splines is already present in the model. However, the inclusion of an SLM or SDM autoregressive structure in a specification composed of a spatial drift and GAM terms does not result in additional prediction gains.

Table 6 Estimates of the spatial parameters, effective degrees of freedom, AIC and standard deviation of residuals

	ρ/λ	edf (total)	edf (long)	edf (lat)	AIC	RSD
HM	–	32	–	–	−30,276.59	0.2365
SLM	0.11***	33	–	–	−30,496.40	0.2341
SEM	0.92***	33	–	–	−32,183.64	0.2160
SDM	0.26***	64	–	–	−31,763.96	0.2200
SDEM	0.85***	64	–	–	−32,167.23	0.2159
PSSD-HM	–	160.48	64.80	60.67	−32,457.89	0.2119
PSSD-SLM	−0.09	161.16	64.69	60.48	−32,457.29	0.2119
PSSD-SDM	−0.19***	182.71	61.20	54.51	−32,524.11	0.2110
GAM	–	79.01	–	–	−32,284.2	0.2145
GAM-SLM	0.06***	79.98	–	–	32,342.23	0.2139
GAM-SDM	0.58***	151.20	–	–	−33,158.39	0.2051
PSSD-GAM	–	196.70	60.08	55.37	−33,430.43	0.2020
PSSD-GAM-SLM	−0.01	198.47	60.43	55.78	−33,431.07	0.2020
PSSD-GAM-SDM	−0.41***	269.64	61.93	56.84	−33,596.31	0.1997

ρ denotes the spatial autoregressive coefficient in SAR-type models, λ the spatial autocorrelation coefficient in SEM and SDEM, *edf* (total) the effective degrees of freedom, *edf* (long) the effective degrees of freedom corresponding to the spatial coordinate "longitude", and *edf* (lat) the effective degrees of freedom corresponding to the spatial coordinate "latitude." *AIC* is the Akaike information criterion defined as $2k - 2\log L\left(\hat{\theta}\middle/y\right)$, where k is the number of estimable parameters (degrees of freedom) and $\log L\left(\hat{\theta}\middle/y\right)$ is the log-likelihood at its maximum point of the model estimated. *RSD* is residual standard deviation defined by $(\mathrm{SSR}/(N - edf(\mathrm{total})))^{1/2}$, where SSR is the sum of the squared residuals and N is the total number of observations

*** Significance at the 0.01 level

In light of the figures listed in Table 5, we can state that the strategies that exhibit the greatest prediction power are those that account for spatial heterogeneity and smooth nonparametrically specified relationships of some of the covariates with the response variable, with the inclusion of a spatial autoregressive SLM or SDM term being irrelevant for this purpose. Accordingly, the PSSD-GAM, the PSSD-GAM-SLM and the PSSD-GAM-SDM can be considered the best options for predicting housing prices, although for the sake of simplicity the PSSD-GAM is preferred.

Finally, Table 6 lists the estimates of the spatial parameters, effective degrees of freedom (degree of nonlinearity), Akaike information criterion (AIC) and standard deviation of the residuals (goodness of fit) for the 14 competing models. Table 7 presents the main descriptive statistics of the residuals of the 14 competing models. Figures 4 and 5 in "Appendix" present the plots of the residuals versus fitted values, in order to check for nonlinearity, unequal error variances and outliers.

Referring to Table 6, it is of note that the models showing the best goodness of fit are those that include a spatial drift and GAM terms, and a spatial autoregressive component in the case of the Durbin specification. Furthermore, the following relevant observations can be made about the variability of the spatial coefficients (column 1). First, for SLM specifications, the estimates of ρ range from a nonsignificant value in

Table 7 Model diagnostics (descriptives of the residuals)

	Mean	Standard deviation	Coeff. skewness	Coeff. kurtosis
HM	0.00	0.2365	−0.04	5.41
SLM	0.00	0.2341	0.02	5.79
SEM	0.00	0.2160	−0.07	6.89
SDM	0.00	0.2200	−0.07	6.73
SDEM	0.00	0.2159	−0.06	6.86
PSSD-HM	0.00	0.2119	−0.06	7.07
PSSD-SLM	0.00	0.2119	−0.06	7.07
PSSD-SDM	0.00	0.2110	−0.05	7.04
GAM	0.00	0.2145	−0.19	7.76
GAM-SLM	0.00	0.2139	−0.20	7.83
GAM-SDM	0.00	0.2051	−0.25	8.15
PSSD-GAM	0.00	0.2020	−0.28	8.21
PSSD-GAM-SLM	0.00	0.2020	−0.28	8.21
PSSD-GAM-SDM	0.00	0.1997	−0.26	8.11

Residual standard deviation is defined as $(\mathrm{SSR}/(N - edf(\mathrm{total})))^{1/2}$, where SSR is the sum of the squared residuals, N is the total number of observations and *edf* (total) is the effective degrees of freedom

PSSD-SLM and PSSD-GAM-SLM (−0.09 and −0.01, respectively) to 0.06 in GAM-SLM and 0.11 in SLM. It is worth noting that this result is to be expected, since the spatial autocorrelation in the response variable is low (0.11) and part of this autocorrelation is captured by the spatial drift and/or the covariates exhibiting a nonlinear relationship with the response variable when they are entered into the model.

Second, the estimates of ρ show a higher variability in SDM-type models than in SLM-type specifications. This is also to be expected due to the over-parameterization inherent in the Durbin-type specifications (see the estimated *edf* of both types of specification in Table 6) and also to the fact that Durbin-type models include not only ρ but also additional spatial coefficients for the lagged covariates, which results in instability when it comes to estimating the individual parameters. In addition, there is a notable change in ρ in the Durbin specifications extended with GAM terms, as a consequence of the expression of the spatially lagged covariates in terms of B-spline bases.[6] It is of note that, in spite of their high *edf* value, the Durbin-type specifications show the best goodness of fit.

Third, the inclusion of PSSD in a hedonic specification with a spatial lag in the response variables results in a significant decrease in the estimate of ρ, especially, as expected, in the Durbin-type specification. This change is even greater when GAM terms are included as well. This finding is in line with the results obtained by Elhorst and Freret (2009) when spatial fixed effects (equivalent to PSSD) are included in a Durbin specification [ρ decreases from 0.282 (significant) to 0.083 (nonsignificant)].

[6] If the *edf* value is very large, the spatial autocorrelation parameter might even be negative (as is the case when PSSD is included in a Durbin-type specification). This finding is in line with the results presented in Tables 1 and 4 of Bivand (2012).

4 Closing remarks

The inclusion of spatial aspects in the modeling of both natural and social phenomena is becoming more popular due to recent theoretical and computational advances in the fields of spatial statistics and spatial econometrics. According to Tobler's first law of geography, everything is related to everything else, but near things are more related than distant things. The only reason for *not* accounting for spatial dependence when modeling geographic phenomena is that theory and/or computation of spatial data are not sufficiently developed. Fortunately, there is a continuously expanding literature on spatial statistics and spatial econometrics that is developing new models, methods, computation routines, etc., for the analysis of spatial data. As a consequence, more and more researchers are including spatial aspects in their analyses, significantly improving their results.

Real estate, particularly with respect to housing prices, is one of the areas where the use of spatial models is becoming very popular. There is no doubt that housing prices are spatially autocorrelated and also exhibit spatial heterogeneity, and these two facts should not be overlooked in their modeling. This is why the "hedonic house price models" have been replaced in the recent literature by extended spatial model versions, the "spatial hedonic house price models," and new models accounting for spatial autocorrelation and spatial heterogeneity have similarly emerged. However, it is not clear how best to include spatial aspects such as spatial autocorrelation and spatial heterogeneity in the traditional hedonic models, since the number of competing possibilities increases as the research on the topic progresses. In addition, these spatial hedonic models should account for potential nonlinear relationships between some of the housing attributes and housing price. This article focuses on the spatial econometric perspective and contributes to the literature on housing price prediction by assessing (via cross-validation) the out-of-sample prediction performance of 14 competing models, 13 of which account for spatial autocorrelation and/or spatial heterogeneity and/or nonlinear relationships (nonparametrically specified) between the response and some of the covariates. As far as we know, this represents the largest number of spatial models ever compared in house price research. We use a massive database composed of the price and other characteristics (19 variables) of 10,512 homes in the city of Madrid (Q1 2010), the largest database ever used in the city.

The results obtained regarding the prediction performance of the competing models indicate that the inclusion of spatially lagged terms in the traditional hedonic specification results in better predictions in the case of the SEM and SDEM. In the case of the SDM there is also a significant—though smaller—gain. The improvement with the SLM, on the other hand, is very limited. The inclusion of a drift (a PSSD term) in the SAR models results in new gains in terms of prediction accuracy, though none of the PSSD-SAR specifications considered provides significantly better predictions than the other. Obviously, the inclusion of a spatial drift in the traditional hedonic model significantly improves predictions. From these findings, it can be concluded that the inclusion of a spatial drift either in the traditional HM, or in the HM extended with an SAR term, results in non-negligible improvements in predictions, though there are no differences in the predictive performance of such models.

When entering some of the covariates into the model as smooth nonparametrically specified functions of their values, the predictive performance improves with respect to the

conventional HM and the SLM and SDM. However, the performance is very similar to that of the SEM, SDEM and the PSSD-type models of group (iii). Extending the GAM specifications with a spatial drift results in significant new prediction gains. Notwithstanding, adding spatially lagged variables to a model with smooth nonparametric functions of the covariates (with or without PSSD) does not yield substantially better predictions.

Our explanation for these findings is that most of the quantitative covariates considered in the case study exhibit a smooth nonlinear relationship with house prices. However, at the same time, some of these covariates are areal covariates and as such they account for most of the spatial effects. If these covariates were included in the model in the usual way, and not in a smooth nonparametric functional form, the drift would be the term of the model capturing nonlinearity and spatial heterogeneity. Spatial autocorrelation is captured by the SAR structures, but in our case the inclusion of an SLM or an SDM term in a strategy containing a spatial drift and GAM terms does not have a significant impact on predictions. This is because the value of the spatial autoregressive parameter of the SLM is very low, and that of the SDM is offset by the value of the spatial parameters associated with the covariates.

The results obtained are in line with the expected results, given that the spatial autocorrelation and heterogeneity in the data, as well as nonlinearity, can be captured in a hedonic model in a number of different, non-exclusive ways, for example, via (i) areal covariates or smooth nonparametric functions of these covariates, (ii) a spatial drift and (iii) spatially lagged terms (for the response variable and/or covariates and/or the disturbances). Unsurprisingly then, with hedonic specifications that include more than one form of capturing the spatial effects, it can be seen that once one such form has been included in the model, adding yet more does not result in great gains in the predictive power of the model. However, this research suggests that the inclusion of a spatial drift term in the model—whichever model it is—significantly improves its predictive power.

Finally, we can identify, among others, the following future lines of research: (i) developing the estimation of new models belonging to the PSSD family (the PSSD-SEM, the PSSD-SDEM and the PSSD-GAM-SDEM could be especially interesting); (ii) extending the list of competing models with specifications coming from research areas other than spatial econometrics, such as models coming from the field of statistical learning, which could be of particular interest; (iii) investigating whether including spatiotemporal and not just spatial effects results in gains in predictive performance; and (iv) performing additional case studies in order to establish whether the conclusions reached in this research paper for the city of Madrid can be extended to other large cities.

Acknowledgements This work was partially supported by the UCLM Grant Program to Research Groups [Group: Applied Economics and Quantitative Methods, GI20173940] and Spanish Ministry of Economy and Competitiveness grants MTM2014-52184 and ECO2015-65826-P. The authors are grateful for the use of the facilities at UCLM's Super Computational Service.

Appendix: Model diagnostics—plots of residuals versus fitted values

See Figs. 4 and 5.

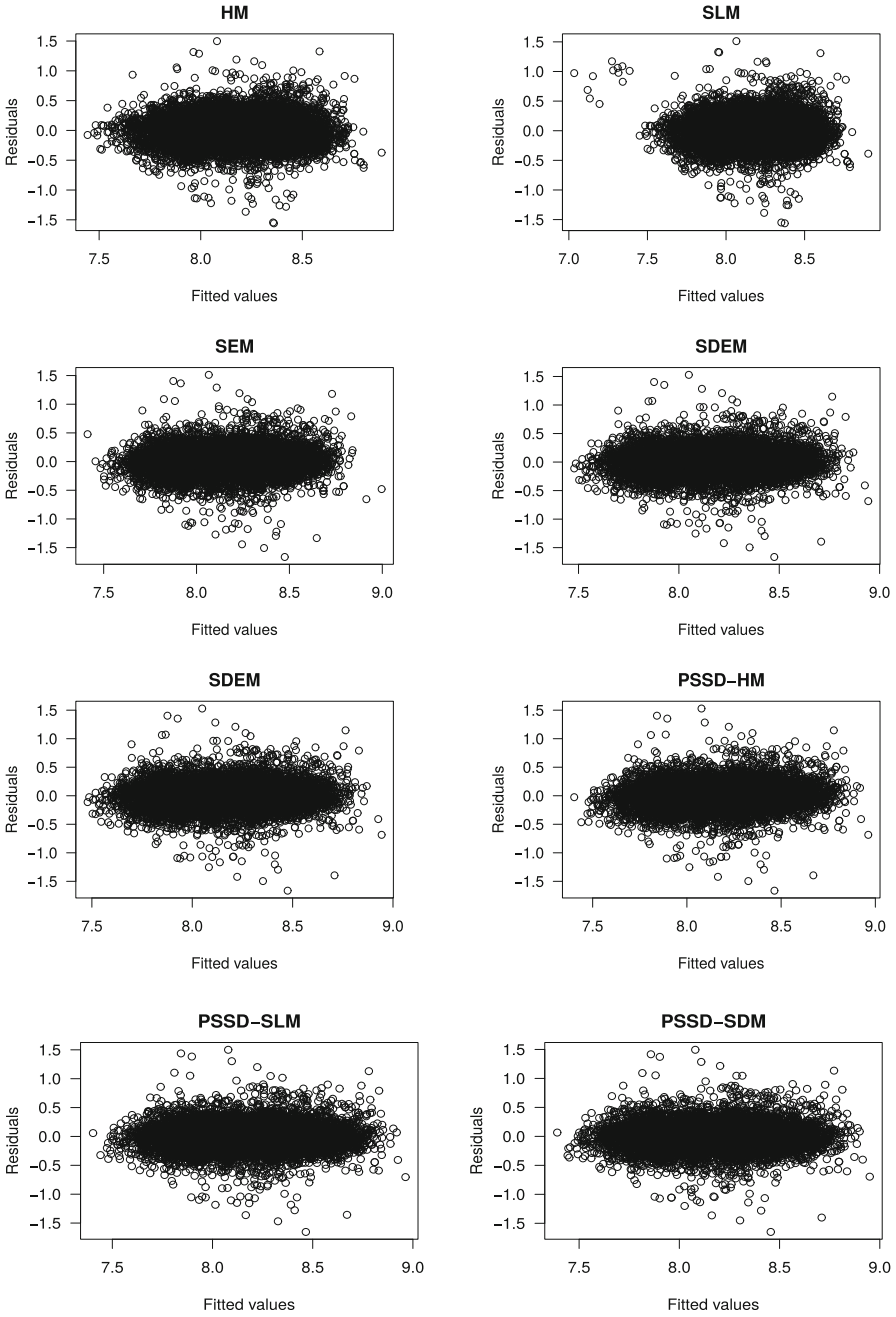

Fig. 4 Plots of residuals versus fitted values. Models of groups (i), (ii) and (iii), see Table 1

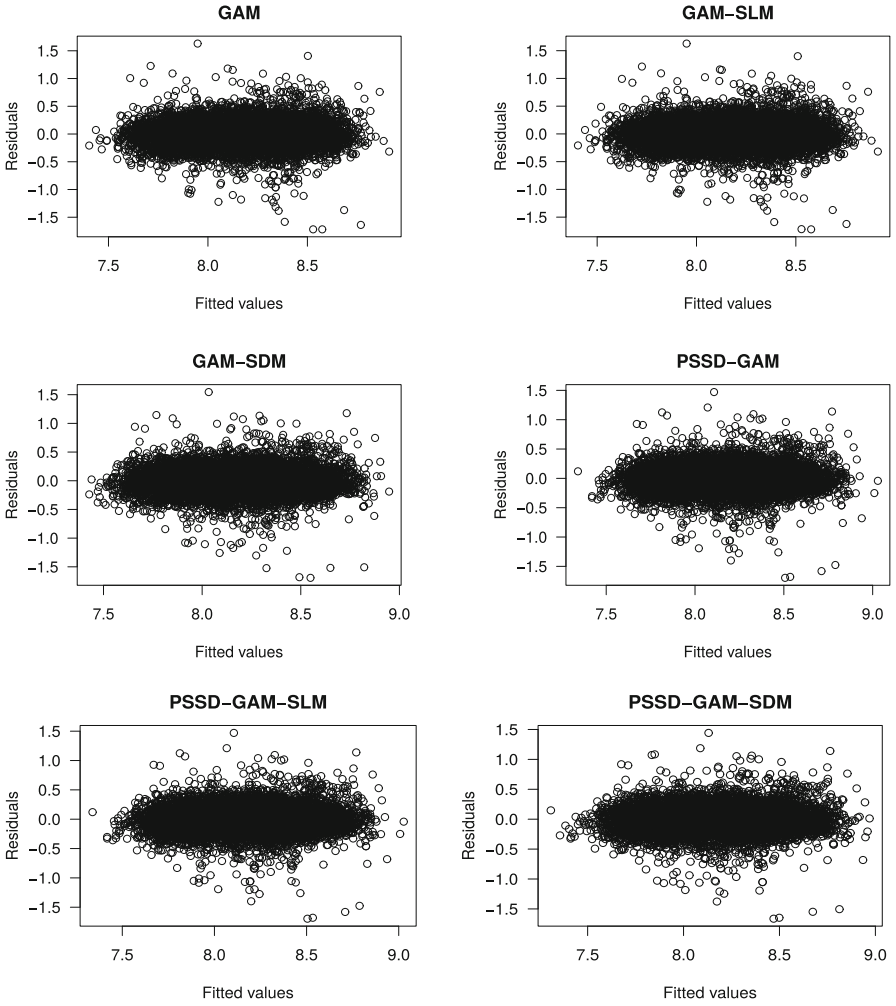

Fig. 5 Plots of residuals versus fitted values. Models of groups (iv), (v), (vi) and (vii), see Table 1

References

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Boston

Anselin L, Le Gallo J (2006) Interpolation of air quality measures in hedonic house price models: spatial aspects. Spat Econ Anal 1(1):31–52

Anselin L, Lozano-Gracia N (2008) Errors in variables and spatial effects in hedonic house price models of ambient air quality. Empir Econ 34(1):5–34

Anselin L, Syabri I, Kho Y (2004) GeoDa: an introduction to spatial data analysis. Geogr Anal 38(1):5–22

Basile R, Durbán M, Mínguez R, Montero JM, Mur J (2014) Modelling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities. J Econ Dyn Control 48:229–245

Bischoff O, Maennig W (2011) Rental housing market segmentation in Germany according to ownership. J Prop Res 28(2):133–149

Bitter C, Mulligan GF, Dall'erba S (2007) Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. J Geogr Syst 9(1):7–27
Bivand RS (2012) After "Raising the Bar": applied maximum likelihood estimation of families of models in spatial econometrics. Estadística Española 44(177):71–88
Bivand RS (2013) Spdep: Spatial dependence: Weighting schemes, statistics and models. R package version 0.5-56. http://CRAN.R-project.org/package=spdep
Bourassa SC, Hamelink F, Hoesli M, MacGregor B (1999) Defining housing submarkets. J Hous Econ 8(2):160–183
Bourassa SC, Cantoni E, Hoesli M (2010) Predicting house prices with spatial dependence: a comparison of alternative Methods. J Real Estate Res 32(2):139–159
Brasington DM, Hite D (2005) Demand for environmental quality: a spatial hedonic analysis. Reg Sci Urban Eco 35(1):57–82
Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr Anal 28(4):281–298
Cebula RJ (2009) The hedonic pricing model applied to the housing market of the city of Savannah and its Savannah historic landmark district. Rev Reg Stud 39(1):9–22
Cellmer R (2014) The possibilities and limitations of geostatistical methods in real estate market analysis. Real Estate Manag Valuat 22(3):54–62
Chica-Olmo J (1995) Spatial estimation of housing prices and locational rents. Urban Stud 32(8):1331–1344
Chica-Olmo J (2007) Prediction of housing location price by a multivariate spatial method: cokriging. J Real Estate Res 29(1):95–114
Chica-Olmo J, Cano-Guervos R, Chica-Olmo M (2013) A coregionalized model to predict housing prices. Urban Geogr 34(3):395–412
Cressie N (1993) Statistics for spatial data. Wiley, New York
Dubin RA (1998) Predicting house prices using multiple listings data. J Real Estate Financ 17(1):35–59
Eilers PHC, Marx BD (1996) Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). Stat Sci 11(2):89–121
Elhorst JP (2014) Spatial econometrics. From cross-sectional data to spatial panels. Springer, Heidelberg
Elhorst JP, Freret S (2009) Evidence of political yardstick competition in France using a two-regime spatial Durbin model with fixed effects. J Reg Sci 49(5):931–951
Fernández-Avilés G, Mínguez R, Montero JM (2012) Geostatistical air pollution indexes in spatial hedonic models: the case of Madrid, Spain. J Real Estate Res 34(2):243–274
Fletcher M, Gallimore P, Mangan J (2000) Heteroscedasticity in hedonic house price models. J Prop Res 17(2):93–108
Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, New York
Gámez M, Montero JM, García N (2000) Kriging methodology for regional economic analysis. Int Adv Econ Res 6(3):438–450
Goodman A, Thibodeau T (2003) Housing market segmentation and hedonic prediction accuracy. J Hous Econ 12(3):181–201
Harville DA (1977) Maximum likelihood approaches to variance component estimation and related problems. J Am Stat Assoc 72(358):320–338
Helbich M, Brunauer W, Hagenauer J, Leitner M (2013a) Data-driven regionalization of housing markets. Ann Assoc Am Geogr 103(4):871–889
Helbich M, Brunauer W, Vaz E, Nijkamp P (2013b) Spatial heterogeneity in hedonic house price models: The case of Austria. Tinbergen Institute Discussion Paper TI 2013-171/VIII. Tinbergen Institute, Rotterdam
Jahanshiri E, Buyong T, Shariff ARM (2011) A review of mass valuation models. Pertanika J Sci Technol 19(S):23–30
Kuntza M, Helbichab M (2014) Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. Int J Geogr Inf Sci 28(9):1904–1921
Lancaster KJ (1966) A new approach to consumer theory. J Polit Econ 74(2):132–157
Lee DJ, Durbán M (2011) P-spline ANOVA-type interaction models for spatio-temporal smoothing. Stat Model 11(1):49–69
LeSage P, Pace RK (2009) Introduction to spatial econometrics. Chapman & Hall/CRC Press, Boca Raton

Lu B, Charlton M, Fotheringam AS (2014) Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. Int J Geogr Inf Sci 28(4):660–681
Manganelli B, Pontrandolfi P, Azzato A, Murgante B (2014) Using geographically weighted regression for housing market segmentation. Int J Bus Intell Data Min 9(2):146–159
Mark J, Goldberg M (1988) Multiple regression analysis and mass assessment: a review of the issues. Apprais J 56(1):89–109
Mason C, Quigley JM (2007) Non-parametric hedonic housing prices. Hous Stud 11(3):373–385
McDonald J (1997) Fundamentals of urban economics. Prentice Hall, Upper Saddle River
Mínguez R, Montero JM, Fernández-Avilés G (2013) Measuring the impact of pollution on property prices in Madrid: objective versus subjective pollution indicators in spatial models. J Geogr Syst 15(2):169–191
Montero JM, Larraz B (2010) Estimating housing prices: a proposal with spatially correlated data. Int Adv Econ Res 16(1):39–51
Montero JM, Larraz B (2011) Interpolation methods for geographical data: housing and commercial establishment markets. J Real Estate Res 33(2):233–244
Montero JM, Larraz B, Páez A (2009) Estimating commercial property prices: an application of cokriging with housing prices as ancillary information. J Geogr Syst 11(4):407–425
Montero JM, Fernández-Avilés G, Mínguez R (2011) Spatial hedonic pricing models for testing the adequacy of acoustic areas in Madrid, Spain. Invest Reg J Reg Res 21:157–181
Montero JM, Mínguez R, Durbán M (2012) SAR models with nonparametric spatial trends. A P-spline approach. Estadística Española 54(177):89–111
Montero JM, Fernández-Avilés G, Mateu J (2015) Spatial and spatio-temporal geostatistical modeling and kriging. Wiley, Chichester
Neill HR, Hassenzahl DM, Assane DD (2007) Estimating the effect of air quality: spatial versus traditional hedonic price models. South Econ J 73:1088–1111
Osland L (2010) An application of spatial econometrics in relation to hedonic house price modeling. J Real Estate Res 32(3):289–320
Pace RK (1993) Nonparametric methods with applications to hedonic models. J Real Estate Finance Econ 7(3):185–204
Pace RK (1998) Appraisal using generalized additive models. J Real Estate Res 15(1):77–99. http://aresjournals.org/doi/pdf/10.5555/rees.15.1.m2g7602885041757
Pavlov A (2000) Space varying regression coefficients: a semi-parametric approach applied to real estate markets. Real Estate Econ 28(2):249–283
R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
Rodriguez-Alvarez MX, Lee DJ, Kneib T, Durbán M, Eilers P (2015) Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. Stat Comput 25(5):941–957
Small KA, Steimetz S (2006) Spatial hedonics and the willingness to pay for residential amenities. Working Paper 050631, University of California–Irvine, Department of Economics, Irvine
Straszheim M (1974) Hedonic estimation of housing market prices: a further comment. Rev Econ Stat 56(3):404–406
von Graevenitz K, Panduro TE (2015) An alternative to the standard spatial econometric approaches in hedonic house price models. Land Econ 91(2):386–409
Wood SN (2006a) An introduction to generalized additive models with R. CRC Press, Boca Raton
Wood SN (2006b) Low-Rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62(4):1025–1036
Wood SN, Augustin NH (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol Model 157(2–3):157–177
Yu DL (2006) Spatially varying development mechanisms in the Greater Beijing area: a geographically weighted regression investigation. Ann Reg Sci 40(1):173–190