Journal of Economic Dynamics & Control 48 (2014) 229–245

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

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

Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities

Roberto Basile [a,*], María Durbán [b], Román Mínguez [c], Jose María Montero [d], Jesús Mur [e]

[a] Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 1 – 81043 Capua, CE, Italy
[b] Department of Statistics, Carlos III University, Madrid, Spain
[c] Statistics Department, University of Castilla-La Mancha, Cuenca, Spain
[d] Statistics Department, University of Castilla-La Mancha, Toledo, Spain
[e] Department of Economic Analysis, University of Zaragoza, Zaragoza, Spain

ARTICLE INFO

Article history:
Received 27 September 2013
Received in revised form
11 May 2014
Accepted 13 June 2014
Available online 24 June 2014

JEL classification:
R11
R12
C14

Keywords:
Spatial econometrics
Nonlinearities
Semiparametric models

ABSTRACT

Spatial modeling of economic phenomena requires the adoption of complex econometric tools, which allow us to deal with important methodological issues, such as spatial dependence, spatial unobserved heterogeneity and nonlinearities. In this paper we describe some recently developed econometric approaches (i.e. Spatial Autoregressive Semiparametric Geoadditive Models), which address the three issues simultaneously. We also illustrate the relative performance of these methods with an application to the case of house prices in the Lucas County.

1. Introduction

Spatial modeling of economic phenomena (growth, unemployment, wages, location, house prices, crime rates and so on) requires the adoption of complex econometric tools which permit us to control for spatial dependence, unknown functional form and unobserved heterogeneity. The dominant paradigm in spatial econometrics is not well equipped to deal simultaneously with the three topics, which instead have been approached separately.

Spatial dependence reflects a situation where values observed at one location depend on the values of neighboring observations. That is, there are externalities known as *global and local spatial spillovers* (Anselin, 2003).[1] Contrary to what one would expect, in only a few cases spatial externalities have been formally predicted by well-defined theoretical models. Ertur and Koch (2007), for example, propose an extension of the multi-region neoclassical growth model that includes

* Corresponding author.
E-mail addresses: roberto.basile@unina2.it (R. Basile), mdurban@est-econ.uc3m.es (M. Durbán), roman.minguez@uclm.es (R. Mínguez), jose.mlorenzo@uclm.es (J. María Montero), jmur@unizar.es (J. Mur).

[1] There is a large list of models devoted to this topic (LeSage and Pace, 2009): the Spatial Lag or Spatial Autoregressive Model (SAR), the Spatial Error Model (SEM), the Spatial Durbin Model (SDM), the Spatial in X-variables Model (SLX) and a mix of the SAR and SEM (SARSAR) are the most popular.

http://dx.doi.org/10.1016/j.jedc.2014.06.011
0165-1889/

technological interdependence across regions. The reduced form of the growth equation predicted in this case is a linear SDM. Brueckner (2006) also presents several theoretical models of spatial interaction among local governments that lead directly to the SAR model for empirical implementation.

In most of the cases, instead, economic theory suggests the existence of network dependence and spatial spillovers, but it does not predict a well structured model. An example is the literature on the regional knowledge production function and on the diffusion of innovation, where spatial (knowledge) spillovers may occur through collaborative networks or other forms of spatial interactions (Autant-Bernard, 2012). These cases are characterized by uncertainty about the functional form of the model. The premise in applied literature is that a linear structure, coupled with some previous transformation of the data, offers enough flexibility to account for the problem.

However, there is growing evidence showing that this is a quite optimistic view. Strong nonlinearities have been detected in studies on regional growth (Arbia and Paelinck, 2003; Azomahou et al., 2011; Basile and Gress, 2005; Basile, 2008, 2009; Basile et al., 2012; Ertur and Gallo, 2009; Fotopoulos, 2012), urban agglomeration economies (Basile et al., 2013), urban environment (Chasco and Le Gallo, 2011), land prices (McMillen, 1996), urban sprawl (Brueckner, 2000; Brueckner et al., 2001; Irwin and Bockstael, 2007), social interaction (Lee et al., 2010) or house prices (Bourassa et al., 2010; Kim and Bhattacharya, 2009; Goodman and Thibodeau, 2003). Thus, in a typical empirical application, the *functional form is unknown* and the linear form, imposed sometimes rather arbitrarily, represents another source of mis-specification bias.

Controlling for *unobserved heterogeneity* is another fundamental challenge in empirical research, as failing to do so can introduce omitted-variable biases and preclude causal inference. To complicate the analysis, spatial dependence may simply be the consequence of (spatially correlated) omitted variables rather than being the result of spillovers. If this is the case, there are no compelling reasons for using traditional parametric models, like the SAR or SEM. As McMillen (2012) shows, a simple semiparametric model, with a smooth interaction between latitude and longitude (the so-called *Geoadditive Model*), can remove unobserved heterogeneity.

However, as mentioned above, in many cases the aim of the empirical study is to assess the impact of spillover effects (for example the global effect of a localized shock in R&D investment) rather than simply compensate for unobserved heterogeneity. In these cases we need to capture the effect of spatial spillover through the inclusion of spatial interaction terms, besides controlling for unobserved heterogeneity and functional form mis-specifications. This is a complex objective that the parametric paradigm, dominant nowadays, can hardly attain. It must be recognized that there have been attempts to develop a more general framework. This is the case of the parametric model proposed by Lambert et al. (2014), which combines spatial dependence and nonlinearity, or the case of Lotka–Volterra prey-predator model proposed by Griffith and Paelinck (2011). The literature on spatial regimes introduces heterogeneity in models with spatial dependence (Fischer and Stumpner, 2010), from which the SALE (Spatial Association Local Estimation) (Pace and LeSage, 2004) and Zoom algorithms (Mur et al., 2010) can be considered as limiting cases. According to our knowledge, few more references can be added. In fact, the history is very short.

Our impression is that there is a genuine need for more general and powerful approaches to model spatial data, and we are not alone in this position. In fact, several prominent scholars have recently called for a review of the methodological basis of the traditional spatial econometrics. McMillen (2010, 2012) points that there is a fundamental contradiction between the severity of the unknowns in the specification (functional form and spatially correlated omitted variables) and the overwhelming use of maximum likelihood methods (which heavily depend on the assumption of a correct specification). Pinkse and Slade (2010) recognize the intrinsic complexity of spatial data which suffer from so many problems (irregular spatial arrangement, varying density, aggregation, and so on) that precludes the use of naively parsimonious specifications, like the family of SAR models. The comparison with time series literature is deceiving because stationarity is a strange concept over space. Their advice can be summarized in avoiding overparameterized specifications and letting the application guide the theory; this has a clear parallel with the position of McMillen.

According to Gibbons and Overman (2012), the dominant approach in spatial econometrics is not convincing because of the many, sometimes unjustified, hypotheses made about the functional form, the presence of omitted factors, the spatial weights, and so on. These authors confer special relevance to the notions of identification and causality. Spatial models that include spatial lags of the endogenous variable, together with a set of contextual variables in the right hand side of the equation, are not identified because of the essential collinearity between the contextual variables and the output variable. This is the *'reflection problem'* posed by Manski (1993) in relation to 'peer effects' models. However, this problem was solved by Pinkse and Slade (2010) with the distinction between 'expected reaction of the individual', a relevant concept in the peer effects literature, and spillover effects, which is the adequate notion in spatial analysis (that is, the spatial lag is no longer a mere sample analogue of the expected reactions of the neighbors). Causality is a 'gold standard' in economics except in the field of spatial econometrics where, surprisingly, the concept of causality is mixed with that of correlation. However, fit well the data may mean nothing but spurious correlation or common factors. Gibbons and Overman (2012) confer some merit to the description of spatial data, but this cannot be the ultimate goal of the analysis. A further critical issue raised by Gibbons and Overman (2012) concerns the use of lagged values of the regressors as instrument variables (IV) for the spatial lag of the endogenous variable in the SAR-type models. The arguments are somewhat familiar with Pinkse and Slade (2010): the first is the unconvincing exclusion of these terms (spatial lagged regressors) from the structural equation; the second is the unjustified claim of exogeneity for the X's variables in a typical spatial model (contrary, they are expected to be endogenous and correlated with the unobserved determinants to the endogenous variable). The last deficiency can be treated more efficiently by using, once again, less structured models.

Given these limitations, it is important to pay attention to other, more flexible, approaches, which help us to overcome part of the deficiencies encountered in the parametric framework. In particular, in this paper we focus on *Spatial Autoregressive Semiparametric Geoadditive Models* developed, among others, by Basile and Gress (2005), Su and Jin (2010), Su (2012), Basile et al. (2012) and Montero et al. (2012). With respect to standard parametric spatial econometric models, these semiparametric approaches offer a more convenient way of addressing simultaneously the three problems mentioned above (substantive spatial dependence, unobserved heterogeneity and unknown functional form). The objective of this paper is to describe the main methodological contributions recently produced in this field and to discuss their potentials and limitations.

Section 2 introduces different specifications of the semiparametric model: (i) the Penalized Spline (PS) Geoadditive Model, (ii) the (PS-SAR), (iii) the (PS-SEM), (iv) the (PS-SDM), (v) the (PS-SLX) and (vi) the (PS-SARSAR). Section 3 discusses various technical aspects related to the identification and estimation of these models. Section 4 includes an application of these models to house price data. The application compares parametric and semiparametric regression estimates. The differences are clearly in favor of the more flexible semiparametric specification. Section fifth recaps and offers some practical suggestions for the application of these models.

2. Semiparametric models

In this section, we present a semiparametric framework which allows us to relax the linearity assumption and, simultaneously, model spatial dependence and unobserved heterogeneity. We start by introducing a basic semiparametric geoadditive model (Section 2.1). In Sections 2.2 and 2.3 we extend this model by introducing the spatial lag of the dependent variable on the right hand side (r.h.s.), the spatial lag of other covariates and a spatial autoregressive error term, thus obtaining the *PS-SAR*, the *PS-SDM*, the *PS-SLX*, the *PS-SEM* and the *PS-SARSAR* specifications.

2.1. Penalized spline (PS) geoadditive models

The starting point is a general form of the semiparametric geoadditive model suitable for large cross-sections of either spatial polygonal or spatial point data[2]:

$$y_i = \mathbf{x}_i^{*\prime}\boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{1i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})l_i + \cdots + h(no_i, e_i) + \varepsilon_i, \quad \varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n \tag{1}$$

where y_i is a continuous univariate output variable in location *i*. $\mathbf{x}_i^{*\prime}\boldsymbol{\beta}^*$ is the linear predictor for any strictly parametric component (including the intercept, all categorical covariates and eventually a set of continuous covariates), with $\boldsymbol{\beta}^*$ being a vector of fixed parameters. $f_k(\cdot)$ are unknown smooth functions of univariate continuous covariates or bivariate interaction surfaces of continuous covariates, capturing nonlinear effects of exogenous variables. Which of the explanatory variables enter the model parametrically or non-parametrically may depend on theoretical priors or can be suggested by the results of model specification tests (Kneib et al., 2009).

$f_4(x_{1i})l_i$ is a varying coefficient term, where l_i is either a continuous or a binary covariate. For example, we may wish to assess whether the smooth effect of x_1 (e.g., population density) is higher in metropolitan areas. In this case l_i is a binary variable taking value one if region *i* belongs to a metropolitan area and zero otherwise.

The term $h(no_i, e_i)$ in Eq. (1) is a smooth spatial trend surface, i.e. a smooth interaction between latitude (*northing*) and longitude (*easting*). It allows us to control for unobserved spatial heterogeneity, which is a primary task when dealing with spatial data.[3] When the term $h(no_i, e_i)$ is interacted with one of the explanatory variables (e.g., $h(no_i, e_i)x_{1i}$), it allows us to estimate spatially varying coefficients (like in the *GWR* model). Finally, ε_i are *iid* normally distributed random shocks.[4]

In the case of the semiparametric geoadditive model (1), if all regressors are manipulated independently of the errors, $\widehat{f}_k(x_k)$ can be interpreted as the conditional expectation of *y* given x_k (net of the effect of the other regressors). Blundell and Powell (2003) use the term Average Structural Function (ASF) with reference to this function.

Omitting the subscript *i*, each *k-th* univariate smooth term in Eq. (1) can be approximated by a linear combination of q_k known basis functions[5] $b_{q_k}(x_k)$:

$$f_k(x_k) = \sum_{q_k} \beta_{q_k} b_{q_k}(x_k)$$

with β_{q_k} unknown parameters to be estimated. To reduce mis-specification bias, $q_k's$ should be large enough, which results in a danger of over-fitting. As we shall clarify further on, the smoothness of the functions can be controlled by penalizing 'wiggly' functions when fitting the model. A measure of 'wiggliness', $J_k \equiv \boldsymbol{\beta}_k'\mathbf{S}_k\boldsymbol{\beta}_k$, is associated with each *k* smooth function, with $\mathbf{S}_k$ a positive semidefinite matrix. Typically, the quadratic penalty term is equivalent to an integral of squared second

[2] Although this model is widely used in environmental studies and in epidemiology (Augustin et al., 2009), it has been rarely considered in economics. A similar model, called 'structured additive regression ((STAR) model'), is proposed in Fahrmer et al. (2013).

[3] Especially when the researcher considers spatial unobservables as potential sources of endogeneity, that is, when there is a suspected correlation between unobserved and observed variables.

[4] This assumption can be relaxed by a more general specification, such as $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2\boldsymbol{\Lambda})$ being $\boldsymbol{\Lambda}$ a covariance matrix reflecting cross-sectional dependence in the errors as, for example, in Pinheiro and Bates (2000).

[5] Each basis function is usually represented through splines (see De Boor, (2001), for details).

derivatives of the function, for example $\int f''(x)^2\,dx$, but there are other possibilities such as the discrete penalties suggested by Eilers and Marx (1996) (see Section 3).

The penalized spline base-learners can be extended to two or more dimensions to handle interactions by using thin-plate regression splines (Wood, 2006, Section 4.1.5) or tensor products (Currie et al., 2006). In the last case, which is the approach followed in our implementation, smooth bases are built up from products of 'marginal' bases functions. For example,

$$f_3(x_3,x_4)=\sum_{q_3}\sum_{q_4}\beta_{q_3,q_4}b_{q_3}(x_3)b_{q_4}(x_4).$$

A similar representation can be given for the smooth spatial trend surface, $h(no,e)$. Corresponding wiggliness measures are derived from marginal penalties (Wood, 2006). Moreover, it is worth mentioning that, when $f(x_3,x_4)$ – or $h(no,e)$ – is represented using a tensor product, the basis for $f(x_3)+f(x_4)$ is strictly nested within the basis for $f(x_3,x_4)$. This means that we do not need to include the two marginal terms $f(x_3)$ and $f(x_4)$ into the equation, in order to test for smooth interaction effects.

In the case of a varying coefficient term like $f_4(x_1)l$, the basis functions $b_{q_4}(x_1)$ are pre-multiplied by a diagonal matrix containing the values of the interaction variable (l). Similarly, in the case of a spatially varying coefficient term like $h(no,e)x_1$, the basis functions $b_{q_{no}}(no)b_{q_e}(e)$ are pre-multiplied by a diagonal matrix containing the values of the interaction variable x_1.

Given the bases for each smooth term, Eq. (1) can be rewritten in matrix form as

$$\begin{aligned}\mathbf{y}&=\mathbf{X}^*\boldsymbol{\beta}^*+\Sigma_{q_1}\beta_{1q_1}b_{1q_1}(x_1)+\Sigma_{q_2}\beta_{2q_2}b_{2q_2}(x_2)+\cdots+\boldsymbol{\varepsilon}\\&=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\end{aligned}\tag{2}$$

where matrix $\mathbf{X}$ includes $\mathbf{X}^*$ and all the basis functions evaluated at the x's covariate values, while $\boldsymbol{\beta}$ contains $\boldsymbol{\beta}^*$ and all the coefficient vectors, β_q, corresponding to the basis functions.

2.2. PS-SAR, PS-SLX and PS-SDM

The Geoadditive model (1) represents a quite general framework to model spatial data taking into account nonlinearities and spatial heterogeneity. However, this model rules out spatial interaction effects. One step in this direction is by introducing spatial lags of the exogenous (X) variables on the r.h.s. of model (1), to capture the so-called *local spatial spillovers*. This model can be termed Penalized-Spline Spatial in X-variable model (or simply PS-SLX).[6] It is also possible to capture *global spatial spillovers* by augmenting the Geoadditive model with the spatial lag of the dependent variable. The structural form of the semiparametric model becomes a Penalized-Spline Spatial Autoregressive Geoadditive Model (PSSAR as called in Mínguez et al., 2012):

$$\begin{aligned}&y_i=\mathbf{x}_i^{*\prime}\boldsymbol{\beta}^*+\rho\sum_{j=1}^{n}w_{ij}y_j+f_1(x_{1i})+f_2(x_{2i})\quad+f_3(x_{3i},x_{4i})+f_4(x_{1i})l_i+\cdots+h(no_i,e_i)+\varepsilon_i\\&\varepsilon_i\sim iid\mathcal{N}(0,\sigma_\varepsilon^2),\quad i=1,\ldots,n\end{aligned}\tag{3}$$

where w_{ij} is the element of a spatial weights matrix $\mathbf{W}_n$, $\sum_{j=1}^{n}w_{ij}y_j$ is the spatial lag of the dependent variable (which always enters the model linearly), and ρ is the spatial spillover parameter. This model was first proposed by Gress (2004) and Basile and Gress (2005) and then reformulated by Basile (2008, 2009), Basile et al. (2012), Montero et al. (2012), Mínguez et al. (2012), Su and Jin (2010) and Su (2012). It reflects the notion of spatial dependence made of two parts: (i) a spatial trend due to unobserved regional characteristics, which is modeled by the smooth function of the coordinates, and (ii) global spatial spillover effects, which are modeled by including the spatial lag of the dependent variable. Su (2012) extends model (3) by allowing for heteroskedasticity and spatial dependence in the error term. The introduction of the spatial lags of the exogenous ($\mathbf{X}$) variables results in what may be called the Penalized-Spline Geoadditive Spatial Durbin Model (PS-SDM).

As in the parametric SAR, also in the PS-SAR the estimated coefficients for the parametric terms ($\widehat{\boldsymbol{\beta}}^*$) cannot be interpreted as marginal effects of the corresponding variables on the dependent variable, due to the autoregressive term (ρ). Direct, indirect (spillover) and total effects must be computed instead using the algorithm described in LeSage and Pace (2009). Similarly, the estimated smooth functions – $\widehat{f}_k(x_k)$ – cannot be interpreted as ASF. Taking advantage of the results obtained for parametric SAR, we can compute the total smooth effect (total-ASF) of x_k as

$$\widehat{f}_k^{\,T}(x_k)=\Sigma_q[\mathbf{I}_n-\widehat{\rho}\mathbf{W}_n]_{ij}^{-1}b_{kq}(x_k)\widehat{\beta}_{kq}\tag{4}$$

[6] It is customary to distinguish between local and global spatial spillovers (Anselin, 2003). The key is the existence of a spatial multiplier matrix in the reduced form of the model. The reduced form of the Spatial Lag Model (SAR) ($\mathbf{y}=\mathbf{A}\mathbf{X}\boldsymbol{\beta}+\mathbf{A}\boldsymbol{\varepsilon}$), for example, contains the spatial multiplier matrix $\mathbf{A}=(\mathbf{I}_n-\rho\mathbf{W}_n)^{-1}$, which implies that a change in a regressor x_k in region i – as well as a change in the error $\boldsymbol{\varepsilon}$ in region i – impacts on the outcome of this region, on the outcome of its neighbors, on that of the neighbors of its neighbors and so on. The impact therefore is global. In the case of the SEM, the global spillover effect concerns only un-modeled random shocks: $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{B}u$, with $\mathbf{B}=(\mathbf{I}_n-\lambda\mathbf{W}_n)^{-1}$. On the contrary, local spatial spillovers in the explanatory variables characterize the spatial cross-regressive model (SLX): $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{W}_n\mathbf{X}\boldsymbol{\delta}+\boldsymbol{\varepsilon}$. In fact, there is not an inverse involved in the reduced form of this model, and the impact of the change dies just after its effect on the neighbors (the structural form of the SLX is in fact the reduced form).

We can also compute direct and indirect (or spillover) effects of smooth terms in the PS-SAR case as

$$\widehat{f}_k^D(x_k) = \Sigma_q[\mathbf{I}_n - \widehat{\rho}\mathbf{W}_n]_{ii}^{-1} b_{kq}(x_k)\widehat{\beta}_{kq} \tag{5}$$

$$\widehat{f}_k^I(x_k) = \widehat{f}_k^T(x_k) - \widehat{f}_k^D(x_k) \tag{6}$$

Similar expressions can be provided for the direct, indirect and total effects of the PS-SDM.

2.3. Penalized spline spatial error models (PS-SEM) and PS-SARSAR

The Spatial Error Geoadditive Model (PS-SEM) proposed by Mínguez et al. (2012) augments the Penalized Spline Geoadditive Model by including a spatial autoregressive error term, while leaving the systematic part of the model unchanged:

$$\begin{aligned} y_i &= \mathbf{x}_i^{*\prime}\boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{2i}) \quad + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})l_i + \cdots + h(no_i, e_i) + u_i \\ u_i &= \lambda \sum_{j=1}^{n} w_{ij}u_j + \varepsilon_i, \quad \varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2), \ \ i = 1, \ldots, n \end{aligned} \tag{7}$$

where λ is a spatial autoregressive parameter. As in the case of the pure PS model (1), if all regressors are exogenous, $\widehat{f}_k(x_k) = \Sigma_q b_{kq}(x_k)\widehat{\beta}_{kq}$ can be directly interpreted as the conditional expectation of y given x_k (ASF).

The PS-SEM allows us to capture spatial externalities in the un-modeled idiosyncratic random shocks. The reduced form of the model is:

$$y = \mathbf{X}^{*\prime}\boldsymbol{\beta}^* + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4) + f_4(x_1)l + \cdots + h(no, e) + (I_n - \lambda W_n)^{-1}\varepsilon$$

Finally, we may consider a Penalized Spline Geoadditive Model which includes both a spatial lag of the dependent variable and a spatial autorregresive error term (PS-SARSAR):

$$\begin{aligned} y_i &= \mathbf{x}_i^{*\prime}\boldsymbol{\beta}^* + \rho \sum_{j=1}^{n} w_{ij}y_j + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})l_i + \cdots + h(no_i, e_i) + u_i \\ u_i &= \lambda \sum_{j=1}^{n} w_{ij}u_j + \varepsilon_i, \quad \varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2), \ \ i = 1, \ldots, n \end{aligned} \tag{8}$$

This model captures global spatial spillovers in the same way as the PS-SAR model.

3. Estimation methods

Let us now discuss the issues concerning estimation and inference in the semiparametric models described above. We begin with model (1). As it is well known, there are two alternative estimators for this case. The first one is the penalized least squares (PLS) method, coupled with a generalized cross validation (GCV) score minimization process to select the smoothing parameters. Alternatively, the semiparametric model (1) can be expressed as a mixed model and, thus, it is possible to estimate all the parameters (including the smoothing parameters) using restricted maximum likelihood methods (REML). Here, we focus on the second method which, in spite of being less popular among practitioners, it appears to be clearly superior (Wood, 2011). Thus, in the next section we present different procedures to deal with general semiparametric models using mixed models. Section 3.2 shows how this methodology can be extended to estimate the parameters of PS-SAR and PS-SEM models in a single step. Finally, in Section 3.3 we present an alternative two-step control function approach to estimate the PS-SAR model.

3.1. Penalized regression splines as mixed models and the REML estimator

The estimation of model (2) can be based on its reparameterization as a mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\varepsilon}, \quad \mathbf{U} \sim \text{i.i.d. } N(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\varepsilon} \sim \text{i.i.d. } N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}) \tag{9}$$

where, again, matrix $\mathbf{X}$ may include parametric components such as the intercept, continuous covariates and categorical covariates, while matrix $\mathbf{Z}$ includes all the nonlinear components of the smooth effects. This is a mixed model where $\boldsymbol{\beta}$ represents the parameters of the fixed part of the equation and $\mathbf{U}$ are the random effects. $\mathbf{G}$ is the covariance matrix of these effects; in our case it is a block-diagonal matrix, which depends on both $\sigma_{u_k}^2$ and σ_ε^2 variances. The smoothing parameters, that control the fit versus smoothness trade-off, are defined by the ratios $\theta_k = \sigma_\varepsilon^2/\sigma_{u_k}^2$.

The reparameterization consists in post-multiplying $\mathbf{X}$ and pre-multiplying $\boldsymbol{\beta}$ in model (2) by an orthogonal matrix resulting from the singular value decomposition of the penalty matrices $\mathbf{S}_k$ (Wand, 2003; Lee and Durbán, 2011; Wood et al., 2012). Therefore, the type of penalizations determines the transformation matrix and, thus, the resulting fixed and random effects. The coefficients associated with the fixed effects ($\boldsymbol{\beta}$) are not penalized, while those associated with the random effects ($\mathbf{U}$) are penalized. The penalization of random effects is given by the variance–covariance matrix of these coefficients.

It is worth pointing out that when the model is a pure additive model $\mathbf{y}=\sum_{k=1}^{K} f(x_k)+\boldsymbol{\varepsilon}$ (i.e. there are no interaction terms), $\mathbf{G}$ is block-diagonal, each block matrix $\mathbf{G}_k$ depending only on θ_k, the smoothing coefficient associated to each variable x_k. Thus, model (9) becomes a variance components model that can be estimated by using standard software. When the model contains interaction terms, it is no longer a pure additive model. In this case, each block $\mathbf{G}_k$ depends on more than one smoothing coefficient θ_k, except in the isotropic case,[7] where coefficients θ_k are the same for all variables (Wood et al., 2012; Lee and Durbán, 2011). As a consequence, the resulting mixed model is not an orthogonal variance component model.

A reparameterization exposed in Lee and Durbán (2011), from a P-Spline approach with a B-Spline basis and penalization matrices for the basis coefficients based on discrete differences, allows us to express a semiparametric model, including additive and interaction effects, as a mixed model with orthogonal variance components. In this way, different degrees of smoothing for interacting variables can be allowed. A recent alternative reparameterization, using only one smoothing coefficient for each term, is proposed by Wood et al. (2012). Two other interesting reparameterizations are based on (i) a truncated polynomial basis and ridge penalizations (Ruppert et al., 2003), and (ii) on a thin-plate regression splines basis and penalizations based on the integral of the second derivatives of the spline functions (Wood, 2003).

Once the mixed model is defined, the parameters associated to fixed ($\boldsymbol{\beta}$) and random effects (θ_k and σ_ε^2) can be estimated by using a ML algorithm. If the random term follows a Gaussian distribution, the log-likelihood function is given by

$$\log L(\boldsymbol{\beta},\theta_1,\ldots,\theta_K,\sigma_\varepsilon^2)=\text{constant}-\frac{1}{2}\log|\mathbf{V}|-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

where $\mathbf{V}=\mathbf{Z}\mathbf{G}\mathbf{Z}'+\sigma_\varepsilon^2\mathbf{I}$; the smoothing parameters θ_k are included in $\mathbf{V}$.

The ML estimates are biased since this method does not take into account the reduction in the degrees of freedom due to the estimation of the fixed effects. The restricted maximum likelihood (REML) method can be used to solve the problem. The REML method looks for the linear combinations of the dependent variable that eliminates the fixed effects from the equation (McCulloch et al., 2008). In this case the objective function to maximize is given by

$$\log L_R(\theta_1,\ldots,\theta_K,\sigma_\varepsilon^2)=\text{constant}-\frac{1}{2}\log|\mathbf{V}|-\frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|-\frac{1}{2}\mathbf{y}'(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}$$

An estimation of the variance components can be obtained after maximizing $\log L_R(\cdot)$. In the second step, the estimates of $\boldsymbol{\beta}$ and $\mathbf{U}$ are given by (McCulloch et al., 2008):

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$
$$\hat{\mathbf{U}}=\hat{\mathbf{G}}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

Finally, the estimated values of the observed variable are obtained as

$$\hat{\mathbf{y}}=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{U}}$$

To build confidence intervals for the estimated values, an approximation of the variance–covariance matrix of the estimation error is given by $V(\mathbf{y}-\hat{\mathbf{y}})=\sigma_\varepsilon^2\mathbf{H}$ where $\mathbf{H}$ is the hat matrix of the model (Ruppert et al., 2003). For the mixed model, it can be shown that

$$\mathbf{H}=\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\end{pmatrix}$$

Recently, Wood (2011) has proposed a Laplace procedure to obtain an approximated REML or ML for any generalized linear model, which is suitable for efficient direct optimization. Simulation results indicate that these novel REML and ML procedures offer, in most cases, significant gains (in terms of mean-square error) with respect to GCV or AIC methods.

3.2. Estimation of the PS-SAR and PS-SEM: extending the REML approach

In a mixed-model form, the PS-SAR can be expressed as

$$\mathbf{y}=\rho\mathbf{W}_n\mathbf{y}+\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{U}+\boldsymbol{\varepsilon}\quad \mathbf{U}\sim\text{i.i.d. } N(\mathbf{0},\mathbf{G})\quad \boldsymbol{\varepsilon}\sim\text{i.i.d. } N(\mathbf{0},\sigma_\varepsilon^2\mathbf{I})$$

Its reduced form is:

$$\mathbf{y}=\mathbf{A}\mathbf{X}\boldsymbol{\beta}+\mathbf{A}\mathbf{Z}\mathbf{U}+\mathbf{A}\boldsymbol{\varepsilon} \tag{10}$$

where $\mathbf{A}=(\mathbf{I}-\rho\mathbf{W}_n)^{-1}$.

[7] Isotropy in this context means that the degree of smoothness is the same for all the covariates. Anisotropy is the most common situation since the covariates are usually expressed in different units or, in the case of equal measurement units (e.g. spatial location variables), the variability of such covariates differs greatly.

As pointed out in Montero et al. (2012) and Mínguez et al. (2012), the log-REML function for model (10) is

$$\log L_R(\rho,\theta_1,\ldots,\theta_K,\sigma_\varepsilon^2) = \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \log|\mathbf{A}| - \frac{1}{2}\mathbf{y}'\mathbf{A}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{A}\mathbf{y} \quad (11)$$

As usual, $\log L_R(\cdot)$ is maximized with respect to the parameter vector $(\rho,\theta_1,\ldots,\theta_K,\sigma_\varepsilon^2)'$. Note that the maximization process requires the computation of the log-determinant of matrix $\mathbf{A}$, a dense $n \times n$ inverse matrix, which depends on ρ; this is a challenging task that can be alleviated, when n is large, by using the Monte Carlo procedures described in LeSage and Pace (2009).

The estimation of the PS-SEM model can be solved also in the context of expression (11), with $\mathbf{A} = \mathbf{I}$ and a complicated covariance matrix:

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_\varepsilon^2(\mathbf{B}\mathbf{B}')$$

being $\mathbf{B} = (I - \lambda\mathbf{W}_n)$. In this case, we need to invert the matrix $\mathbf{V}$.

Finally, fixed and random effects can be estimated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{A}}\mathbf{y}$$
$$\hat{\mathbf{U}} = \hat{\mathbf{G}}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\hat{\mathbf{A}}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Unlike model (9), the PS-SAR and PS-SEM models cannot be estimated by using standard software. Nevertheless, Montero et al. (2012) and Mínguez et al. (2012) have developed some *R* codes which are available upon request.

The same methodology can be used to estimate the PS-SARSAR model. Nevertheless, as for the linear econometric SARSAR, there are some difficulties in the identification of both spatial parameters (ρ and λ) and, thus, there are often problems of numerical instability.

3.3. Estimation of the PS-SAR: a control function approach

In the PS-SAR model, the spatial lag term $\mathbf{W}_n\mathbf{y}$ and the error term $\boldsymbol{\varepsilon}$ are correlated. In Section 3.2 we have described a possible solution to this endogeneity bias based on a 1-step REML approach. As suggested by Basile (2009), an alternative way of dealing with the simultaneity bias in PS-SAR is the 2-step "control function" (CF) approach (Blundell and Powell, 2003).[8]

The CF approach is an alternative to standard instrumental variable (IV) methods (either two-stage-least squares – 2SLS or GMM). It is a two-step procedure: in the first step the endogenous explanatory variables ($\mathbf{X}$) are regressed on a set of instrumental variables ($\mathbf{Q}$); the residuals from the first step are then included in the original equation to "control" for the endogeneity bias. In linear models ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$), the CF approach relies on the same identification (orthogonality) conditions – i.e. unconditional moment restriction $E(\mathbf{Q}'u) = 0$ – as the IV methods and leads to the usual 2SLS estimator. The CF approach treats endogeneity as an omitted variable problem, where the inclusion of the first-stage residuals $\mathbf{v}$ (the part of the regressors $\mathbf{X}$ that is correlated with $\mathbf{Q}$) as a covariate corrects the inconsistency of least-squares regression of $\mathbf{y}$ on $\mathbf{X}$.

In the case of nonparametric and semiparametric additive models, the CF approach imposes extra identification assumptions – i.e. conditional mean restrictions $E(\mathbf{u}|\mathbf{Q}) = 0$ and $E(\mathbf{u}|\mathbf{X},\mathbf{Q}) = E(\mathbf{u}|\mathbf{X},\mathbf{v}) = E(\mathbf{u}|\mathbf{v})$ – not imposed in a standard IV approach. However, in the case of nonparametric additive models, the CF approach offers a critical advantage over the IV method (Wooldridge, 2010). In particular, the application of the standard 2-SLS method to nonparametric additive models (i.e. the substitution of the fitted values from the first-stage nonparametric regression of $\mathbf{X}$ on $\mathbf{Q}$ into nonlinear structural functions) generally yields inconsistent estimates of the structural parameters. Instead, alternative procedure involving the use of the residuals $\mathbf{v}$ from the first-stage regression to control for the endogeneity of the regressors $\mathbf{X}$ do yield identification of the ASF (Blundell and Powell, 2003).

Using the CF approach to estimate the PS-SAR model implies to run the following first-step semiparametric regression

$$\mathbf{W}_n\mathbf{y} = \beta_0 + \sum_m g_m(\mathbf{Q}) + \boldsymbol{\nu}$$

where $\boldsymbol{\nu}$ is a random vector satisfying conditional mean restrictions $E(\boldsymbol{\nu}|\mathbf{Q}) = 0$ and $\mathbf{Q}$ is a set of m conformable instruments. In line with Kelejian and Prucha (1997), $\mathbf{Q}$ may contain all exogenous terms included in the model and some of their spatial lags. The functions g_m define generic representations of different types of covariate effects, including both linear and nonparametric smooth components.

[8] A semiparametric spatial lag model has also been proposed within a partial linear framework. Su and Jin (2010) develop a profile quasi-maximum likelihood estimator for the partially linear spatial autoregressive model which combines the spatial autoregressive model and the nonparametric (local polynomial) regression model. Furthermore, Su (2012) proposes a semiparametric GMM estimator of the SAR model under weak moment conditions which allows for both heteroskedasticity and spatial dependence in the error terms.

The residuals from the first step are then included in the original PS-SAR equation to control for the endogeneity of $\mathbf{W}_n\mathbf{y}$[9]:

$$\mathbf{y} = \mathbf{X}^{*'}\boldsymbol{\beta}^* + \rho\mathbf{W}_n\mathbf{y} + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4) + f_4(x_1)l + \cdots + h(no, e) + c(\widehat{\nu}) + \boldsymbol{\varepsilon} \tag{12}$$

Obviously, the endogeneity of any other continuously distributed regressor in the PS-SAR model can also be addressed via the CF approach if valid instruments are available.[10] Since the second-step regression contains generated regressors (i.e. the first-step residuals), a bootstrap procedure is recommended to compute the *p*-values. Following Fiaschi et al. (2013), this procedure may consist of the following steps:

1. Obtain a bootstrap sample ($\mathbf{y}_b^*$, $\mathbf{X}_b^*$, $\mathbf{Q}_b^*$) drawn with replacement from ($\mathbf{y}$, $\mathbf{X}$, $\mathbf{Q}$).
2. Run a semiparametric regression of the bootstrapped endogenous variable on the bootstrapped exogenous variables and the instruments.
3. Insert the first-step bootstrapped residuals in the original semiparametric regression.
4. Repeat $B=1000$ times points 1–3.
5. For each estimated parametric coefficient, compute the corresponding equal-tail bootstrap *p*-value:

$$P^*(\hat{\beta}) = 2 \times \min\left(\frac{1}{B}\sum_{b=1}^{B} \#\{\hat{\beta}_b^* \le 0\}, \frac{1}{B}\sum_{b=1}^{B} \#\{\hat{\beta}_b^* > 0\}\right)$$

   where $\#(\cdot)$ is the indicator function with a value of 1 if the argument is true.
6. For each estimated nonparametric function, compute the average partial effect at the 95% confidence bands.

4. An application to lucas county house pricing data

We investigate the performance of the *semiparametric spatial autoregressive models* (PS-SAR, PS-SEM, PS-SDM and PS-SLX) described above using the Lucas County (Ohio) database on house prices. In Section 4.1 we describe the dataset and briefly discuss some issues related to modeling housing prices. In Section 4.2 we report the results of the analysis.

4.1. Data and model specification issues

Lucas County (Ohio) database on housing prices contains 18,378 observations of single family houses sold during 1995–1998, and is fully described in the Spatial Econometrics toolbox for *Matlab*™ (*data/house.txt*). It has been widely used for different purposes. LeSage and Pace (2009) adopted it to illustrate the Bayesian version of the Matrix Exponential Spatial model (MESS). Bivand (2010, 2012) used it to compare functions for fitting spatial econometric models in the `R spdep` package with those in the Spatial Econometrics toolbox for *Matlab*™, in *OpenGeoDa* and in the *STATA*™ *ado* file *sppack*. Zhu et al. (2011) used the dataset to illustrate a new methodology developed to capture anisotropic spatial autocorrelation in the context of the simultaneous autoregressive model. Finally, Dubé and Legros (2013) used Lucas County data to propose a simple way to take into account the unidirectional temporal effect and the multi-directional spatial effect in the estimation process.

In all these applications, hedonic equations for single-family houses are estimated using parametric regression models relating the logarithm of the transaction price (the dependent variable) to the property's characteristics, such as the dwelling age, its squared term (and sometimes its cubic term), the logarithms of the lot size and of the total living area in square feet, and the number of rooms, bathrooms and bedrooms. Unfortunately, the data set does not contain information on various neighborhood amenities and proximity variables. The list of neighbors provided with the data set in `spdep` is a sphere-of-influence (*soi*) graph constructed from a triangulation of the point coordinates of the houses after projection to the Ohio North NAD83 (HARN) Lambert Conformal Conical specification (EPSG:2834). The resulting spatial weights matrix is relatively sparse, with less than three neighbors per observation on average.[11]

Let us note that the housing market is an adequate case for our purposes because of the simultaneous occurrence of spatial spillovers, unobserved (spatially autocorrelated) heterogeneity and nonlinearities. Empirical evidence regarding *spatial externalities* – or *adjacency effects*, as called by Can (1992) – in housing price formation is quite strong. One reason is that, due to uncertainty, real estate agents (buyers and/or sellers) use prices in the neighborhood as reference prices. Thus, the price of one house influences the prices of other houses located nearby and vice-versa. Spatial dependence may also arise because of the so-called "maintenance/repair" effect (Can and Megbolugbe, 1997), according to which the decision of one agent in relation to a variable (i.e., maintenance) affects the utility of this agent as well as the utility of neighboring

[9] Both first and second step equations can be estimated by using the REML estimator.

[10] The requirement that the endogenous regressor be continuously distributed is the most important limitation for the applicability of the CF approach in this context.

[11] We have also computed other spatial weights matrices, namely a binary k-nearest-neighbor matrix (with k=10), a binary distance-based matrix (using as threshold the minimum distance needed to make sure that all houses are linked to at least one neighbor), an inverse-distance matrix and an exponential inverse-distance matrix. The results obtained (available upon request) are very robust to the choice of the weights matrix.

Table 1
Model comparison.

Model	*AIC*				*BIC*			
	1995	1996	1997	1998	1995	1996	1997	1998
Parametric models								
A-spatial	5.170	5.455	5.595	5.329	5.181	5.465	5.604	5.339
SLX	5.033	5.345	5.482	5.212	5.052	5.362	5.498	5.230
SDM	4.784	5.116	5.233	5.019	4.803	5.132	5.250	5.038
SEM	4.916	5.206	5.326	5.126	4.926	5.215	5.335	5.136
SAR	4.831	5.168	5.295	5.067	4.841	5.177	5.304	5.077
Semiparametric models								
A-spatial PS	5.054	5.341	5.462	5.205	5.083	5.368	5.485	5.233
PS-Geoadditive	4.941	5.235	5.322	5.105	5.014	5.301	5.393	5.178
PS-SLX	4.842	5.160	5.250	5.016	4.928	5.240	5.334	5.104
PS-SDM	4.692	5.010	5.010	4.896	4.793	5.108	5.108	4.997
PS-SEM	4.771	5.064	5.064	4.975	4.868	5.154	5.154	5.070
PS-SAR (1-step)	4.708	5.030	5.030	4.919	4.799	5.114	5.114	5.001
PS-SAR (2-step)	4.708	5.025	5.108	4.927	4.784	5.094	5.179	5.000
No. of obs.	3510	4112	4276	3721	3510	4112	4276	3721

Notes: the parametric a-spatial and SLX models are estimated by Ordinary Least Squares (*OLS*). Parametric spatial regression models (*SAR*, *SEM*, *SDM*) are estimated through Maximum Likelihood (*ML*). All semiparametric and geoadditive models are estimated by Restricted Maximum Likelihood (*REML*). A Control Function (CF) approach is applied for the 2-step SAR, using the REML estimator for the estimation of the smoothing parameters in each step. The number of knots used for the smooth terms $f_1(age)$, $f_2(\log(lotsize))$, $f_3(\log(livingarea))$ and $h(no, e)$ are always 8, 10, 10 and 8, respectively.

agents. Furthermore, information flows and expectations are likely to reinforce horizontal transmissions between agents which, in addition to commuting and migration, favor the appearance of strong dependence (Brady, 2011; Holly et al., 2011; Kuethe and Pede, 2011). Following Can and Megbolugbe (1997), these kinds of spatial spillover effects can be captured only by the spatial lag of the dependent variable ($\mathbf{W}_n\mathbf{y}$).

A second type of mis-specification in house price models comes from the *unobserved effect* of local amenities. These *neighborhood effects* (Can, 1992) are the array of location characteristics (neighbors, accessibility, public service provision) that lead to different household housing demand for certain locations. These unobserved effects also generate spatial dependence between the house prices. In our approach, these effects are mostly captured by the spatial trend $h(no, e)$. Both adjacency effects and neighborhood effects are capitalized into housing prices directly as a "premium".

Finally, it must be recognized that the nature of the relationship between house prices and its attributes is complex and nonlinear, so it would be better represented by semiparametric models (Ekeland et al., 2004). For example, Goodman and Thibodeau (1995) suggest that housing depreciation (the relationship between dwelling age and the market value of owner-occupied housing) is nonlinear and possibly non-monotonic. The three issues (spatial dependence, unobserved heterogeneity and nonlinearities) clearly raise the need to modeling housing prices by using flexible PS specifications.

4.2. Model selection and econometric results

We use the Lucas County housing price data divided by year (starting from 1995) to compare the performance of different competing parametric and semiparametric models.[12] The most restricted specification is the simple *parametric* model without spatial effects (we call it the *a-spatial* model), relating the logarithm of the house price to the age of the house, its squared term, the log of the lot size and of the total living area in square feet, and the number of bathrooms. This model does not contain a spatial trend and is based on the assumption of spatial independence of the residuals. The other parametric models are the *SAR*, the *SEM*, the *SDM* and the *SLX*.[13] The first three models are estimated through *ML*, while the *SLX* is estimated by *OLS* (see the Appendix for a complete list of the different equations used in this study). The introduction of the spatial lag of the dependent variable or of the spatial error term leads to a dramatic reduction of Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics. For example comparing the performance of OLS and SAR parametric models, we obtain a reduction of the BIC value between 4 and 5% (Table 1).

[12] After a first inspection of the residuals of the simple a-spatial parametric model, we have trimmed 10% from the left-hand tail and 5% from the right-hand tail of the distribution of house prices, in order to reduce the effect of extreme values.

[13] We have also estimated the *SARSAR* model both in a parametric and a semiparametric framework, using two different spatial weights matrices (namely the sphere-of-influence, *soi*, matrix and the 10-nearest-neighbor, *10-nn*, matrix) for the $\mathbf{W}_n\mathbf{y}$ term and the error term. However, the results of these models turned out to be very sensitive to the selection of the $\mathbf{W}_n$ matrix. For example, using the *soi* matrix for computing the $\mathbf{W}_n\mathbf{y}$ term and the *10-nn* matrix for the error term, we obtained a ρ parameter of about 0.2 and a λ parameter of about 0.5. Changing the order of the two matrices, we obtained a ρ parameter of about 0.5 and a λ parameter of about 0.2.

Table 2
Likelihood ratio tests for the smooth terms of the a-spatial PS model.

Smooth terms	Restricted form	Deviance	*p-value*
1995			
$f_1(age)$	$age+age^2$	11.453	0.000
$f_2(\log(lotsize))$	$\log(lotsize)$	6.515	0.000
$f_3(\log(livingarea))$	$\log(livingarea)$	1.621	0.000
1996			
$f_1(age)$	$age+age^2$	17.881	0.000
$f_2(\log(lotsize))$	$\log(lotsize)$	9.285	0.000
$f_3(\log(livingarea))$	$\log(livingarea)$	1.547	0.000
1997			
$f_1(age)$	$age+age^2$	24.889	0.000
$f_2(\log(lotsize))$	$\log(lotsize)$	11.030	0.000
$f_3(\log(livingarea))$	$\log(livingarea)$	0.721	0.001
1998			
$f_1(age)$	$age+age^2$	17.497	0.000
$f_2(\log(lotsize))$	$\log(lotsize)$	6.875	0.000
$f_3(\log(livingarea))$	$\log(livingarea)$	1.641	0.000

Table 3
Estimates of ρ and λ parameters.

Model	ρ				λ			
	1995	1996	1997	1998	1995	1996	1997	1998
Parametric models								
SDM	0.396	0.387	0.398	0.358				
SEM					0.444	0.437	0.448	0.403
SAR	0.334	0.307	0.320	0.294				
Semiparametric models								
PS-SDM	0.287	0.288	0.288	0.263				
PS-SEM					0.319	0.316	0.316	0.275
PS-SAR (1-step)	0.260	0.242	0.242	0.239				
PS-SAR (2-step)	0.260	0.261	0.258	0.269				

These parametric models are compared to their semiparametric counterparts: the *semiparametric P-spline a-spatial model* (PS), the *Geoadditive model*, the *PS-SAR*, the *PS-SEM*, the *PS-SDM*, and the *PS-SLX*. The last five specifications include a nonparametric smooth spatial trend. All semiparametric models are estimated by *REML*; the *PS-SAR* is estimated using both a one-step *REML* approach and a two-stage control function approach (using the *REML* method to estimate the parameters in both steps).

Comparing parametric and semiparametric models, we firstly observe that the a-spatial PS model outperforms its parametric counterpart, indicating that the functional form imposed in the parametric model does not capture all the nonlinearities in the relationship between house prices and the characteristics of the houses. This evidence is reinforced by the *Likelihood Ratio* tests reported in Table 2, where we compare the a-spatial PS model with models imposing a restricted functional form for each term (a quadratic form for *age* and a log-linear term in the other covariates). The LR tests reject the restricted specifications in all the cases.

Significant gains in model performance are obtained once the geoadditive component is included in the model, thus highlighting the importance of controlling for unobserved spatial heterogeneity. Like in the parametric framework, the inclusion of a spatial interaction term ($\mathbf{W}_n\mathbf{y}$) produces significant improvements in the goodness of fit. Using the Geoadditive model as benchmark and comparing it with the PS-SAR (2-step), we observe a decrease of the BIC value between 3.4 and 4.6%. It is also worth noticing that the estimate of the ρ parameter drops from 0.294–0.334, obtained with the parametric SAR, to 0.258–0.269 estimated with the PS-SAR specification (2-step) (Table 3). This is an expected result, since part of the spatial dependence is now captured by the spatial trend surface. All in all, our results clearly display the superiority of semiparametric spatial geoadditive models at least for this case study. Referring to the AIC values, the model that best fit the data seems to be the PS-SDM. Nevertheless, using the BIC criterion, in three out of the four years the best

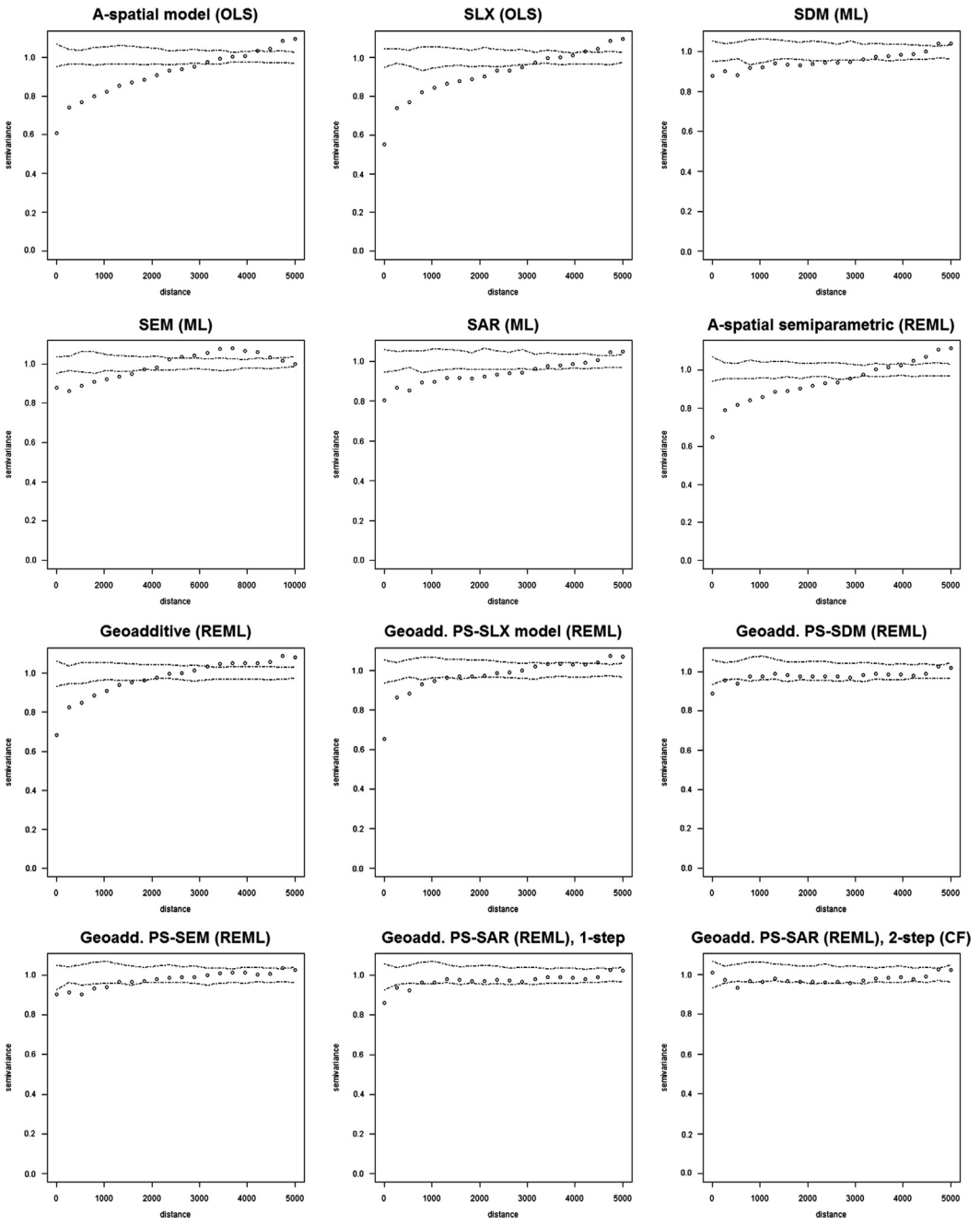

Fig. 1. Semi-variograms of the residuals (1995).

specification is the PS-SAR (2-step). Since the BIC penalizes more the degrees of freedom, we conclude that the PS-SDM is over-specified.

Following Augustin et al. (2009), we check for the gross violation of distributional assumptions of the residuals of the different models by using the empirical semi-variogram of the residuals. For the estimation of the semivariogram, we use the `variogram` function of the `geoR` package in `R`. We calculate the empirical semivariogram of the observed residuals. Then, these residuals are permuted 99 times and envelopes are computed by taking, at each spatial lag, the maximum and minimum values of the semi-variograms for the permuted residuals. Fig. 1 shows the semivariograms of the residuals resulting from the a-spatial model (OLS), traditional SAR, SEM and SDM, and new PS-SAR, PS-SEM and PS-SDM for 1995 (results for the other years are very similar). It emerges that the spatial dependence, underlying the a-spatial (OLS) model, is not captured by the traditional specifications. In fact, the semivariograms for the residuals of these specifications are Gaussian-type and does not stabilize even for large distances. According to Journel and Huijbrets (1978), this behavior is

Table 4
Control function estimates of the semiparametric PS–SAR Model. Year: 1995.

Parametric terms	First stage	Second stage
	Estimate (Bootstrap p-value)	
(Intercept)	10.955	8.159
	(0.000)	(0.000)
$\mathbf{W}_n y$		0.260
		(0.000)
baths	0.011	0.064
	(0.000)	(0.000)
$\mathbf{W}_n$*baths*	0.127	
	(0.000)	
Smooth terms	*edf*	*edf*
$f_1(age)$	5.273	5.135
$f_2(\log(lotsize))$	2.407	4.739
$f_3(\log(livingarea))$	1.002	3.356
$h(no, e)$	35.416	23.886
$f_4(res)$		3.188
$f_5(\mathbf{W}_n \log(lotsize))$	5.718	
$f_6(\mathbf{W}_n \log(livingarea))$	3.043	

Notes: bootstrap *p*-values for the significance of the parametric coefficients are reported in parenthesis. Smooth terms are specified using P-spline basis functions. Smoothing parameters are estimated using the REML. The number of knots used for the smooth terms $f_1(age)$, $f_2(\log(lotsize))$, $f_3(\log(living.area))$, $h(no, e)$, $f_4(\mathbf{W}_n age)$, $f_5(\mathbf{W}_n \log(lotsize))$, $f_6(\mathbf{W}_n \log(living.area))$, are 8, 10, 10, 8, 8, 10, and 10 respectively. *edf* means *Effective Degrees of Freedom*.

typical of spatial datasets with a (quadratic) trend and cannot be confused with the parabolic behavior at the origin exhibited by the Gaussian semivariogram model. That is, the parametric traditional specifications cannot account for the global drift of the data. However, when the PS-SAR, PS-SEM and PS-SDM are used, the experimental semivariograms turn in pure nugget semivariogram (the theoretical semivariogram representing the situation of absence of spatial correlation), which means that all these PS specifications are able to account for the spatial correlation. The estimation method does not seem to have any significant influence when it comes to capturing spatial dependence, although in our particular case study, the PS-SAR 2-step appears to be superior.

In sum, the PS-SAR model estimated using the 2-stage control function approach performs better than the other models. The results corresponding to the estimation of this model for the year 1995 appear in Table 4. First, we run a semiparametric regression of the endogenous term $\mathbf{W}_n\mathbf{y}$ on the exogenous variables and their spatial lags used as external instruments.[14] Then, we insert the first-stage residuals in the original semiparametric regression to correct the inconsistency due to the endogeneity problem. All terms, but $\mathbf{W}_n y$, *baths* and $\mathbf{W}_n$ *baths* are introduced as smooth terms. The model also includes a spatial trend surface, $h(no, e)$, constructed by using the spatial coordinates in re-scaled form. All smooth terms are specified using P-spline basis functions. Both stages are estimated using the REML method. Second-stage results show that all smooth terms have an *edf* higher than 1, confirming that not only *age*, but also log(*lotsize*) and log(*livingarea*) enter nonlinearly the model.

Fig. 2 reports the plots of total, direct and indirect effects in the PS-SAR (2-step) computed using equations (4)–(6).[15] The point-wise 95% confidence bands (obtained using the bootstrap procedure described in Section 3.3) show that all effects are also significant in most part of the variable domain. As expected, indirect effects are always lower than the direct effects.

Finally, a picture of the spatial trend surface – $h(no, e)$ – estimated with the PS-SAR (2-step) is reported in Fig. 3. It emerges quite clearly that, even after having controlled for the effect of the characteristics of the houses (in terms of age, number of bathrooms, lot size and living area) and for the spillover effects (through the spatial lag term), there are

[14] Unfortunately there is not a well-known and widely accepted test for the validity of the conditional mean restrictions imposed by the CF approach. A practically feasible way of testing such restrictions consists of including some of the excluded instruments in the control function (CF) estimate and check for the significance of their coefficients (we thank Jeffrey Wooldridge for having suggested us this method). These coefficients should not be significant because the CF should pick up all of the correlation between the structural error term and $(\mathbf{W}_n\mathbf{y}, \mathbf{Z})$, where $(\mathbf{Z} = (\mathbf{X}, \mathbf{W}_n\mathbf{X}))$. In particular, if $\mathbf{y} = \rho\mathbf{W}_n\mathbf{y} + f(\mathbf{X}) + u_1$ and $\mathbf{W}_n\mathbf{y} = \mathbf{X}\delta + v_2$, then $\mathbf{W}_n\mathbf{X}$ can be added to the CF estimation. This means that we can regress $\mathbf{y}$ on $\mathbf{W}_n\mathbf{y}$, $\mathbf{X}$, $\hat{v}_2$ and $\mathbf{W}_n\mathbf{X}$, using whatever model/estimation method, and test coefficients on $\mathbf{W}_n\mathbf{X}$. In our case, the only spatial lag variable ($\mathbf{W}_n\mathbf{X}$) which turned out to be not strictly exogenous was $\mathbf{W}_n$(*age*); this variable was removed from the set of excluded instruments.

[15] Actually, total, direct and indirect effects are not smooth over the domain of variable x_k due to the presence of the spatial multiplier matrix in the algorithms. A wiggly profile of direct, indirect and total effects would appear even if the model were linear. Therefore, in the spirit of this paper, we have applied a spline smoother to obtain smooth curves.

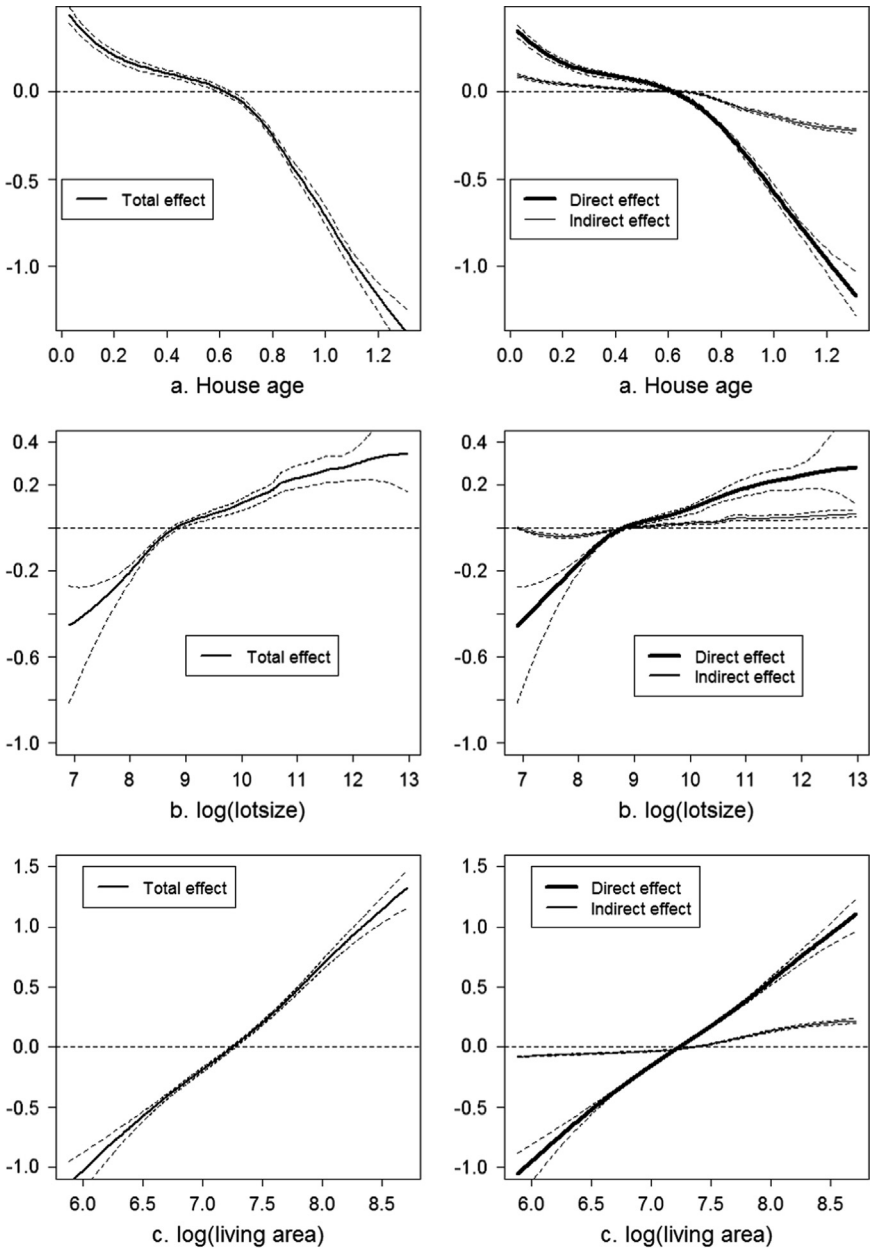

Fig. 2. Total effects (left panels) and direct and indirect (right panels) – PS–SAR. Bootstrapped confidence bands in dotted lines.

significant house-price differences over space. In particular, the house prices are significantly higher in the southern part of the selected area. This confirms the role of local (unobserved) characteristics in house prices formation.

5. Conclusions and some practical advice for users

The standard spatial econometrics approach relies on a classical statistical model in which the true model is known *a-priori*. Recently, McMillen (2012) has criticized this approach. He argues that traditional spatial autoregressive models serve basically to compensate for the effects of omitted variables that are correlated over space and for the effects of functional form misspecification. He points out that there are alternative approaches (namely semiparametric regression approaches) which can be used instead of standard parametric spatial regression models. These approaches admit at the start that the true model structure is unknown.

Our contribution is very much in line with this view. However, we also consider the cases where the theory (more or less formally) suggests the existence of spatial spillovers, but it does not predict a highly structured model, so that the functional

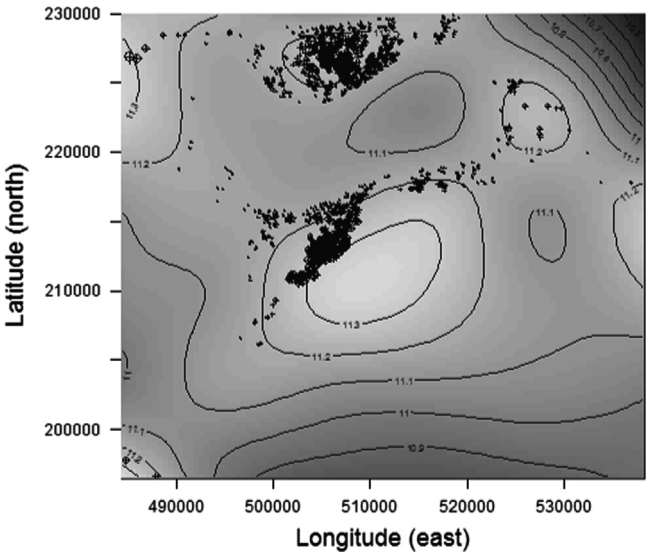

Fig. 3. Spatial trend surface.

form of the empirical equation remains unknown. In these cases, it becomes important to identify spatial interaction effects (introducing spatial lag terms in the model) rather than simply smooth the data over space to remove spatial autocorrelation. Thus, flexible semiparametric spatial autocorrelation models (such as PS-SAR and PS-SDM) remain a very important tool to identify the functional form of the relationship between the response variable and its predictors and to identify spatial interaction effects.

Moreover, even if the theory predicts a highly structured model, its empirical implementation always requires to control for the effect of omitted factors. Thus, the inclusion of a spatial trend surface turns out to be a useful tool to disentangle substantive spatial autocorrelation (spatial spillovers), captured by the parameter of the spatial lag term, and spatial autocorrelation generated by (spatially autocorrelated) omitted variables.

We are not proposing the use of Spatial Autoregressive Semiparametric Geoadditive Models only for description purposes. In fact, we suggest that these models may help to draw substantive inferences on the backdrop of the underlying theoretical models and, eventually, to explore opportunities to improve the theoretical models.

Obviously, we also recognize the existence of some major practical problems associated with the implementation of Spatial Autoregressive Semiparametric Geoadditive Models. First of all, it is well known that nonparametric estimates may be spurious due to *outliers*, although in the case of penalized splines the effect of the extreme values is often mitigated. In practice, as in our application, it is often necessary to trim the data (either symmetrically or asymmetrically) in order to reduce the effect of extreme values.

Second, as in the case of traditional parametric spatial econometric models, the *spatial interaction network underlying the* $\mathbf{W}_n$ *matrix* is not known beforehand, which introduces uncertainty in the building of spatial autoregressive models. Usual practice in the applied literature is to simply impose the spatial weight matrix as a maintained hypothesis; this solution implies a certain degree of arbitrariness. However, recently some procedures have been suggested in order to approach the problem in a more formalized way (see,for example, Harris et al., 2011; LeSage and Pace, 2009; Kostov, 2010). These methods can also be easily applied to the semiparametric approach proposed in the present paper. Nevertheless, in our application, the results of PS-SAR, PS-SDM and PS-SEM turned out to be very robust to the choice of the $\mathbf{W}_n$ matrix. Only the results of the PS-SARSAR appears to be sensitive to the $\mathbf{W}_n$ matrix used, as indicated in Section 4.2.

Third, regarding the problem of *model selection*, it seems preferable to simply compare the performance of the different models in terms of Schwarz' Bayesian Information Criterion and to check for the gross violation of distributional assumptions of the residuals of the different models by using the empirical semi-variogram of the residuals. We do not provide a battery of diagnostic tests for Spatial Autoregressive Semiparametric Geoadditive Models like the Lagrange Multiplier tests widely used in the traditional parametric spatial econometric literature (LM-SEM, LM-SAR, LM-SARSAR). Indeed, the use and abuse of LM tests for the spatial autocorrelation of the residuals has been largely criticized, as it may yield to a mechanical selection process.

Fourth, concerning the specification of the *smooth terms* in the semiparametric models, it seems preferable to use the B-spline bases, given their desirable numerical properties, with a discrete penalty on the basis coefficients as proposed by Eilers and Marx (1996). Moreover, with a reparameterization of the penalized additive model as a mixed model, the smoothing parameters (that control the trade-off between fidelity to the data and smoothness of the fitted spline) are treated as variance parameters and can be estimated by REML together with the other parameters, avoiding any arbitrariness in the choice of the degree of smoothing of each term. Another advantage of the reparametrization proposed in this paper is that it allows for anisotropy.

Fifth, an obvious advantage of the 2-step control function approach with respect to the 1-step REML approach in relation to the estimation of the PS-SAR is that it allows to eventually control for the *endogeneity of r.h.s. terms* different from $\mathbf{W}_n\mathbf{y}$, if valid instruments are available.

Sixth, the PS-SAR, PS-SEM, PS-SDM and PS-SLX models can be easily extended to a *static longitudinal data* framework when panel data are available. Eventually, a random spatial effect can be included in the semiparametric model when it is reparameterized as a mixed model.

Finally, a major practical problem associated with the implementation of Spatial Autoregressive Semiparametric Geoadditive Models is the lack of a standard *software* to estimate the entire set of semiparametric models. However, we have developed specific functions in `R` software that allow the estimation of the large range of semiparametric models included in the paper (1-step PS-SAR, PS-SEM, PS-SDM, PS-SLX, PS-SARSAR). It is also relatively easy to adapt the `R` library `mgcv` developed by Simon Wood to implement the 2-step control function approach for the SAR model.

Acknowledgments

This paper has been presented at the EU-COST Meeting in Lisboa. We thank all participants for useful comments. We are also grateful to two referees that with their comments helped us to reformulate the analysis. We are responsible for any remaining errors. The work of Román Mínguez and María Durbán was supported by the research project MTM-2011-28285-C02-C2 from the Spanish Government's Ministry of Economy and Competitiveness. Jesús Mur likes to thank the financial support of the research project ECO2012-36032-C03-02 from the Ministerio de Educación, Cultura y Deporte del Reino de España. Jose María Montero wants to acknowledge the Junta de Comunidades de Castilla La Mancha for funding part of this research with the Project POII10-0250-6975.

Appendix A. Models specification

1. Parametric a-spatial Model:

$$y_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 \log(lotsize)_i + \beta_4 \log(livingarea)_i + \beta_5 baths_i + \varepsilon_i,$$
$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2), \quad i = 1, \ldots, n.$$

2. Parametric Spatial in X Model (SLX):

$$\begin{aligned} y_i = {} & \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 \log(lotsize)_i + \beta_4 \log(livingarea)_i + \beta_5 baths_i \\ & + \delta_1 \sum_{j=1}^{n} w_{ij} age_j + \delta_2 \sum_{j=1}^{n} w_{ij} age_j^2 + \delta_3 \sum_{j=1}^{n} w_{ij} \log(lotsize)_j \\ & + \delta_4 \sum_{j=1}^{n} w_{ij} \log(livingarea)_j + \delta_5 \sum_{j=1}^{n} w_{ij} baths_j + \varepsilon_i, \\ & \varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2), \quad i = 1, \ldots, n. \end{aligned}$$

3. Parametric Spatial Durbin Model (SDM):

$$\begin{aligned} y_i = {} & \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 \log(lotsize)_i + \beta_4 \log(livingarea)_i + \beta_5 baths_i \\ & + \delta_1 \sum_{j=1}^{n} w_{ij} age_j + \delta_2 \sum_{j=1}^{n} w_{ij} age_j^2 + \delta_3 \sum_{j=1}^{n} w_{ij} \log(lotsize)_j \\ & + \delta_4 \sum_{j=1}^{n} w_{ij} \log(livingarea)_j + \delta_5 \sum_{j=1}^{n} w_{ij} baths_j + \rho \sum_{j=1}^{n} w_{ij} y_j + \varepsilon_i, \\ & \varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2), \quad i = 1, \ldots, n. \end{aligned}$$

4. Parametric Spatial Error Model (SEM):

$$\begin{aligned} & y_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 \log(lotsize)_i + \beta_4 \log(livingarea)_i + \beta_5 baths_i + u_i, \\ & u_i = \lambda \sum_{j=1}^{n} w_{ij} u_j + \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned}$$

5. Parametric Spatial Lag Model (SAR):

$$y_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 \log(lotsize)_i + \beta_4 \log(livingarea)_i + \beta_5 baths_i + \rho \sum_{j=1}^{n} w_{ij} y_j + \varepsilon_i,$$
$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2), \quad i = 1, \ldots, n.$$

6. Semiparametric Additive Model (a-spatial PS):

$$y_i = \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i + \varepsilon_i,$$
$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n.$$

7. Semiparametric Geoadditive Model (PS-Geo):

$$y_i = \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i + h(no_i, e_i) + \varepsilon_i,$$
$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n.$$

8. Semiparametric Geoadditive SLX (PS-Geo-SLX):

$$\begin{aligned} y_i = {} & \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i \\ & + g_1\left(\sum_{j=1}^{n} w_{ij} age_j\right) + g_2\left(\sum_{j=1}^{n} w_{ij} \log(lotsize)_j\right) + g_3\left(\sum_{j=1}^{n} w_{ij} \log(livingarea)_j\right) \\ & + \delta_1 \sum_{j=1}^{n} w_{ij} baths_j + h(no_i, e_i) + \varepsilon_i, \\ & \varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n. \end{aligned}$$

9. Semiparametric Geoadditive SDM (PS-Geo-SDM):

$$\begin{aligned} y_i = {} & \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i \\ & + g_1\left(\sum_{j=1}^{n} w_{ij} age_j\right) + g_2\left(\sum_{j=1}^{n} w_{ij} \log(lotsize)_j\right) + g_3\left(\sum_{j=1}^{n} w_{ij} \log(livingarea)_j\right) \\ & + \delta_1 \sum_{j=1}^{n} w_{ij} baths_j + \rho \sum_{j=1}^{n} w_{ij} y_j + h(no_i, e_i) + \varepsilon_i, \\ & \varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n. \end{aligned}$$

10. Semiparametric Geoadditive SEM (PS-Geo-SEM):

$$\begin{aligned} & y_i = \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i + h(no_i, e_i) + u_i, \\ & u_i = \lambda \sum_{j=1}^{n} w_{ij} u_j + \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned}$$

11. Semiparametric Geoadditive SAR (PS-Geo-SAR):

$$y_i = \beta_0 + f_1(age_i) + f_2(\log(lotsize)_i) + f_3(\log(livingarea)_i) + \beta_1 baths_i + \rho \sum_{j=1}^{n} w_{ij} y_j + h(no_i, e_i) + \varepsilon_i,$$
$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n.$$

References

Anselin, L., 2003. Spatial externalities, spatial multipliers and spatial econometrics. Int. Reg. Sci. Rev. 26, 153–166.

Arbia, G., Paelinck, J., 2003. Spatial econometric modeling of regional convergence in continuous time. Int. Reg. Sci. Rev. 26, 342–362.

Augustin, N., Musio, M., Wilpert, K.V., Kublin, E., Wood, S., Schumacher, M., 2009. Modeling spatio-temporal forest health monitoring data. J. Am. Stat. Assoc. 104, 899–911.

Autant-Bernard, C., 2012. Spatial econometrics of innovation: recent contributions and research perspectives. Spat. Econ. Anal. 7, 403–419.

Azomahou, T., Ouardighi, J.E., Nguyen-Van, P., Pham, T., 2011. Testing convergence of european regions: a semiparametric approach. Econ. Model. 28, 1202–1210.

Basile, R., 2008. Regional economic growth in Europe: a semiparametric spatial dependence approach. Pap. Reg. Sci. 87, 527–544.

Basile, R., 2009. Productivity polarization across regions in Europe: the role of nonlinearities and spatial dependence. Int. Reg. Sci. Rev. 32, 92–115.

Basile, R., Capello, R., Caragliu, A., 2012. Technological interdependence and regional growth in Europe. Pap. Reg. Sci. 91, 697–722.

Basile, R., Donati, C., Pittiglio, R., 2013. Industry structure and employment growth: evidence from semiparametric geoadditive models. Rég. Dév. 38, 121–160.

Basile, R., Gress, B., 2005. Semi-parametric spatial auto-covariance models of regional growth behavior in Europe. Rég. Dév. 21, 93–118.

Bivand, R., 2010. Comparing Estimation Methods for Spatial Econometrics Techniques Using R. Discussion paper, 26, Department of Economics, Norwegian School of Economics and Business Administration.

Bivand, R., 2012. After raising the bar: applied maximum likelihood estimation of families of models in spatial econometrics. Estad. Espa. 54, 71–88.

Blundell, R., Powell, J., 2003. Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, M., Hansen, L., Turnsovsky, S.J. (Eds.), Advances in economics and econometrics, Cambridge University Press, Cambridge.
Bourassa, S.C., Cantoni, E., Hoesli, M., 2010. Predicting house prices with spatial dependence: a comparison of alternative methods. J. Real Estate Res. 32, 139–159.
Brady, R.R., 2011. Measuring the diffusion of housing prices across space and over time. J. Appl. Econom. 26, 213–231.
Brueckner, J.K., 2000. Urban sprawl: diagnosis and remedies. Int. Reg. Sci. Rev. 23, 160–171.
Brueckner, J.K., 2006. A companion to urban economics. Strategic Interaction Among Governments. Blackwell, Malden, MA, pp. 332–347.
Brueckner, J.K., Mills, E., Kremer, M., 2001. Urban Sprawl: Lessons from Urban Economics [with Comments]. Brookings–Wharton Papers on Urban Affairs, pp. 65–97.
Can, A., 1992. Specification and estimation of hedonic housing price models. Reg. Sci. Urban Econ. 22, 453–474.
Can, A., Megbolugbe, I., 1997. Spatial dependence and house price index construction. J. Real Estate Financ. Econ. 14, 203–222.
Chasco, C., Le Gallo, J., 2011. The Impact of Objective and Subjective Measures of Air Quality and Noise on House Prices: A Multilevel Approach for Downtown Madrid. Discussion paper, European Regional Science Association.
Currie, I., Durbán, M., Eilers, P., 2006. Generalized array models with application to multidimensional smoothing. J. R. Stat. Soc. B 68, 259–280.
De Boor, C., 2001. A practical guide to splines. Applied Mathematical Sciences, vol. 27 (Revised Edition). Springer, New York.
Dubé, J., Legros, D., 2013. Dealing with spatial data pooled over time in statistical models. Lett. Spat. Resour. Sci. 6, 1–18.
Eilers, P., Marx, B., 1996. Flexible smoothing with b-splines and penalties. Stat. Sci. 11, 89–121.
Ekeland, I., Heckman, J., Nesheim, L., 2004. Identification and estimation of hedonic models. J. Polit. Econ. 112, 60–109.
Ertur, C., Gallo, J.L., 2009. Handbook of regional growth and development theories. Regional Growth and Convergence: Heterogenous Reaction versus Interaction Spatial Econometric Approaches. Edward Elgar, Cheltenham, pp. 374–388.
Ertur, C., Koch, W., 2007. Growth, technological interdependence and spatial externalities: theory and evidence. J. Appl. Econom. 22, 1033–1062.
Fahrmer, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Models, Methods and Applications. Springer, Berlin.
Fiaschi, D., Lavezzi, M., Parenti, A., 2013. On the Determinants of Distribution Dynamics. Mimeo.
Fischer, M., Stumpner, P., 2010. Income distribution dynamics and cross-region convergence in Europe. In: Handbook of Applied Spatial Analysis, vol. 4, pp. 599–628. Springer, Berlin, Heidelberg and New York.
Fotopoulos, G., 2012. Nonlinearities in regional economic growth and convergence: the role of entrepreneurship in the European union regions. Ann. Reg. Sci. 48, 719–741.
Gibbons, S., Overman, H.G., 2012. Mostly pointless spatial econometrics. J. Reg. Sci. 52, 172–191.
Goodman, A.C., Thibodeau, T.G., 1995. Age-related heteroskedasticity in hedonic house price equation. J. Hous. Res. 6, 25–42.
Goodman, A.C., Thibodeau, T.G., 2003. Housing market segmentation and hedonic prediction accuracy. J. Hous. Econ. 12, 181–201.
Gress, B., 2004. Using Semi-Parametric Spatial Autocorrelation Models to Improve Hedonic Housing Price Prediction. Mimeo, Department of Economics, University of California.
Griffith, D., Paelinck, J., 2011. Non-Standard Spatial Statistics and Spatial Econometrics. Springer-Verlag, Berlin vol. 1.
Harris, R., Moffat, J., Kravtsova, V., 2011. In search of W. Spat. Econ. Anal. 6, 249–270.
Holly, S., Pesaran, H.M., Yamagata, T., 2011. The spatial and temporal diffusion of house prices in the UK. J. Urban Econ. 69, 2–23.
Irwin, E.G., Bockstael, N.E., 2007. The evolution of urban sprawl: evidence of spatial heterogeneity and increasing land fragmentation. Proc. Natl. Acad. Sci. 104 (52), 20672–20677.
Journel, A., Huijbrets, C., 1978. Mining Geostatistics. Academic Press, New York.
Kelejian, H., Prucha, I., 1997. Estimation of spatial regression models with autoregressive errors by two-stage least squares procedures: a serious problem. Int. Reg. Sci. Rev. 20, 103–111.
Kim, S.W., Bhattacharya, R., 2009. Regional housing prices in the USA: an empirical investigation of nonlinearity. J. Real Estate Financ. Econ. 38, 443–460.
Kneib, T., Hothorn, T., Tutz, G., 2009. Variable selection and model choice in geoadditive regression models. Biometrics 65, 626–634.
Kostov, P., 2010. Model boosting for spatial weighting matrix selection in spatial lag models. Environ. Plan. B: Plan. Des. 37 (3), 533.
Kuethe, T.H., Pede, V.O., 2011. Regional housing price cycles: a spatio-temporal analysis using US state-level data. Reg. Stud. 45, 563–574.
Lambert, D.M., Xu, W., Florax, R.J., 2014. Partial adjustment analysis of income and jobs, and growth regimes in the appalachian region with smooth transition spatial process models. Int. Reg. Sci. Rev. 37, 328–364.
Lee, D., Durbán, M., 2011. P-Spline ANOVA type interaction models for spatio-temporal smoothing. Stat. Model. 11, 49–69.
Lee, L.F., Liu, X., Lin, X., 2010. Specification and estimation of social interaction models with network structures. Econom. J. 13, 145–176.
LeSage, J., Pace, K., 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton.
Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60, 531–542.
McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized, linear, and mixed models, 2nd edition. Wiley Series in Probability and Statistics, Wiley, Hoboken, New Jork.
McMillen, D.P., 1996. One hundred fifty years of land values in Chicago: a nonparametric approach. J. Urban Econ. 40, 100–124.
McMillen, D.P., 2010. Issues in spatial data analysis. J. Reg. Sci. 50, 119–141.
McMillen, D.P., 2012. Perspectives on spatial econometrics: linear smoothing with structured models. J. Reg. Sci. 52, 192–209.
Mínguez, R., Durbán, M., Montero, J., Lee, D., 2012. Competing Spatial Parametric and Non-Parametric Specifications. Mimeo.
Montero, J., Mínguez, R., Durbán, M., 2012. SAR models with nonparametric spatial trends. A P-spline approach. Estad. Espa. 54, 89–111.
Mur, J., López, F., Angulo, A., 2010. Instability in spatial error models: an application to the hypothesis of convergence in the European case. J. Geograph. Syst. 12, 259–280.
Pace, K., LeSage, J., 2004. Spatial econometrics and spatial statistics. Spatial Autoregressive Local Estimation. Palgrave Macmillan, Basingstoke, pp. 31–51.
Pinheiro, J., Bates, D., 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York.
Pinkse, J., Slade, M., 2010. The future of spatial econometrics. J. Reg. Sci. 50, 103–117.
Ruppert, D., Wand, M., Carroll, R., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.
Su, L., 2012. Semiparametric GMM estimation of spatial autoregressive models. J. Econom. 167, 543–560.
Su, L., Jin, S., 2010. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. J. Econom. 157, 18–33.
Wand, M., 2003. Smoothing and mixed models. Comput. Stat. 18, 223–249.
Wood, S., 2003. Thin plate regression splines. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 65, 95–114.
Wood, S., 2006. Generalized Additive Models. An Introduction with R. Chapman and Hall, London.
Wood, S., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 73, 3–36.
Wood, S., Scheipl, F., Faraway, J., 2012. Straightforward intermediate rank tensor product smoothing in mixed models. Stat. Comput. 23, 341–360.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge, Massachusetts.
Zhu, B., Füss, R., Rottke, N., 2011. The predictive power of anisotropic spatial correlation modeling in housing prices. J. Real Estate Financ. Econ. 42, 542–565.