Modelling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities

Roberto Basile* Román Mínguez† Jose María Montero‡ Jesús Mur§

July 11, 2013

Abstract

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

Keywords: Spatial econometrics, nonlinearities, semiparametric models.
Jel codes: R11, R12, C14

*Corresponding author. Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 1 - 81043, Capua (CE), Italy. *Email*: roberto.basile@unina2.it

†Statistics Department, University of Castilla-La Mancha. Cuenca, Spain. roman.minguez@uclm.es

‡Statistics Department, University of Castilla-La Mancha. Toledo, Spain. jose.mlorenzo@uclm.es

§Department of Economic Analysis, University of Zaragoza, Spain. jmur@unizar.es

1 Introduction

Spatial dependence and spatial heterogeneity are two characteristics of spatial data that need to be taken into account when modelling spatial economic dynamics. This is a *mantra* insistently repeated since the influential textbook of Anselin (1988). Spatial dependence reflects a situation where values observed at one location or region depend on the values of neighbouring observations at nearby locations (LeSage and Pace, 2009); spatial heterogeneity points to the lack of spatial stability of the relationships under study (functional forms and parameters vary with location and are not homogeneous throughout the data set) (Fotheringham, Brunsdon, and Charlton, 2002). By far, spatial dependence has been the main attractor and concentrates the most singular results in the literature of spatial econometrics, although there remain important problems to be solved. Spatial heterogeneity, on the contrary, has attracted more attention in the realm of theoretical spatial economics than in the field of econometric modelling.

When it comes to the point of building spatial econometric models, it is necessary to consider a specific form of (spatial) heterogeneity which characterises spatial data analysis, that is nonlinearity. This is not a point of consensus in the literature on spatial econometrics, where linearity clearly dominates. The premise is that a linear structure, possibly coupled with some previous functional transformation of the variables warrants enough flexibility to account for the complexity of spatial data formation. However, there is growing evidence, coming from different fields, showing that this is a quite optimistic view. Strong nonlinearities have been detected in studies on regional growth (Arbia and Paelinck, 2003; Azomahou, Ouardighi, Nguyen-Van, and Pham, 2011; Basile and Gress, 2005; Basile, 2008, 2009; Basile, Capello, and Caragliu, 2012; Ertur and Gallo, 2009; Fotopoulos, 2012), urban agglomeration economies (Basile, Donati, and Pittiglio, 2013), urban environment (Chasco Yrigoyen and Le Gallo, 2011), land prices (McMillen, 1996), urban sprawl (Brueckner, 2000; Brueckner, Mills, and Kremer, 2001; Irwin and Bockstael, 2007), social interaction (Lee, Liu, and Lin, 2010) and house prices (Bourassa, Cantoni, and Hoesli, 2010; Kim and Bhattacharya, 2009; Goodman and Thibodeau, 2003).

The dominant parametric approach in spatial econometrics is not well equipped to deal *simultaneously* with the three topics (spatial dependence, spatial heterogeneity and nonlinearities). They have been approached separately and only recently there have been attempts to mix some of them. This is the case of Lambert, Xu, and Florax (2013), which combine spatial dependence and nonlinerity in a STAR model, or the case of the Lotka-Volterra prey-predator model discussed in 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, López, and Angulo, 2010) can be considered limiting cases. To our knowledge, few more references can be added. In fact, the history is very short.

Given the limitations of the parametric framework, it is important to pay attention to other, more flexible, approaches, which offer a more convenient way of addressing simultaneously the three problems: dependence, heterogeneity and nonlinearity. This is the case of the *Spatial Autoregressive Semiparametric Geoadditive Models* developed, among others, by Basile and Gress (2005), Su and Jin (2010), Su (2012), Basile, Capello, and Caragliu (2012) and Montero, Mínguez, and Durbán (2012). The objective of this paper is to describe the main methodological

contributions produced recently in this field, which help us to overcome some of the deficiencies encountered in a parametric framework. We also illustrate the discussion with an application to the case of house prices in Lucas County.

Section 2 introduces different specifications of the semiparametric model: (*i*) the Penalized Spline (PS) Geoadditive Model, (*ii*) the Penalized Spline Spatial Lag (or Spatial Autoregressive) Geoadditive Model (PS-SAR) and (*iii*) the Penalized Spline Spatial Error Geoadditive Model (PS-SEM). Section 3 discusses different technical aspects related to the estimation of these models, to the choice of the smoothing parameters and to the solution of identification issues. Section 4 includes an application to the case of spatial modelling house prices in Lucas County. The application compares parametric and semiparametric regression estimates. The differences are clearly in favour of the more flexible semiparametric specification. Section fifth recaps and summarizes the main findings in our work.

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 spatial heterogeneity. We start by introducing a general specification of the semiparametric geoadditive model (Subsect. 2.1). In Subsections 2.2 and 2.3 we extend this model by introducing the spatial lag of the dependent variable on the *rhs*, or alternatively, a spatial autoregressive error term, thus obtaining the (*PS-SAR*) and the (*PS-SEM*) specifications, respectively.

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:[1]

$$\begin{aligned} y_i &= \mathbf{x}_i^{*'}\beta^* + f_1(x_{1i}) + f_2(x_{1i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + ... \\ &\quad + h(no_i, e_i) + \varepsilon_i \qquad \varepsilon_i \sim iid\mathcal{N}\left(0, \sigma_\varepsilon^2\right) \qquad i = 1, ..., n \end{aligned} \tag{1}$$

where y_i is a continuous univariate response variable measuring, for example, the average annual productivity growth rate of region i or the price of the house i. $\mathbf{x}_i^{*'}\beta^*$ is the linear predictor for any strictly parametric component (including the intercept, all categorical covariates and eventually some continuous covariates), with β^* being a vector of fixed parameters. $f_k(.)$ 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, Hothorn, and Tutz, 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 want to test whether the smooth effect of x_1 (e.g., population density) is different in the North and in the South. In this case l_i is a binary variable taking value one if region

[1] Although this model is widely used in environmental studies and in epidemiology (Augustin, Musio, Wilpert, Kublin, Wood, and Schumacher, 2009), it is rarely considered for modelling economic data.

i belongs to the North and zero if it belongs to the South. Thus, if $l_i = 0$, the effect of x_1 is given by $f_1(x_{1i})$, whereas for $l_i = 1$, the effect is composed as the sum $f_1(x_{1i}) + f_4(x_{1i})$, and $f_4(x_{1i})$ can be interpreted as the deviation of x_1 for the North. The term $h(no_i, e_i)$ in equation (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.[2] 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). For example, by using this interaction term, we can test the assumption that the effect of urbanization economies on local productivity in Italy varies moving from the South to the North, or from North–Western to North–Eastern regions. Finally, ε_i are *iid* normally distributed random shocks.[3]

In the case of the pure penalized regression spline 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), that is it can be interpreted as the Average Structural Function (ASF) (Blundell and Powell, 2003).

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

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

with β_{q_k} unknown parameters to be estimated. To reduce mis-specification bias, $q_k's$ must be made fairly large. But this may generate a danger of over-fitting. As we shall clarify further on, by penalizing 'wiggly' functions when fitting the model, the smoothness of the functions can be controlled. Thus, a measure of 'wiggliness' $J \equiv \beta' \mathbf{S} \beta$ is associated with each k smooth function, with $\mathbf{S}$ a positive semidefinite matrix. Typically, the quadratic penalty term is equivalent to an integral of squared second 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).

The penalized spline base–learners can be extended to two or more dimensions to handle interactions by using thin–plate regression splines or tensor products (Wood, 2006a, Section 4.1.5). In the case of a tensor product, 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, 2006a). 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 $f(x_3, x_4)$. Thus, in order to test for

[2] Removing *unobserved spatial patterns* is a primary task, 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.

[3] Equation (1) can be augmented by relaxing the *iid* assumption for the error term, that is assuming an error vector $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \Lambda)$ with a covariance matrix Λ reflecting spatial error correlation as, for example, in Pinheiro and Bates (2000).

smooth interaction effects, we do not need to include in the model the two further terms $f(x_3)$ and $f(x_4)$.

In the case of a varying coefficient term like $f_4(x_1)\,l$, the basis functions $b_{q_4}(x_1)$ are premultiplied 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 premultiplied by a diagonal matrix containing the values of the interaction variable x_1.

Given the bases for each smooth term, equation (1) can be rewritten in matrix form as a large linear model,

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

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

2.2 Penalized Spline Spatial Lag Geoadditive Models (PS-SAR)

The Geoadditive model (1) represents a quite general framework to model spatial data taking account of nonlinearities and spatial heterogeneity. However, this model rules out spatial interaction or spatial spillover effects. One way of dealing with this issue is the introduction of spatial lags of the exogenous variables on the *rhs* of model (1), thus capturing so-called *local spatial externalities*.[4] However, 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 Spatial Autoregressive Geoadditive Model or Penalized-Spline SAR model (PS-SAR as called in Mínguez, Durbán, Montero, and Lee, 2012):

$$\begin{aligned} y_i &= \mathbf{x}_i^{*'}\beta^* + \rho\sum_{j=1}^{n} w_{ij}y_j + f_1(x_{1i}) + f_2(x_{2i}) \\ &\quad + f_3(x_{3i},x_{4i}) + f_4(x_{1i})\,l_i + ... + h(no_i,e_i) + \varepsilon_i \end{aligned} \tag{3}$$

$$\varepsilon_i \sim iid\,\mathcal{N}(0,\sigma_\varepsilon^2) \qquad i = 1,...,n$$

where w_{ij} are the elements of a spatial weights matrix $\mathbf{W}_n$, $\sum_{j=1}^{n} w_{ij}y_j$ captures 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, Capello, and Caragliu (2012), Montero, Mínguez,

[4] In spatial econometrics, 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) ($\mathbf{y} = \mathbf{A}\mathbf{X}'\beta + \mathbf{A}\varepsilon$), 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 a change in the error ε in region i – impacts on the outcome of this region, on the outcome of its neighbours, on that of the neighbours of its neighbours and so on. The impact therefore is global. In the case of SEM, the global spillover effect concerns only unmodelled random shocks ($\mathbf{y} = \mathbf{X}'\beta + \mathbf{B}u$, 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) ($\mathbf{y} = \mathbf{X}'\beta + \mathbf{W}_n\mathbf{X}'\delta + \varepsilon$). In fact, there is no inverse involved in the reduced form of this model, so as the impact of the change dies just after its effect on the neighbours (the structural form of the SLX is in fact the reduced form).

and Durbán (2012), Mínguez, Durbán, Montero, and Lee (2012), Su and Jin (2010) and Su (2012). It reflects the notion of a spatial correlation comprised of two parts: (*i*) a spatial trend due to unobserved regional characteristics, which is modelled by the smooth function of the coordinates, and (*ii*) global spatial spillover effects, which is modelled by including the spatial lag of the dependent variable. Su (2012) extends this model to allow for both heteroskedasticity and spatial dependence in the error term.

As in the parametric SAR, in the PS-SAR also the estimated coefficients of parametric terms ($\widehat{\beta}^*$) cannot be interpreted as marginal effects of the explanatory variables on the dependent variable, due to the presence of a significant spatial autoregressive parameter (ρ). Direct, indirect (spillover) and total effects must be computed instead after estimation using the algorithms described in LeSage and Pace (2009) for parametric SAR. For the same reason, the estimated smooth functions — $\widehat{f}_k(x_k)$ — cannot be interpreted as ASF, that is as conditional expectations of y given x_k. Taking advantage of the results obtained for parametric SAR, we can compute the total smooth effect (total–ASF) of x_k as

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

Finally, we can compute direct and indirect (or spillover) effects of smooth terms in PS-SAR as follows:

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

$$\widehat{f}_k^{I_k}(x_k) = \widehat{f}_k^{T_k}(x_k) - \widehat{f}_k^{D_k}(x_k) \tag{6}$$

2.3 Penalized Spline Spatial Error Geoadditive Models (PS-SEM)

An alternative specification of the semiparametric model that captures global spatial externalities (in the form of spatial diffusion of idiosyncratic random shocks), together with the nonlinearities and spatial unobserved heterogeneity, is the Spatial Error Geoadditive Model or PS-SEM proposed by Mínguez, Durbán, Montero, and Lee (2012). This specification augments the Penalized Spline Geoadditive Model by including a spatial autoregressive error term, while leaving the systematic part of the model unchanged:

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

where, again, w_{ij} is the element of a spatial weights matrix $\mathbf{W}_n$ and 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). In other words, the model cannot capture any spatial spillover of shocks in a modelled x_k factor. Nevertheless, the PS-SEM allows us to capture spatial externalities in un-modelled idiosyncratic random shock, since the reduced form of the model is

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

3 Estimation Methods

Let us now discuss the issues concerning the estimation of model parameters in the semiparametric models described above and the related inference starting from the assumptions of an independent error structure and strict exogeneneity of all explanatory variables, that is starting from the estimation of model (1). Specifically, we describe two alternative estimators of model (1). 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 (Section 3.1). The semiparametric model (1) can also be expressed as a mixed model. Consequently, it is possible to estimate all the parameters of this model using restricted maximum likelihood methods (REML). In Section (Section 3.2), we present some ways to deal with general semiparametric models using mixed models. Subsection 3.3 shows how this methodology can be applied to estimate the parameters of PS-SAR and PS-SEM models in a single step. Finally, in Section 3.4 we present an alternative two-step control function approach to estimate the PS-SAR model.

3.1 PLS and GCV Score Minimization

As already mentioned, the number of parameters for each smooth term in a semiparametric model must be large enough to reduce misspecification bias, but not too large to escape overfitting. To solve this trade–off, we need to penalize lack of smoothness. Thus, starting from the assumption of exogeneity of all the *rhs* variables, model (2) can be estimated by solving the following optimization problem

$$\min \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_k \theta_k \beta' \mathbf{S}_k \beta \qquad w.r.t. \qquad \beta \tag{8}$$

subject to any constraints associated with the bases plus any constraints needed to ensure that the model is identifiable. $\|.\|^2$ is the Euclidean norm and $\theta_k \geq 0$ are the smoothing parameters that control the fit vs. smoothness trade–off. Employing a large number of basis functions yields a flexible representation of the nonparametric effect $f_k(.)$ where the actual degree of smoothness can be adaptively chosen by varying θ_k.[5]

Given smoothing parameters, θ_k, the solution to (8) is the following penalized least square estimator:

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X} + \sum_k \theta_k \mathbf{S}_k \right)^{-1} \mathbf{X}'\mathbf{y}$$

The covariance matrix of $\widehat{\beta}$ can be derived from that of $\mathbf{y}$

$$V_{\widehat{\beta}} \quad = \quad \sigma_\varepsilon^2 \left(\mathbf{X}'\mathbf{X} + \sum_k \theta_k \mathbf{S}_k \right)^{-1} \mathbf{X}'\mathbf{X} \left(\mathbf{X}'\mathbf{X} + \sum_k \theta_k \mathbf{S}_k \right)^{-1}$$

[5] It is worth noticing that in expression (8), for interactive terms, the penalty matrix $\mathbf{S}_k$ usually depends on both interacting variables, and the associated θ_k will have two components allowing for different degrees of smoothing.

If we also assume normality, that is $\varepsilon \sim \mathcal{N}\left(0, \mathbf{I}_n \sigma_\varepsilon^2\right)$, then

$$\widehat{\beta} \quad \sim \quad \mathcal{N}\left(E\left(\widehat{\beta}\right), V_{\widehat{\beta}}\right)$$

It has been observed, however, that frequentist confidence intervals based on the naive use of $\widehat{\beta}$ and the corresponding covariance matrix perform quite poorly in terms of realized coverage probability (Wood, 2006b). Thus, in practice, in additive models based on penalized regression splines, frequentist inference yields us to reject the null hypothesis too often. To overcome this problem, and following Wahba (1983) and Silverman (1985), Wood (2006a,b) has implemented a Bayesian approach to coefficient uncertainty estimation. This strategy recognizes that, by imposing a particular penalty, we are effectively including some prior beliefs about the likely characteristics of the correct model. This can be translated into a Bayesian framework by specifying a prior distribution for the parameters β. Specifically, Wood (2006b) shows that using a Bayesian approach to uncertainty estimation results in a Bayesian posterior distribution of the parameters

$$\beta|\mathbf{y} \quad \sim \quad \mathcal{N}\left(E\left(\widehat{\beta}\right), \sigma_\varepsilon^2 \left(\mathbf{X}'\mathbf{X} + \sum_k \theta_k \mathbf{S}_k\right)^{-1}\right)$$

This latter result can be used directly to calculate credibility intervals for any parameter. Moreover, the credibility intervals derived via Bayesian theory are well behaved also from a frequentist point of view, i.e. their average coverage probability is very close to the nominal level $1-\alpha$, where α is the significance level.

A crucial issue in the use of penalized regression splines within an additive semiparametric model is the selection of the smoothing parameters, θ_k, controlling the trade–off between fidelity to the data and smoothness of the fitted spline. How should these values be selected? There are two main approaches to identify the optimum smoothing parameters. First, we can use prediction error criteria, such as generalized cross validation (GCV), Akaike information criterion (AIC), Bayesian information criterion (BIC) and so on. Alternatively we can rewrite the penalized additive model as a mixed model by decomposing each smooth term into fixed effect and random effect components and estimate the model by ML or restricted maximum likelihood (REML), treating θ_k as variance parameters (see Sect. 3.2).

As for the first method, we might select the values of $\widehat{\theta}_k$ that minimize the GCV score:

$$GCV(\theta_k) \quad = \quad \frac{n \parallel \mathbf{y} - \mathbf{H}(\theta_k)\mathbf{y} \parallel^2}{[n - tr(\mathbf{H}(\theta_k))]^2} \tag{9}$$

where $\mathbf{H}(\theta_k) = \mathbf{X}(\mathbf{X}'\mathbf{X} + \sum \theta_k \mathbf{S}_k)^{-1}\mathbf{X}'$ is the hat matrix for the model being fitted and its trace, $tr(\mathbf{H}(\theta_k))$, gives the effective degrees of freedom edf (i.e. the number of identifiable parameters in the model). The edf are a general measure for the complexity of a function estimates, which allows us to compare the smoothness, even for different types of effects (e.g. nonparametric versus parametric effects). If θ_k=0, then edf is equal to the size of the β vector minus the number of constraints (i.e. $edf = K$). Positive values of θ_k lead to an effective reduction of the number of parameters (i.e. $edf < K$). If θ_k is high, we have very few edf.

Actually, multiple smoothing parameter selection based on the minimization of the GCV score (9) is often too computationally demanding. To overcome this problem, Wood (2000) extended the 'performance iteration' method proposed by Gu and Wahba (1991) for automatically select multiple smoothing parameters to the case of computationally efficient low–rank additive models based on penalized regression splines. First, the multiple smoothing parameter model fitting problem is re-written with an extra *overall* smoothing parameter (δ) controlling the trade-off between model fit and overall smoothness, while retaining smoothing parameters multiplying each individual penalty, which now control only the relative weights given to the different penalties. The following steps are then iterated: (1) given the current estimates of the relative smoothing parameters (θ_k/δ), estimate the overall smoothing parameter; and (2) given the overall smoothing parameter, update $log(\theta_k)$ by Newton's method. In this way, the smoothing parameters for each smooth term in the model are chosen simultaneously and automatically as part of the model fitting. A drawback of this method is that it does not allow users to fix some smoothing parameters and estimate others or to bound smoothing parameters from below. Moreover, the method is not optimally stable numerically.

More recently, Wood (2004) proposed an improved (optimally stable) version of the 'performance iteration' method which is more robust to collinearity or concurvity problems and which can deal with fixed penalties. The second issue is very important when fully automatic smoothing parameter selection result in one or more model terms clearly over-fitted and thus it is necessary to fix or bound smoothing parameters. The issue is also particularly relevant in geoadditive models, since the smooth function of spatial location ($h(no_i, e_i)$), which enter the model as a nuisance term – i.e. only to explain variability that cannot be explained by the covariates that are really of interest – is often estimated with bounded smoothing parameters, while the 'interesting' terms are left with free smoothing parameters: in this way the 'interesting' covariates can be forced to do as much of the explanatory work as possible.

3.2 Penalized Regression Splines as Mixed Models and the REML estimator

The estimation of model (2) can be based on the reparameterization of such a model in the form of a mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \varepsilon \quad \mathbf{U} \sim \text{ i.i.d. } N(\mathbf{0}, \mathbf{G}) \quad \varepsilon \sim \text{ i.i.d. } N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \tag{10}$$

where $\mathbf{G}$ is a block–diagonal matrix, which depends on both $\sigma_{u_k}^2$ and σ_ε^2 variances. This model is a mixed model where β represents the parameters vector of fixed part and $\mathbf{U}$ are the random effects. The smoothing parameters are defined by the ratios $\theta_k = \frac{\sigma_\varepsilon^2}{\sigma_{u_k}^2}$. Again, matrix $\mathbf{X}$ may include parametric components such as the intercept, continuous covariates and categorical covariates.

This reparameterization consists in postmultiplying $\mathbf{X}$ and premultiplying β 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, Scheipl, and Faraway, 2012). Therefore, the type of penalizations determines the transformation matrix and, thus, the fixed and random effects obtained in the mixed model. The resulting coefficients associated with the fixed effects

(β) 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 $\mathbf{y} = \sum_{k=1}^{K} f(x_k) + \varepsilon$ (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 (10) becomes a variance components model that can be estimated by using standard software on the topic. When the model contains interaction terms, it is not longer a pure additive model. Therefore, each block $\mathbf{G}_k$ depends on more than one smoothing coefficient θ_k, except in the isotropic case,[6] where coefficients θ_k are the same for all variables (Wood, Scheipl, and Faraway, 2012; Lee and Durbán, 2011). As a consequence, the resulting mixed model is not an orthogonal variance component model.

A recent reparameterization, proposed by Wood, Scheipl, and Faraway (2012), 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 by using only one smoothing coefficient for each term. An alternative reparameterization from a P–Spline approach with a B–Spline basis and penalization matrices for the basis coefficients based on discrete differences is considered in Eilers and Marx (1996) and Lee and Durbán (2011). Two other interesting reparameterizations are based on (i) a truncated polynomial basis and ridge penalizations (Ruppert, Wand, and Carroll, 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, 2003b). The last three alternatives cannot be estimated with standard software on mixed models when the interactions between the variables are considered (except in the isotropic case).

Once the mixed model is defined, the parameters associated to fixed (β) and random effects (θ_k and σ_ε^2) can be estimated by using a ML algorithm. If the noise term follows a Gaussian distribution, the log–likelihood function is given by:

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

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

However, 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 in the model (McCulloch, Searle, and Neuhaus, 2008). In this case the objective function to maximize is given by:

$$\begin{aligned} \log L_{\mathrm{R}}(\theta_1, \cdots, \theta_K, \sigma_\varepsilon^2) &= \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \\ &\quad -\frac{1}{2}\mathbf{y}'\left(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right)\mathbf{y} \end{aligned}$$

[6] For the sake of clarity, isotropy means that the degree of smoothness is the same for all the covariates, that is the degree of flexibility in all of them is the same. Nevertheless, the usual situation in real cases is anisotropy, since the covariates are usually measured in different units of measure or, in the case of equal measurement units (e.g. spatial location variables), the variability of such covariates differs greatly.

An estimation of the variance components parameters can be obtained after maximizing $\log L_{\mathrm{R}}(.)$. In a second step, the estimates of β and $\mathbf{U}$ are given by (McCulloch, Searle, and Neuhaus, 2008):

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{U}} &= \hat{\mathbf{G}}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})\end{aligned}$$

Finally, the estimated values of the observed variable can be obtained as:

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

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, as shown previously in the GCV method, $\mathbf{H}$ is the hat matrix of the model (Ruppert, Wand, and Carroll, 2003). For the mixed model, it can be proved 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 approximation 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.

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

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

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

In reduced form we have:

$$\mathbf{y} = \mathbf{A}\mathbf{X}\beta + \mathbf{A}\mathbf{Z}\mathbf{U} + \mathbf{A}\varepsilon \tag{11}$$

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

As pointed out in Montero, Mínguez, and Durbán (2012) and Mínguez, Durbán, Montero, and Lee (2012), the log–REML function for model (11) is:

$$\begin{aligned}\log L_{\mathrm{R}}(\rho,\theta_1,\cdots,\theta_K,\sigma_\varepsilon^2) &= \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \log|\mathbf{A}| \\ &\quad -\frac{1}{2}\mathbf{y}'\mathbf{A}'\left(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right)\mathbf{A}\mathbf{y}\end{aligned} \tag{12}$$

As usual, $\log L_{\mathrm{R}}(.)$ is maximized with respect to the parameter vector $(\theta_1,\cdots,\theta_K,\sigma_\varepsilon^2)'$. Note that the maximization process requires the computation of the log-determinant of matrix $\mathbf{A}$, a dense $n\times n$ inverse matrix depending on ρ. As a consequence, the maximization of such a function

constitutes a challenging task. Nevertheless, to evaluate $\mathbf{A}$ for different values of ρ when n is large, it is possible to use Monte Carlo procedures (LeSage and Pace, 2009).

With respect to the estimation of PS-SEM model, we could also maximize the log–REML function (12) but, now, the covariance matrix is given by:

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

and $\mathbf{B} = I - \lambda \mathbf{W}_n$. Now, the maximization problem is also challenging due to the difficulty of the inversion of $\mathbf{V}$ matrix.

Finally, fixed and random effects can be estimated as:

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

Unlike model (10), the spatial lag model (11) cannot be estimated by using standard software regardless of the type of reparameterization used to express it as a mixed model.[7]

3.4 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 ε are correlated. In Sect. 3.3 we have described a possible solution to this endogeneity bias based on the REML estimation approach. As suggested by Basile (2009), an alternative way of dealing with the simultaneity bias in PS-SAR is the “control function" (CF) approach (Blundell and Powell, 2003).[8]

Generally speaking, 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 ($\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$), where the endogenous explanatory variables appear linearly, 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 estimates of the first-stage errors $\mathbf{v}$ (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 by IV approaches. However, in these cases the CF approach offers some distinct advantage over the IV methods (Wooldridge, 2007). In

[7] Nevertheless, there are some *R* codes using *spdep* package available from Montero, Mínguez, and Durbán (2012) and Mínguez, Durbán, Montero, and Lee (2012).

[8] It is important to mention that a semiparametric spatial lag model has also been proposed within a partial linear framework. For example, 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.

particular, the application of the standard 2-SLS fitted-value 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 this 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

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

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

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

$$\begin{aligned}\mathbf{y} = {} & \mathbf{X}^{*'}\beta^* + \rho\mathbf{W}_n\mathbf{y} + f_1(x_1) + f_2(x_2) \\ & + f_3(x_3, x_4) + f_4(x_1)l + ... + h(no, e) + c(\widehat{\nu}) + \varepsilon \end{aligned} \tag{13}$$

Obviously, the endogeneity of any other continuously distributed regressor in the PS-SAR model can also addressed via the control function 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 standard errors. This procedure may consist of the following steps:

1. Select 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 each endogenous variable on the exogenous variables and the instruments;
3. Insert the first-step residuals in the original semiparametric regression;
4. Repeat $B = 1000$ times points (*i*)–(*iii*);
5. For each estimated parametric coefficients compute the corresponding equal–tail bootstrap p-value:
$$P^*(\hat{\beta}) = 2 \times min\left(\frac{1}{B}\sum_{b=1}^{B}\#\{\hat{\beta}_b^* \leq 0\}, \frac{1}{B}\sum_{b=1}^{B}\#\{\hat{\beta}_b^* > 0\}\right)$$
6. For each estimated nonparametric coefficients compute the average partial effect at the 95% confidence bands.

[9] Both first and second step equations can be estimated by using, for example, PLS or REML estimators.

[10] The requirement that the endogenous regressor be continuously distributed is the most important limitation of the applicability of the CF approach to estimation of nonparametric and semiparametric models with endogenous regressors.

4 An Application to Lucas County House Pricing Data

We investigate the performance of the *semiparametric spatial autoregressive geoadditive models* (PS-SAR and PS-SEM) 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 modelling housing prices. In Section 4.2 we report the results of the analysis.

4.1 Data and model specififfcation issues

Lucas County (Ohio) database on housing prices contains 18,378 observations of single family homes sold during 1995-1998, and is fully described in the Spatial Econometrics toolbox for $Matlab^{TM}$ (*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 econometrics models in the *R spdep* package with those in the Spatial Econometrics toolbox for $Matlab^{TM}$, in *OpenGeoDa* and in the $STATA^{TM}$ *ado* file *sppack*. Zhu, Füss, and Rottke (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 multidirectional spatial effect in the estimation process.

In all these applications, hedonic equations for single-family homes are estimated mainly 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 cube term), the logarithms of the lot size and of the total living area in square feet, and numbers of rooms, bathrooms and bedrooms. Unfortunately, the dataset does not contain information on various neighbourhood amenities and proximity variables. The list of neighbours 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 neighbours per observation on average (Figure 1).

Figure 1 about here

Before discussing the results of our analysis, it is important to remark that modelling housing markets is a complex task because of the simultaneous occurrence of strong spatial dependence, strong heterogeneity and strong nonlinearities. Empirical evidence regarding *spatial dependence* 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 neighbourhood as reference price. Thus, the price of one house influences the prices of other houses located nearby and vice versa (Can, 1997). 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 neighbouring agents. Furthermore, information flows and expectations are likely to reinforce horizontal transmissions between agents which, in addition to commuting and migration, favour the appearance

of strong dependence (Brady, 2011; Holly, Hashem Pesaran, and Yamagata, 2011; Kuethe and Pede, 2011).

Spatial heterogeneity is also very common in housing price analysis. First, geographical constraints and urbanization regulations restrict the supply side of the market in a different way from zone to zone. Second, search costs and differences in the channels transmitting price information create information asymmetries across space (Dieleman, Clark, and Deurloo, 2000; Gray, 2012; van Dijk, Franses, Paap, and van Dijk, 2011). Third, Wood (2003a) points to differences in liquidity across different spatial markets; combined with a different capacity to absorb national shocks, these differences result in powerful potential sources of heterogeneity. Forth, the disparities can be accentuated by differences in the patterns of spatial mobility, in terms of commuting and migration (Kosfeld, 2007; Molloy, Smith, and Wozniak, 2011).

Finally, it must be recognized that the nature of the relationship between house prices and the various associated attributes is complex and nonlinear, so it would be better represented by nonparametric models rather than the classical parametric specifications (Ekeland, Heckman, and Nesheim, 2004). For example, Goodman and Thibodeau (1995) suggest that housing depreciation, that is the relationship between dwelling age and the market value of owner-occupied housing, is nonlinear and possibly non-monotonic. These three issues (spatial dependence, spatial heterogeneity and nonlinearities) clearly raise the need to modelling housing prices by using flexible PS-SAR and PS-SEM specifications.

4.2 Econometric results

We use the subset of Lucas County housing price data referring to the 1996 year to compare the performance of different competing parametric and semiparametric models. The starting point is the pure a-spatial parametric model, relating the logarithm of house price to the age of the house, its squared term, the logarithms of the lot size and of the total living area in square feet, and the number of bathrooms. OLS coefficients of this model are all significant and have the expected sign (Table 1). Moreover, a quadratic effect of age is clearly detected.

Table 1 about here

As observed above, no contextual variables about the neighbourhood of the houses are available in the dataset, so one would expect a strong spatial autocorrelation reflecting this misspecification. This expected result is clearly corroborated by the Lagrange Multiplier tests for spatial autocorrelation in OLS residuals, which strongly work in favour of the SAR model.[11] Direct, indirect and total effects from the estimated parametric SAR are reported in Table 2.

Table 2 about here

[11] This is not a new finding. As it is well known from the literature on the topic, the conventional hedonic regression model (the A-spatial model) is not capable to capture spatial dependence on house prices (even when location variables are included). However, we claim that the absence of contextual variables must engender a range of spatial processes which could not be fully captured by the spatially lagged dependent variable. It may also be the case that the very sparse spatial weights used are insufficiently dense to mop up the existent autocorrelation. To account for these critical issues, we estimate more flexible semiparametric PS-SAR and PS-SEM models.

The estimate of the spatial correlation coefficient ($\widehat{\rho}$) in the parametric SAR model is 0.466, which clearly indicates that houses prices in Lucas County are highly spatially dependent. Obviously, this large $\widehat{\rho}$ traduces in a dramatic reduction of mean square error (MSE), Akaike information criterion (AIC), Bayesian information criterion (BIC) statistics: MSE reduces by 36.4%, and AIC and BIC comes down by 6.5% (Table 3). Something similar occurs when the spatial correlation is considered in the error term and, consequently, the spatial strategy is a parametric SEM. In this case, the estimate of the spatial correlation coefficient is $\widehat{\lambda} = 0.512$, and the inclusion of the autoregressive term in the error reduces the MSE by 31.3% and both AIC and BIC by 5.4%. It is of note that the spatial coefficient in SEM exceeds that of SAR, which could be attributed to the omission of relevant variables.

Table 3 about here

Comparing parametric and semiparametric models, we firstly observe that the a-spatial nonparametric model outperforms its parametric counterpart, although it performs worse than the parametric SAR and SEM models. As can be seen in Table 3, the a-spatial model specified using cubic regression spline basis functions and the GCV score to identify smoothing parameters reduces the MSE by 16.6% and the AIC and BIC statistics by 2.5% and 2.2%, respectively. These results are even better when using P-spline basis functions and estimating the model with REML: MSE reduces by 19.3%, AIC and BIC fall by 3.0 and 2.7%, respectively.

Significant gains in model performance are observable once the geoadditive component is included in the model, which highlights the importance of controlling for unobserved spatial heterogeneity. In particular, it is worth noticing that the P-Spline Geoadditive model estimated with RMEL provides similar MSE, AIC and BIC statistics as the pararametric SAR (the best of the spatial parametric specifications). These statistics are even better when using cubic splines and carrying out a GCV estimation.

Although the inclusion of a spatially autocorrelated term in the traditional parametric specification yields significant gains, the benefits gained from semiparametric spatial autoregressive models are certainly greater. More specifically, the MSE of the PS-SAR model is about 0.11 (the MSE of the parametric SAR is 0.14), irrespective of the type of spline basis function used and the estimation method adopted (i.e. one-stage REML method or two-stage control function approach). It is also worth noticing that the estimate of the ρ parameter decreases from 0.47 for the parametric SAR model to 0.32-0.38 for the nonparametric PS-SAR specifications. This is a expected result, since part of the spatial dependence is captured by the spatial trend surface. Surprisingly, some further improvement in terms of model fitting is obtained with the PS-SEM. This last finding can be attributed to the short number of explanatory variables included in the model. The spatial coefficient λ in the PS-SEM (0.446) is lower than in the parametric SEM (0.521), again probably because part of the spatial dependence is captured by the spatial trend surface. All in all, our results clearly display the superiority of semiparametric spatial geoadditive models (PS-SAR and PS-SEM) at least for this case study.

Finally, having shown the superiority of semiparametric spatial geoadditive models, we briefly discuss the results of the PS-SAR model estimated using the two-stage control func-

tion approach (Table 4).[12] First, we run a semiparametric regression of the endogenous term $W_n y$ on the exogenous variables and their spatial lags used as external instruments. Then, we insert the first-stage residuals in the original semiparametric regression to correct the inconsistency of the regression of the dependent variable y on the endogenous explanatory variable $W_n y$. All terms, but $W_n y$, $bath$ and $W_n bath$ are introduced as smooth terms. The model also includes a spatial trend surface, $h(no, e)$, constructed by using the spatial coordinates in rescaled form. All smooth terms are specified using P-spline basis functions. Both stages are estimated using the REML method.

Table 4 about here

Second-stage results show that all smooth terms have an edf higher than 1, indicating that not only age, but also $log(lotsize)$ and $log(livingarea)$ enter nonlinearly the model. This is clearly displayed in Figure 2, which reports the plots of estimated additive smooth components along with the total, direct and indirect effects computed using equations (4), (5) and (6).[13] The pointwise 95% confidence bands (obtained using the bootstrap procedure described in Section 3.4) show that all effects are also significant in most part of the variable domain. As expected, indirect effects are always lower that direct ones. Finally, a picture of the unobserved spatial heterogeneity captured by the geoadditive component — $h(no, e)$ — is reported in Figure 3.

Figures 2 and 3 about here

5 Conclusions

In this paper we have reviewed recently developed *semiparametric spatial autoregressive geoadditive models*, called Penalized Spline Spatial Lag model (PS-SAR) and Penalized Spline Spatial Error model (PS-SEM) model. These methods play a prominent role in those context in which the theory suggests the existence of spatial interdependence and heterogeneous behavior of the spatial units. We have showed their relative performance with respect to parametric a-spatial and spatial regression models using a large dataset on house prices. Natural directions in which these methods can be extended are a specification for longitudinal data and, eventually, a dynamic framework.

[12] As observed above, the performances of the PS-SAR estimated with the one-step REML method are very similar. Thus, the evidence reported for the two-stage control function approach can be safely extended to those with the one-step REML method.

[13] Actually, total, direct and indirect effects are not smooth at all 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.

References

ANSELIN, L. (1988): *Spatial econometrics: methods and models*, vol. 4. Kluwer Academic Pub.

ANSELIN, L. (2003): "Spatial Externalities, Spatial Multipliers and Spatial Econometrics," *International Regional Science Review*, 26, 153–166.

ARBIA, G., AND J. PAELINCK (2003): "Spatial Econometric Modeling of Regional Convergence in Continuous Time," *International Regional Science Review*, 26, 342–362.

AUGUSTIN, N., M. MUSIO, K. V. WILPERT, E. KUBLIN, S. WOOD, AND M. SCHUMACHER (2009): "Modeling Spatio-Temporal Forest Health Monitoring Data," *Journal of the American Statistical Association*, 104, 899–911.

AZOMAHOU, T., J. E. OUARDIGHI, P. NGUYEN-VAN, AND T. PHAM (2011): "Testing Convergence of European Regions: A semiparametric approach," *Economic Modelling*, 28, 1202–1210.

BASILE, R. (2008): "Regional Economic Growth in Europe: a Semiparametric Spatial Dependence Approach," *Papers in Regional Science*, 87, 527–544.

——— (2009): "Productivity Polarization across Regions in Europe: The Role of Nonlinearities and Spatial Dependence," *International Regional Science Review*, 32, 92–115.

BASILE, R., R. CAPELLO, AND A. CARAGLIU (2012): "Technological Interdependence and Regional Growth in Europe," *Papers in Regional Science*, 91, 697–722.

BASILE, R., C. DONATI, AND R. PITTIGLIO (2013): "Industry Structure and Employment Growth: Evidence from Semiparametric Geoadditive Models," Mimeo.

BASILE, R., AND B. GRESS (2005): "Semi-parametric Spatial Auto-covariance Models of Regional Growth Behavior in Europe," *Region et Developement*, 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.

——— (2012): "After Raising the Bar: applied maximum likelihood estimation of families of models in spatial econometrics," *Estadística Española*, 54, 71–88.

BLUNDELL, R., AND J. POWELL (2003): *Advances in Economics and Econometrics Theory and Application*chap. Endogeneity in Nonparametric and Semiparametric Regression Models. Cambridge University Press.

BOURASSA, S. C., E. CANTONI, AND M. HOESLI (2010): "Predicting house prices with spatial dependence: a comparison of alternative methods," *Journal of Real Estate Research*, 32(2), 139–159.

BRADY, R. R. (2011): "Measuring the diffusion of housing prices across space and over time," *Journal of Applied Econometrics*, 26(2), 213–231.

BRUECKNER, J. K. (2000): "Urban sprawl: diagnosis and remedies," *International regional science review*, 23(2), 160–171.

BRUECKNER, J. K., E. MILLS, AND M. KREMER (2001): "Urban Sprawl: Lessons from Urban Economics [with Comments]," *Brookings-Wharton papers on urban affairs*, pp. 65–97.

CAN, A. (1997): "Specification and Estimation of Hedonic Housing Price Models," *Regional Science and Urban Economics*, 22, 453?74.

CAN, A., AND I. MEGBOLUGBE (1997): "Spatial dependence and house price index construction," *The Journal of Real Estate Finance and Economics*, 14(1-2), 203–222.

CHASCO YRIGOYEN, C., AND J. LE GALLO (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.

DIELEMAN, F. M., W. A. CLARK, AND M. C. DEURLOO (2000): "The geography of residential turnover in twenty-seven large US metropolitan housing markets, 1985-95," *Urban Studies*, 37(2), 223–245.

DUBÉ, J., AND D. LEGROS (2013): "Dealing with spatial data pooled over time in statistical models," *Letters in Spatial and Resource Sciences*, 6, 1–18.

EILERS, P., AND B. MARX (1996): "Flexible Smoothing with B-Splines and Penalties," *Statistical Science*, 11, 89–121.

EKELAND, I., J. HECKMAN, AND L. NESHEIM (2004): "Identification and estimation of hedonic models," *Journal of Political Economy*, 112, 60–109.

ERTUR, C., AND J. L. GALLO (2009): *Handbook of Regional Growth and Development Theories*chap. Regional Growth and Convergence: Heterogenous Reaction versus Interaction Spatial Econometric Approaches, pp. 374–388. Edward Elgar, Cheltenham.

FISCHER, M., AND P. STUMPNER (2010): *Handbook of Applied Spatial Analysis*vol. 4, chap. Income distribution dynamics and cross-region convergence in Europe, pp. 599–628. Springer, Berlin, Heidelberg and New York.

FOTHERINGHAM, A., C. BRUNSDON, AND M. CHARLTON (2002): *Geographically Weighted Regression*. Wiley, Chichester.

FOTOPOULOS, G. (2012): "Nonlinearities in Regional Economic Growth and Convergence: the Role of Entrepreneurship in the European Union Regions," *The Annals of Regional Science*, 48, 719–741.

GOODMAN, A. C., AND T. G. THIBODEAU (1995): "Age-related heteroskedasticity in hedonic house price equation," *Journal of Housing Research*, 6(3), 25–42.

——— (2003): "Housing market segmentation and hedonic prediction accuracy," *Journal of Housing Economics*, 12(3), 181–201.

GRAY, D. (2012): "District house price movements in England and Wales 1997–2007: An exploratory spatial data analysis approach," *Urban Studies*, 49(7), 1411–1434.

GRESS, B. (2004): "Using Semi-Parametric Spatial Autocorrelation Models to Improve Hedonic Housing Price Prediction," Mimeo. Department of Economics, University of California.

GRIFFITH, D., AND J. PAELINCK (2011): *Non-standard spatial statistics and spatial econometrics*, vol. 1. Springerverlag Berlin Heidelberg.

GU, C., AND G. WAHBA (1991): "Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method," *SIAM Journal on Scientific and Statistical Computing*, 12(2), 383–398.

HOLLY, S., M. HASHEM PESARAN, AND T. YAMAGATA (2011): "The spatial and temporal diffusion of house prices in the UK," *Journal of Urban Economics*, 69(1), 2–23.

IRWIN, E. G., AND N. E. BOCKSTAEL (2007): "The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation," *Proceedings of the National Academy of Sciences*, 104(52), 20672–20677.

KELEJIAN, H., AND I. PRUCHA (1997): "Estimation of Spatial Regression Models with Autoregressive Errors by Two-Stage Least Squares Procedures: A Serious Problem," *International Regional Science Review*, 20, 103–111.

KIM, S.-W., AND R. BHATTACHARYA (2009): "Regional housing prices in the USA: an empirical investigation of nonlinearity," *The Journal of Real Estate Finance and Economics*, 38(4), 443–460.

KNEIB, T., T. HOTHORN, AND G. TUTZ (2009): "Variable selection and model choice in geoadditive regression models," *Biometrics*, 65(2), 626–634.

KOSFELD, R. (2007): "Regional Spillovers and Spatial Heterogeneity in Matching Workers and Employers in Germany," *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)*, 227(3), 236–253.

KUETHE, T. H., AND V. O. PEDE (2011): "Regional housing price cycles: a spatio-temporal analysis using US state-level data," *Regional Studies*, 45(5), 563–574.

LAMBERT, D. M., W. XU, AND R. J. FLORAX (2013): "Partial Adjustment Analysis of Income and Jobs, and Growth Regimes in the Appalachian Region with Smooth Transition Spatial Process Models," *International Regional Science Review*.

LEE, D., AND M. DURBÁN (2011): "P-Spline ANOVA Type Interaction Models for Spatio-Temporal Smoothing," *Statistical Modelling*, 11, 49–69.

LEE, L.-F., X. LIU, AND X. LIN (2010): "Specification and estimation of social interaction models with network structures," *The Econometrics Journal*, 13(2), 145–176.

LESAGE, J., AND K. PACE (2009): *Introduction to Spatial Econometrics*. CRC Press, Boca Raton.

MCCULLOCH, C., S. SEARLE, AND J. NEUHAUS (2008): *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics, Chichester, 2nd edn.

MCMILLEN, D. P. (1996): "One hundred fifty years of land values in Chicago: a nonparametric approach," *Journal of Urban Economics*, 40(1), 100–124.

MÍNGUEZ, R., M. DURBÁN, J. MONTERO, AND D. LEE (2012): "Competing Spatial Parametric and Non-Parametric Specifications," Mimeo.

MOLLOY, R., C. L. SMITH, AND A. K. WOZNIAK (2011): "Internal Migration in the United States," Working Paper 17307, National Bureau of Economic Research.

MONTERO, J., R. MÍNGUEZ, AND M. DURBÁN (2012): "SAR Models with Nonparametric Spatial Trends. A P-Spline Approach," *Estadística Española*, 54, 89–111.

MUR, J., F. LÓPEZ, AND A. ANGULO (2010): "Instability in spatial error models: an application to the hypothesis of convergence in the European case," *Journal of geographical systems*, 12(3), 259–280.

PACE, K., AND J. LESAGE (2004): *Spatial Econometrics and Spatial Statistics*chap. Spatial Autoregressive Local Estimation, pp. 31–51. Palgrave Macmillan, Basingstoke.

PINHEIRO, J., AND D. BATES (2000): *Mixed-effects Models in S and S-PLUS*. Springer-Verlag, New York.

RUPPERT, D., M. WAND, AND R. CARROLL (2003): *Semiparametric Regression*. Cambridge University Press, Cambridge.

SILVERMAN, B. (1985): "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting," *Journal of the Royal Statistical Society. Series B*, 47, 1–53.

SU, L. (2012): "Semiparametric GMM Estimation of Spatial Autoregressive Models," *Journal of Econometrics*, 167, 543–560.

SU, L., AND S. JIN (2010): "Profile Quasi-Maximum Likelihood Estimation of Partially Linear Spatial Autoregressive Models," *Journal of Econometrics*, 157, 18–33.

VAN DIJK, B., P. H. FRANSES, R. PAAP, AND D. VAN DIJK (2011): "Modelling regional house prices," *Applied Economics*, 43(17), 2097–2110.

WAHBA, G. (1983): "Bayesian Confidence Intervals for the Cross Validated Smoothing Spline," *Journal of the Royal Statistical Society. Series B*, 45, 133–150.

WAND, M. (2003): "Smoothing and Mixed Models," *Computational Statistics*, 18, 223–249.

WOOD, R. (2003a): "The information content of regional house prices: can they be used to improve national house price forecasts?," *Bank of England Quarterly Bulletin, Autumn*.

WOOD, S. (2003b): "Thin Plate Regression Splines," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65, 95–114.

——— (2004): "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686.

——— (2006a): *Generalized Additive Models. An Introduction with R*. Chapman and Hall, London.

——— (2006b): "On Confidence Intervals for Generalized Additive Models based on Penalized Regression Splines," *Australian and New Zealand Journal of Statistics*, 48, 445–464.

——— (2011): "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73, 3–36.

WOOD, S., F. SCHEIPL, AND J. FARAWAY (2012): "Straightforward Intermediate Rank Tensor Product Smoothing in Mixed Models," *Statistics and Computing*, In press. DOI=10.1007/s11222-012-9314-z.

WOOD, S. N. (2000): "Modelling and smoothing parameter estimation with multiple quadratic penalties," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 413–428.

ZHU, B., R. FÜSS, AND N. ROTTKE (2011): "The Predictive Power of Anisotropic Spatial Correlation Modeling in Housing Prices," *Journal of Real Estate Finance and Economics*, 42, 542?565.

TABLE 1
OLS estimation results

	Estimate	Std. Error	t value	Pr(>|t|)
$(Intercept)$	3.5371	0.1804	19.61	0.0000
age	1.0626	0.0922	11.53	0.0000
age^2	-1.9078	0.0739	-25.81	0.0000
$log(lotsize)$	0.1701	0.0099	17.26	0.0000
$log(livingarea)$	0.8353	0.0250	33.44	0.0000
$baths$	0.0475	0.0190	2.49	0.0127
Lagrange Multiplier Spatial Dependence Tests on OLS Residuals				
	Statistic	p-value		
LM_err	1064.9	0.000		
LM_lag	1552.6	0.000		
Robust-LM_err	4.5	0.034		
Robust-LM_lag	492.2	0.000		

Notes: Dependent variable: log of house price. * , ** and *** indicate significance at the 10, 5 and 1% levels, respectively. LM_err = H_0: OLS and H_1: SEM; LM_lag = H_0: OLS and H_1: SAR.

TABLE 2
Direct, Indirect an Total impacts from the parametric SAR model

Variable	Direct	Indirect	Total
age	1.028***	0.710***	1.738***
age^2	-1.472***	-1.016***	-2.488***
$log(lotsize)$	0.092***	0.063***	0.155***
$log(livingarea)$	0.691***	0.477***	1.169***
$bath$	0.013	0.009	0.023

Notes: * , ** and *** indicate significance at the 10, 5 and 1% levels, respectively.

TABLE 3
Model comparison

Model	Method	MSE	AIC	BIC	EDF	Rho	Lambda
		Parametric models					
A-spatial	OLS	0.217	6.960	6.968	6.000		
SAR	ML	0.138	6.505	6.513	6.000	0.466	
SEM	ML	0.149	6.583	6.591	6.000		0.521
		Semiparametric models					
A-spatial	P-Splines, REML	0.175	6.751	6.779	20.138		
A-spatial	CR-Splines, GCV	0.181	6.783	6.813	23.008		
Geoadditive	P-Splines, REML	0.137	6.519	6.585	49.047		
Geoadditive	CR-Splines, GCV	0.129	6.463	6.554	68.041		
PS-SAR	P-Splines, REML	0.110	6.293	6.347	39.687	0.354	
PS-SAR	CR-Splines, GCV, CF	0.111	6.316	6.404	65.835	0.324	
PS-SAR	P-Splines, REML, CF	0.112	6.316	6.380	48.183	0.379	
PS-SEM	P-Splines, REML	0.108	6.274	6.327	39.485		0.446

Notes: OLS = Ordinary Least Squares; ML = Maximum likelihood; GCV = Generalized Cross Validation; REML = Restricted maximum Likelihood; CF = Control Function; CR-Splines = Cubic Regression Splines. The number of knots used for the smooth terms $f_1(age)$, $f_2(log(lotsize))$, $f_3(log(living.area))$ and $h(no,e)$ are 8, 10, 10 and 8, respectively.

TABLE 4
Control function estimates of the semiparametric PS–SAR Model

	First stage	Second stage
Parametric terms	*Estimate (Bootstrap p-value)*	
(Intercept)	11.019	6.806
	(0.000)	0.000
$W_n y$		0.379
		0.000
$baths$	0.012	0.043
	(0.420)	0.000
$W_n baths$	0.011	
	(0.500)	
Smooth terms	*edf*	*edf*
$f_1(age)$	5.238	6.643
$f_2(log(lotsize))$	4.216	4.978
$f_3(log(livingarea))$	1.888	3.728
$h(no,e)$	38.342	24.986
$f_4(res)$		4.848
$f_4(W_n age)$	6.467	
$f_5(W_n log(lotsize))$	5.256	
$f_6(W_n log(livingarea))$	6.637	
$R^2_{adj.}$	0.826	0.811
Deviance explained	82.9%	81.2%
REML score	1194	1686

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 $f_1(age)$, $f_2(log(lotsize))$, $f_3(log(living.area))$, $h(no,e)$, $f_4(W_n age)$, $f_5(W_n log(lotsize))$, $f_6(W_n log(living.area))$, are 8, 10, 10, 8, 8, 10, and 10 respectively.

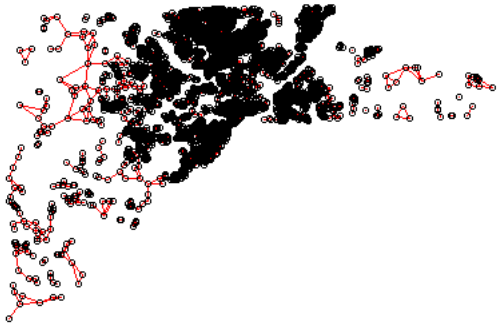

FIGURE 1
Sphere of influence graph

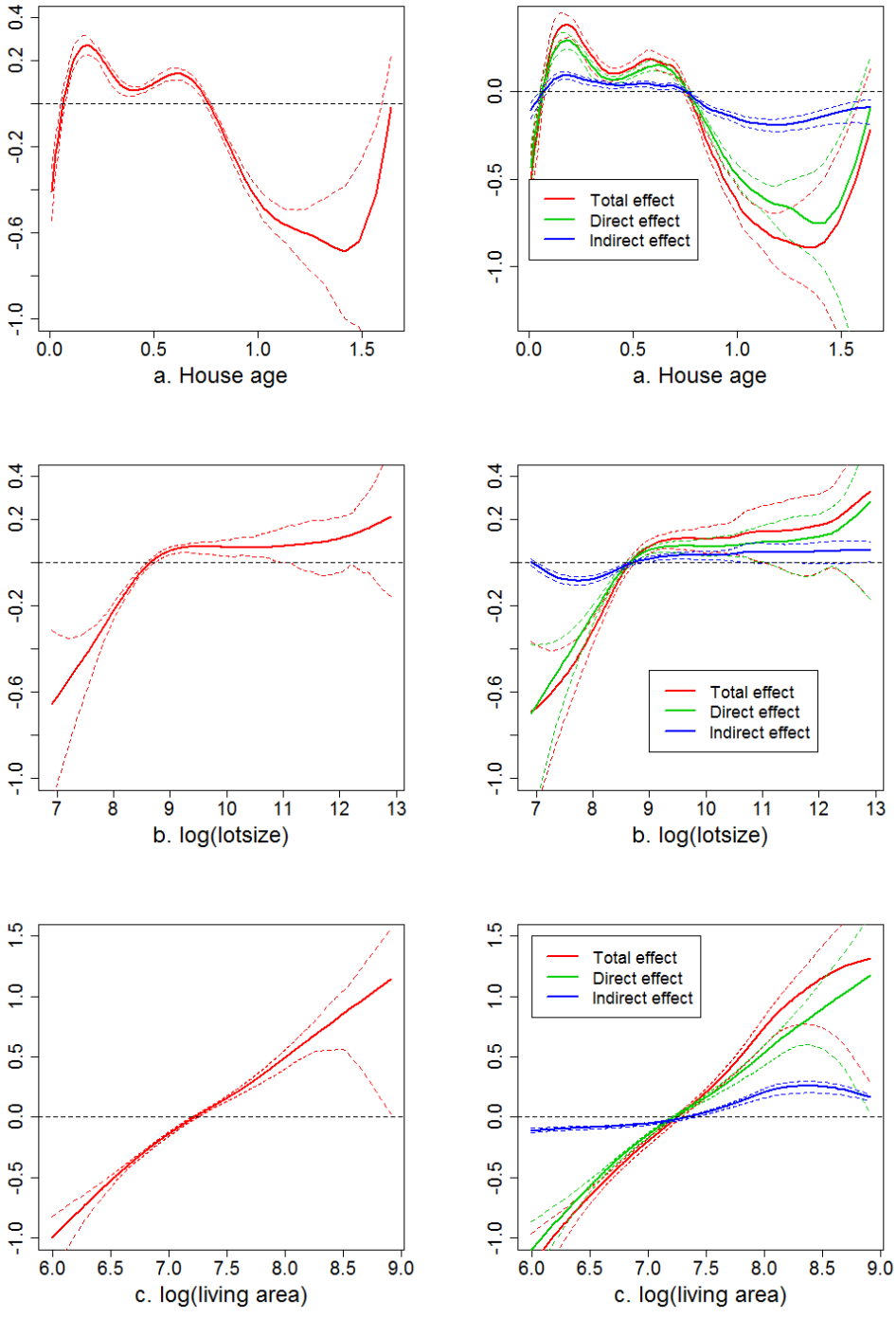

FIGURE 2
Estimated Effect (Left panels) and Direct, Indirect and Total Effects (Right Panels) - PS–SAR

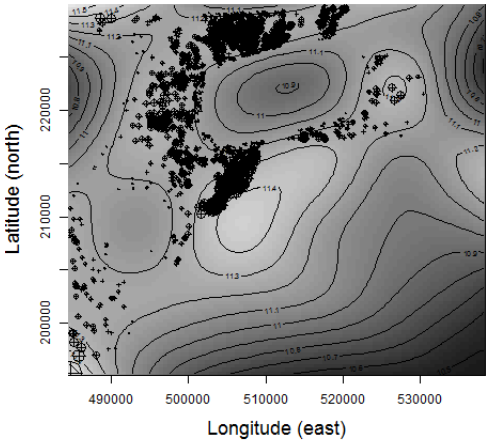

FIGURE 3
Spatial trend surface