Functional regression over irregular domains

Functional regression over irregular domains: Variation in the shadow price of living space

Arnab Bhattacharjee[#] Liqian Cai[+] Taps Maiti[+#] *

December 1, 2016

Abstract

Hedonic house price models need to account for spatial heterogeneity – the variation in the functional surface of shadow prices. In this context, the complexity of spatial domains raises issues for the traditional spatial econometric methods. Specifically, discontinuities in the spatial surface need to be accounted for, including irregular boundaries, peninsulas and interior holes. Motivated by an application to housing markets, we develop a method for estimating the functional surface of a regression coefficient that varies over such an irregular spatial domain. Spatially varying coefficients for a specific regressor are estimated by a combination of three smoothing problems using splines based on finite element analysis. The effect of additional regressors is also allowed. We verify finite sample performance using a simulation study, and develop an application to the Aveiro-Ílhavo urban housing market in Portugal.

Keywords: Hedonic pricing model; Delaunay triangulation; Finite element analysis; Housing markets; Spatial regression; Spline smoothing.

*Correspondence: a.bhattacharjee@hw.ac.uk. # Spatial Economics and Econometrics Centre, Heriot-Watt University, UK. + Department of Statistics and Probability, Michigan State University, East Lansing MI, USA.

We thank the the Co-Editor and the Editor-in-Chief for their encouragement, and two anonymous referees for many constructive comments and suggestions that helped us revise and improve upon the paper. Sincere thanks are also due to organisers and participants of Spatial Statistics 2013 Conference (Columbus, Ohio, USA) and 14th International Workshop on Spatial Econometrics and Statistics 2015 (Paris, France), and in particular, Giuseppe Arbia, Bernie Fingleton, Raymond Florax and Abhimanyu Gupta for their constructive comments that helped formulate and shape our work. The usual disclaimer applies.

1 Introduction

The topic of this paper is modeling spatial heterogeneity in hedonic house price models in an urban context with complex spatial domain. Here complex or irregular spatial domain refers to contexts where there are irregular boundaries, peninsulas and/or interior holes, and therefore Euclidean distances are not meaningful measures of the true distance between locations within the domain. Hedonic regression is a popular revealed preference method for estimating demand or value for a heterogenous good. In particular, housing is a very heterogenous commodity, with value depending potentially on hundreds of characteristics, including internal features and dimension, external facilities, access to infrastructure and services, and proximity to local facilities and the urban labour market. Hence, hedonic house price models have been useful for assessing the value of housing, and are frequently used for real estate appraisal and valuation, housing economics, urban planning and in the construction of housing and consumer price indices (Maclennan, 1977; Malpezzi, 2003; Banerjee et al., 2004).

However, housing prices are also subject to another fundamental issue, that they are explained by many determinants in micro-geographic and structural characteristic space that are not all observable to the analyst (Cheshire and Sheppard, 1995). This implies that hedonic regressions are usually affected by unobserved spatial heterogeneity. Simplistic applications of hedonic models are therefore problematic because spatial variation in many applications is rather complex. Hence, estimates of hedonic house price models have been criticized for lack of adequate attention to heterogeneity of slopes and complexity of the spatial domain over which they are estimated. Adequate modelling requires the adoption of complex econometric tools, which allow us to deal with important methodological issues, such as spatial dependence, spatial heterogeneity and nonlinearities (Basile et al.. 2015). This paper focuses specifically on the latter two issues in a context where, in addition to complex spatial variation, the spatial domain is also irregular.

Spatial heterogeneity implies that the slope parameter corresponding of a regressor potentially varies over the spatial domain. In housing markets, the "shadow price" of living space usually varies over space.[1] Further, this

[1] Typically, a hedonic house price model is estimated in double logarithmic form, where logarithm of house prices are regressed on the logarithm of living space, logarithm of distance to the centre, and other hedonic characteristics. In this case, the "shadow price" of living space is measured by the corresponding regression coefficient, which has the interpretation of living space elasticity of house price. Then the coefficient measures the number of percentage points by which the price of a house would increase, if there were a 1 percent increase in living space.

variation is related to the spatial structure of the domain, combining both topography and structure of the built environment. In turn, this spatial structure can be quite complex and needs to be explicitly modeled. Price per unit area is generally higher in the centre of an urban area, where jobs, public transport and civic facilities are concentrated, and becomes less expensive towards the periphery, where the patterns of such decrease are closely related to road networks, the structure of housing development, and more generally the urban geography of the city.

To deal with spatial heterogeneity and spatial structure, as well as spatial dependence, several complex econometric models have been proposed in the literature. In particular, in an excellent recent review, Basile et al. (2015) propose the very general semiparametric spatial autoregressive geoadditive models. There are also two other closely related alternatives to multiple regression in this context: locally weighted regression (McMillen and Redfearn, 2010) and a specific type of spatial kernel regression called geographically weighted regression (GWR) (Fotheringham *et al.*, 2002).

However, these methods are not readily available when the spatial domain is complex, with irregular boundaries, peninsulas and interior holes.[2] Locally weighted regressions or GWR modeling requires the use of very complex spatial kernels, and not simply a kernel truncated across the spatial discontinuities. Similar considerations apply to spatial spline methods where the spline function has to take explicit account of irregularities in the spatial domain. Further, existing methods often do not incorporate an explicit spatial model, which renders the underlying assumption of independent and identically distributed regression errors particularly suspect. At the same time, there have been major recent advances in methods for spatial smoothing in such situations. Basile et al. (2015) consider a very general geoadditive spatial model where nonlinearities are modeled in a semiparametric manner using spatial splines. The methodological contribution in this paper lies in taking one of these spatial spline smoothing methods and adapting this to the context of regression with heterogenous slopes and spatial fixed effects. The possibility of using spatial kernels in combination with geodesic distances

[2]Let us consider GWR as an example. If there is an interior hole, the underlying kernel can be modified to take account of this. Typically, one may use a truncated kernel. However, this approach is not suitable if the discontinuity is narrow but relatively long. In the application considered in this paper, the spatial domain is divided by a river that has only a single connection across it (a bridge). Then, a truncated kernel is not appropriate for locations close to the banks of the river (see Figures 1 and 3), or consider locations $(3, -0.1)$ and $(3, 0.1)$ in our simulation study (Figure 2a). One needs a kernel that places opposite banks of the river at substantial distance to each other, because the only way to go from one bank to the other is to go along the river all the way upto the bridge, and then back again along the other bank of the river.

proposed in the above literature also holds substantial promise; this lies in the domain of future work.

Together, several approaches have been proposed in the recent literature for spatial smoothing on irregular domains; see, for example, Ramsay (2002), Eilers (2006), Wang and Ranalli (2007), Wood et al. (2008) and Sangalli et al. (2013). In this paper, we address a related but distinctly different problem. Like the above literature, our spatial domain is complex, but unlike recent studies, our inference problem relates to spatial semiparametric regression, and not to spatial smoothing. We develop statistical methodology for estimating a functional regression model (Ramsay and Silverman, 2005, 2006) where the effect of one specific regressor varies over the spatial domain (Majumdar et al., 2006; Bhattacharjee et al., 2016). In addition, there may be other regressors with spatially fixed coefficients. The proposed estimator is related to spatial smoothing methods developed in Ramsay (2002) and Wang and Ranalli (2007), but extended to the context of spatial functional regression. Locally weighted regression provides an alternate approach to spatial smoothing (McMillen, 2010). However, current methods in this literature rely too strongly on kernels based on Euclidean distances, which do not take the complexity of the spatial domain fully into account (Wang and Ranalli, 2007; O'Donnell et al., 2014). Though not the main focus of this paper, our method can also be adapted to local linear or polynomial estimation.

The remainder of the paper is organized as follows. Section 2 discusses issues of irregular spatial domain, followed in section 3 by discussion of the current literature on spatial smoothing over irregular domains and our specific modeling context. Section 4 develops the proposed method to estimate the functional surface of a spatially varying coefficient. Section 5 reports on a simulation study, and section 6 applies the proposed methods to the urban housing market of the twin municipalities of Aveiro and Ílhavo in central Portugal. Finally, section 7 concludes.

2 Specificities and Issues of Irregular spatial domain

Our method is motivated by a study of an urban housing market in Portugal: the Aveiro-Ílhavo housing market in central Portugal. Figure 1 shows the spatial domain, including the twin municipalities of Aveiro and Ílhavo and the adjoining peri-urban and rural area. We can see that the spatial domain is complex, including irregular boundaries, peninsulas and interior holes. Specifically, it includes coastal areas to the west that are separated

Figure 1: The Aveiro-Ílhavo housing market

from the centre by a large lagoon (Ria de Aveiro). The central area (central business district, CBD) is the region with the highest density of submarkets, the centroids of each submarket being denoted by a small circle. Two branches of water extend from the lagoon. The western branch extends southwards and westwards to the Atlantic Ocean, trisecting the landmass into three parts. There is limited connection of the beach areas (the strips of land on the western margin) to the mainland – specifically, a bridge connects to one area of the coast (the southern part), and a ferry crossing connects the other (the northern beach). The northern beach is also connected to the mainland by a road which circumvents the lagoon to the north; this road connection is very long[3] and hence does not constitute an alternate transport connection. The other southward extension of the lagoon (towards the centre in Figure 1) is a narrow canal traversed by many bridges, and therefore does not increase the complexity of the spatial domain in any significant way; see Bhattacharjee et al. (2012, 2016) for more discussion on this specific urban housing market.

Our application is based on a spatial functional regression model where

[3]The road link to the centre of Aveiro is 55 km long, while the Euclidean distance ("as the crow flies") is only about 6 km, and the road distance from the ferry crossing to the centre of Aveiro is about 10 km.

the dependent variable is logarithm of house prices in the Aveiro-Ílhavo housing market and the regressors are hedonic housing features. The central object of inference are the partial effects of these housing features, and these are expected to vary over the irregular two-dimensional territory. Initially, our focus lies on a single regressor, logarithm of living area, so that the functional regression coefficient $\beta(s)$ is interpreted as the price elasticity of living area. Generally, $0 \leq \beta(s) < 1$; however the elasticity can exceed unity in small localities where living space is at a very large premium, and likewise can perhaps fall to negative values in small areas where large living space is either unavailable or undesired; we discuss this issue later in the context of our application.

Estimation of this functional regression model is important for identifying housing submarkets and inferring on spatial dependence. In the housing studies literature, submarkets have been defined either by similarity in hedonic housing attributes (Rothenberg et al., 1991) or by substitutability of housing units (Pryce, 2013). In the context of a hedonic house price model with homogenous slopes, where logarithm of house prices are regressed on a collection of hedonic attributes, the two definitions are equivalent.[4] However, this is not true when there are spatially varying coefficients (spatial heterogeneity). This heterogeneity is not necessarily in hedonic characters, but more importantly in the shadow prices (coefficients) assigned to such features. Assuming that house prices (in logarithms) are determined by a single feature (say, living area) and that the partial effect of this feature is potentially heterogeneous over the territory, the principle of substitutability will imply that two locations i and j are in the same submarket if $\partial y_i/\partial y_j = 1$.This implies that submarkets should be delineated by clustering jointly on the surface of the functional partial effect $\beta(s)$ and the surface of the spatially varying hedonic feature (Bhattacharjee et al., 2016). In reality, however, some other factors that affect house prices through fixed coefficients should also be taken into account, in order to eliminate potential omitted variable bias and inefficiency in inferences on the partial effects of primary interest.

[4] A hedonic regression model (Rosen, 1974) is a popular revealed preference method of estimating demand or value of a heterogenous product. The prices of a product (for example, a house) is regressed on a host of constituent characteristics or attributes, called hedonic attributes. The estimated regression coefficients can be interpreted as (functions of) implicit or shadow prices of these attributes.

3 Spatial smoothing over irregular domains

3.1 Literature on spatial smoothing

While the literature on spatial regression on irregular domains is rather limited, the same is not true for spatial smoothing. In the recent literature, particular focus has been placed on smoothing over irregular spatial domains which present considerable challenges in many real applications. Well-known methods for spatial data analysis, such as kriging, thin plate splines and kernel smoothing are difficult to apply in these circumstances. This is mainly because these methods do not take into account the shape of the domain, and therefore cannot smooth around the discontinuities or across concavities in the domain. To deal with such challenges, the literature has proposed several useful approaches for smoothing a functional random variable over a complex spatial domain. These include: Delaunay triangulations combined with quadratic or linear smoothers (Ramsay, 2002), domain morphing (Eilers, 2006), low-rank smoothing splines based on within-area distance (Wang and Ranalli, 2007) and soap film smoothers (Wood et al., 2008). Sangalli et al. (2013) extended the approach in Ramsay (2002) to a semi-parametric model with regressors that have spatially fixed coefficients. O'Donnell et al. (2014) investigate flexible regression models based on kernel methods and penalized splines.

The above methods take a nonparametric or functional regression approach, where the response random variable is a nonlinear (nonparametric) function of x- and y-coordinates (denoted a and b, respectively). However, different contributions in this literature differ significantly in how the specific irregularities or discontinuities of the region are taken explicitly into account. Following Ramsay (2002), the standard model is a spatial spline smoothing function of the form:

$$y_i = f_y(s_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the data are of the form $\{(s_1, y_1), (s_2, y_2), \ldots, (s_n, y_n)\} \subset \Omega \times \mathbb{R}$ generated over a bounded but irregularly shaped spatial domain $\Omega \subseteq \mathbb{R}^2$. Here $s_i = (a_i, b_i)$ is the point on the spatial domain at which the function value y_i is observed. The ε_i are independent errors with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$, and $f(.)$ is a sufficiently smooth function from $\Omega \to \mathbb{R}$. The object of inference, $f(.)$, is estimated by $\widehat{f}(.)$ which minimizes the penalized sum-of-squares functional

$$\sum_{i=1}^{n} [y_i - f_y(s_i)]^2 + \lambda_y \int_{\Omega} (\Delta f_y)^2 \tag{2}$$

subject to some appropriate boundary conditions. λ_y is a positive smoothing parameter which controls the trade-off between the fit and smoothness. The penalty is defined in terms of a Laplacian operator Δ, where Δ is defined by $\Delta f_y = f_{aa} + f_{bb}$ for all f_y to make sure that the penalty is invariant to both rotation and translation. Estimation of $f_y(.)$ is conducted by first partitioning the spatial domain into disjoint triangulations, and then constructing a separate (quadratic) polynomial function on each piece of the partition, using finite element analysis, in such a way that the union of these quadratic polynomials approximates the solution. To be specific, an arbitrary quadratic polynomial g on one piece of partition triangulation $\mathbf{t}_k$ can be written as

$$g(a,b) = \sum_{i=1}^{n} g(\mathbf{n}_{ik})\psi_{ik}(a,b)$$

where

$$\psi_{ik}(\mathbf{n}_{jk}) = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

and here $\mathbf{n}_{ik}$ is the i-th node on the k-th partition. Note that each triangular partition has 6 nodes which are the vertices and edge midpoints. A quadratic polynomial in two variables (the location coordinates a and b) is uniquely determined by its value at six distinct points, so the definition above can define a unique and continuous surface on the target spatial domain.

The model is extended in Sangalli et al. (2013) to include regressors with spatially fixed coefficients. Let $w_i = (w_{1i}, w_{2i}, \ldots, w_{qi})'$ be a q-vector of covariates associated with observed response y_i. Then the semiparametric model for the data is

$$y_i = w_i'\gamma + f_y(s_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

Now, the spatially fixed coefficient vector γ and the surface f is estimated by minimizing the penalized sum-of-squares functional

$$\sum_{i=1}^{n} [y_i - w_i'\gamma - f_y(s_i)]^2 + \lambda_y \int_{\Omega} (\Delta f_y)^2 . \tag{4}$$

Besides the advantage of accounting for additional covariates, this method can also admit different restrictions on the boundary of the domain. The focus of the above methods is to explicitly take into account the spatial structure of the irregular domain Ω. This is achieved by partitioning the spatial domain by Delaunay triangulations.

Another way to deal with the irregular spatial structure is to measure the intrinsic similarity (or dissimilarity) between two locations on the domain by

a suitable construction of distances (Wang and Ranalli, 2007). In place of the traditional Euclidean distance, they introduce a new measure called geodesic distance between points on the spatial domain. It measures the distance only over the domain and around the gaps and estimates the surface based on Low-Rank Thin Plate Splines. An alternate approach is based on flow-weighted distances (O'Donnell et al., 2014).

3.2 Our modeling context

The above modeling frameworks and methodologies are not directly applicable to contexts of the kind considered in this paper. First, our spatial context is markedly distinct, comprising spatial subregions whose boundaries are not necessarily known *a priori*. Second, our interest lies in estimating the functional surface of the spatially varying effect, $\beta(s)$, of a regressor $X(s)$, as opposed to smoothing a function over a complex spatial domain, or estimating a spatially fixed regression coefficient. Note that, from an econometric point of view, an irregular spatial domain may not be a problem, so long as the complexity of the domain is modeled appropriately using suitable complex kernels or splines. However, these methods would typically be very difficult to implement because such suitable kernels or splines are very complex. The approach taken in this paper is computationally simpler and may be viewed as an exploration as to how far methods of spatial smoothing over irregular domains can be useful in a spatially varying regression context.

In our housing market application, house prices are recorded at several locations $(s_i, i = 1, \ldots, n)$, together with explanatory variables like living space and proximity to the city centre. Our spatial hedonic house price model has logarithm of house prices per square meter of living space as the dependent variable (y_i), logarithm of living space (x_i) as the leading explanatory variable, and other housing characteristics, including proximity to the centre, as additional regressors (w_i); see Anselin et al.. (2010) and Bhattacharjee et al. (2016) for further discussion on spatial hedonic pricing models. Clearly, smoothing methods appropriate for models discussed above are inadequate here. This is a regression context with spatially varying coefficients, where the house price elasticity of living space, the regression coefficient β, is expected to vary over the spatial domain. In addition, we might expect variation over space in unobserved land prices, which can be modeled by spatial fixed effects $f(s_i)$. Because of the above reasons, we propose the following regression model:

$$y_i = w_i'\gamma + x_i\beta(s_i) + f(s_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where the data $\{(s_1, y_1, x_1, w_1), (s_2, y_2, x_2, w_2), \ldots, (s_n, y_n, x_n, w_n)\}$ are gen-

erated over the spatial domain $\Omega \subseteq \mathbb{R}^2, s_i = (a_i, b_i) \in \Omega$. The slope parameter $\beta(.)$ on the main regressor x is allowed to vary over space; hence it has potentially different coefficients $\beta(s_i)$ at different locations s_i within the spatial domain Ω. In addition, there are also spatial fixed effects $f(s_i)$ that potentially vary over the spatial domain. While the main regression coefficient and spatial fixed effects are allowed to vary over space, the other regressors (w) are assumed to have fixed coefficients. This semiparametric feature constitutes a major benefit of the proposed approach, where some regressors are allowed to have spatially varying coefficients, while the others have fixed coefficients.

In the recent literature, spatial heterogeneity (or spatially varying coefficients) has been modeled using spatial splines (Basile et al., 2015), locally weighted regressions (McMillen, 2010; McMillen and Redfearn, 2010) or Geographically Weighted Regression (GWR) (Fotheringham et al., 1998, 2002). The main advantage of the above methods over spatial smoothing is allowing the regression coefficient to vary over space, and to provide methods for estimating a smooth surface of spatially varying coefficients. However, there is no simple way to account for the irregular or complex nature of the spatial domain, for which we draw upon the recent literature on spatial smoothing.

Inclusion of spatial fixed effects $f(s_i)$ modeled in a flexible nonparametric way is a key innovation in our approach. These fixed effects capture unmeasured location (neighbourhood) features that can potentially be correlated with the included regressors. This ensures that the assumption of independent identically distributed errors in our model (5) is defensible. More importantly, $f(s_i)$ also capture the effect of potentially non-uniform sampling over the spatial domain. Such sample selection effects are typically modeled by including a spatial inverse Mill's ratio (Heckman, 1979), which in this case is encompassed in the fixed effects. Hence, whereas reasonable statistical properties of locally (geographically) weighted regressions under in-fill asymptotics are very difficult to obtain (Lahiri, 1996), our methods do not suffer equally from this limitation because sampling variation is already modeled.

This paper places spatial smoothing and spatial heterogenous slope regression within a unified framework. Specifically, our interest lies in estimating the effect of a regressor with potentially spatially varying coefficients; the object of inference is to understand such spatial variation within a complex and irregular spatial domain. In addition, we allow for location-specific fixed effects and other covariates with fixed coefficients.

4 Functional regression over irregular domains

Now, we develop inferences for a functional regression model on an irregular spatial domain, where a functional regression coefficient varying over the domain is the main object of inference. Starting from spatial spline smoothing methods developed in Ramsay (2002) and Sangalli et al. (2013), we extend this to the estimation of a functional regression model.

We first discuss the intuition behind our approach, then describe the proposed method for a single regressor with functional coefficient, and finally extend this to the context of the semiparametric model (5).

4.1 Intuition

Initially, consider a simple linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

For estimation of the slope coefficient β by (ordinary) least squares (OLS), we first eliminate the intercept by writing this model in the form of deviations from mean:

$$(y_i - \overline{y}) = \beta (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}).$$

Under the assumption that errors are homoscedastic, the OLS estimator follows. However, more generally, one can rewrite the model as

$$\frac{y_i - \overline{y}}{x_i - \overline{x}} = \beta + \frac{\varepsilon_i - \overline{\varepsilon}}{x_i - \overline{x}}.$$

A consistent estimator for β can then be obtained by taking weighted averages of the ratio of deviations $(y_i - \overline{y}) / (x_i - \overline{x})$, ensuring that observations with zero (small) deviations $(x_i - \overline{x})$ correspondingly receive zero (small) weights.

For the one way fixed effects model, a similar approach follows. In this case, deviations are computed from group means, rather than the overall means, $\overline{y}$ and $\overline{x}$. This is exactly the fixed effects or within transformation. In the panel data case, one takes deviations of y and x from their respective group means. This removes the group-specific fixed effects from the model; see, for example, Hsiao (1986, section 3.2) or Wooldridge (2002, section 10.5.1). Further, if the slopes were potentially different across the groups, the averages need to be separately computed for each group. Then, the ratio of the deviations recovers the slope for the specific group, and this approach allows for slope heterogeneity.

The context of this paper is somewhat different. We do not have separate groups, but the slope and intercept potentially vary over the spatial

domain, reflecting spatially varying coefficients and spatial fixed effects respectively. This is where smoothing approaches are useful. Our proposed method is to first smooth the surface of y and x, obtaining smoothed surfaces $\widehat{y}(s)$ and $\widehat{x}(s)$ respectively. Next, we compute deviations from these smoothed values $(y-\widehat{y})_i$ and $(x-\widehat{x})_i$ respectively, and the ratio of deviations, $(y-\widehat{y})_i / (x-\widehat{x})_i$. Finally, the smoothed surface of this ratio provides estimates of the spatially varying coefficient $\beta(s)$. However, in each of these smoothing steps, we need to take into account the complex nature of the spatial domain.

The above demeaning approach does not hold for a nonlinear model. However, the way we view our model (5) is not as a nonlinear model but as a semiparametric model with an infinite dimensional parameter space that is locally linear in the parameters. Following Ramsay (2002), our interpretation of the nonparametric components $\beta(s)$ and $f(s)$ are in least square sense. In other words, one may view our approach to semiparametric functional estimation as a local histogram sieve approach.

4.2 Single regressor with functional coefficient

Consider the model (5) with only a single regressor x in a neighborhood h_s of $s \in \Omega$. Within this neighborhood, we consider or model as locally linear, with

$$y_t = f_s + \beta_s x_t + \varepsilon_t, \quad t \in h_s, s \in h_s \subset \Omega,$$

where f_s and β_s are unknown but fixed scalars. Then, we can apply a local smoother to compute the local mean of y and x, (locally) demean the two variables, then take the ratio, and finally apply the local smoother again to recover the slope coefficient β_s.

The underlying methods are based on an initial Delaunay triangulation of the irregular domain. In the case of Euclidean space with sampled point locations, as is the case here as well as in the spatial smoothing literature, Delaunay triangulation is the dual graph for a Voronoi tessellation applied to the same set of points. In turn, Voronoi tessellations are defined as follows. Let Ω be a space (a nonempty set) endowed with a distance function d . Let K be a set of indices and let $(s_k)_{k\in K}$ be a tuple of nonempty subsets (the locations) in the space Ω. The Voronoi cell, or Voronoi region, R_k, associated with the location s_k is the set of all points in Ω whose distance to s_k is not greater than their distance to the other locations s_j, where j is any index different from k. In other words, if $d(s,A) = \inf\{d(s,a)|a \in A\}$ denotes the distance between the location s and the subset A, then

$$R_k = \{s \in \Omega | d(s,s_k) \leq d(s,s_j) \text{ for all } j \neq k\}.$$

Having obtained the Delaunay triangulation of the complex and irregular spatial domain Ω, Ramsay (2002) proceeds with estimation of model (1). This is done by finite element analysis to numerically approximate the simplest bivariate L-spline function that minimizes the least squares criterion subject to a roughness penalty:

$$\widehat{f} = \underset{f \in H^2(\Omega)}{\arg\min} \sum_{i=1}^{n} \left(y_i - f_y\left(s_i\right)\right)^2 + \lambda_y \int_{\Omega} \left(\Delta f_y\right)^2 ds,$$

where $H^2\left(\Omega\right)$ is the space of all piecewise quadratic functions that match the values taken by f_y on the nodes of the Delaunay triangles. Sangalli et al. (2013) extend the method to the model (3) where there are additional regressors with spatially fixed coefficients, and where for simplicity, the finite element approximation is based on piecewise linear functions.

Our inference problem is different. As such, we are not interested in smoothing the surface of y or x, but seek to estimate the functional surface for the coefficient of a regressor under the model (5). However, as discussed above, there is a link between the regression problem and smoothing. Therefore, we measure y and x in terms of deviations from their locally smoothed values, $(y - \widehat{y})_i$ and $(x - \widehat{x})_i$ respectively. Conceptually, this is equivalent to measuring deviations from spatial fixed effects, which is achieved in the spatial smoothing literature by estimating the smoothed surface $\widehat{y}_i = \widehat{f}\left(s_i\right)$ in the context of model (1); we estimate $\widehat{x}$ similarly by smoothing the functional surface of x. We first apply spatial smoothing to y and x, and denote the smoothed values $\widehat{y}$ and $\widehat{x}$, respectively. Applying the fixed effects transformation, that is, taking deviations $(y - \widehat{y})$ and $(x - \widehat{x})$, we have then removed the spatial fixed effects $f\left(s_i\right)$. Now, when we consider the ratio $(y - \widehat{y}) / (x - \widehat{x})$, the right hand side has $\beta\left(s_i\right)$ as the spatial fixed effects intercept and no regressor. Therefore, we can now apply smoothing to this ratio to recover the spatially varying slope, $\beta\left(s_i\right)$.

Expressed in terms of the above deviations, without the spatially fixed coefficients, our regression model (5) is therefore transformed as:

$$\begin{aligned} \left(y - \widehat{y}\right)_i &= \left(x - \widehat{x}\right)_i \beta\left(s_i\right) + \left(\varepsilon - \widehat{\varepsilon}\right)_i \\ &\Rightarrow \left(\frac{y - \widehat{y}}{x - \widehat{x}}\right)_i = \beta\left(s_i\right) + \left(\frac{\varepsilon - \widehat{\varepsilon}}{x - \widehat{x}}\right)_i . \end{aligned} \tag{6}$$

This is similar to the spatial smoothing model (1), with the functional coefficient $\beta\left(s_i\right)$ now taking the role of a spatial fixed effect intercept $f\left(s_i\right)$, but with two important differences. First, the errors are now heteroscedastic, which implies that spatial smoothing estimates of $\beta\left(s_i\right)$ will not be efficient.

Second, the errors are also not independent, because the estimation of $\widehat{x}$ uses all the data. However, since $\widehat{x}$ is consistent, in large samples, when $\widehat{x}$ is close to the true smoothed surface $x(s)$, the effect of such dependence vanishes. Thus, $\beta(s_i)$ can be consistently estimated in large samples by spatial smoothing. In section 4, we verify finite sample performance of our proposed estimator.

Thus, our estimation proceeds as follows. In step 1, we obtain a representation of the irregular spatial domain using Delaunay triangulations. In step 2, based on the triangulation above, we use the method in Ramsay (2002) to obtain spatially smoothed surfaces for both $y(s)$ and $x(s)$, denoted by $\widehat{y}$ and $\widehat{x}$, respectively. For the smoothing in step 2, we use the following two penalised objective functions:

$$\widehat{y} = \arg\min \sum_{i=1}^{n} \left[y_i - f_y(s_i)\right]^2 + \lambda_y \int (\Delta f_y)^2 \, ds$$

and

$$\widehat{x} = \arg\min \sum_{i=1}^{n} \left[x_i - f_x(s_i)\right]^2 + \lambda_x \int (\Delta f_x)^2 \, ds.$$

Thus, we compute $(y - \widehat{y}) / (x - \widehat{x})$. Finally, the functional surface of β is estimated in step 3, where we apply Ramsay (2002) smoothing to this ratio using the following penalised objective function:

$$\widehat{\beta} = \arg\min \sum_{i=1}^{n} \left[\left(\frac{y - \widehat{y}}{x - \widehat{x}}\right)_i - \beta(s_i)\right]^2 + \lambda \int (\Delta\beta)^2 \, ds.$$

In implementing steps 2 and 3, we require three penalty parameters: λ_y, λ_x and λ. We set optimal values for these three roughness penalties by cross-validation (Härdle et al., 1988; Ramsay and Silverman, 2005).

Conceptually, our proposed estimation involves two main transformations. First, we compute deviations of y and x from spatially smoothed values, which is equivalent to fixed effects transformation. Second, smoothing the ratio of these deviations recovers the true underlying spatially varying regression coefficient. The proposed method suggests a natural extension to local least squares estimation that takes into account irregularities of the spatial domain. Starting from (6), an alternate estimator would constitute first smoothing $[(y - \widehat{y})(x - \widehat{x})]$ and $(x - \widehat{x})^2$, and then estimating $\beta(s_i)$ by local least squares as the ratio of the above two functional surfaces. The development of this estimator is retained for future research.

4.3 Multiple regressors

In our housing market example, the main focus is on the regressor living area (in logarithm) which is expected to have spatially heterogeneous effects on housing prices (in logarithm). In addition, we have six covariate measures, denoted w, that are assumed to have constant effects on price. Inclusion of additional regressors with spatially fixed coefficients requires only minor modifications to the above estimation procedure. In the presence of additional regressors w, we can use the method in Sangalli et al. (2013) in step 2 to obtain smoothed surfaces for y and x, in each case partialling out the dependence of w. In other words, in step 2, we can use the following objective functions:

$$\sum_{i=1}^{n} \left[y_i - w_i'\gamma_y - f_y\left(s_i\right)\right]^2 + \lambda_y \int \left(\Delta f_y\right)^2 ds$$

and

$$\sum_{i=1}^{n} \left[x_i - w_i'\gamma_x - f_x\left(s_i\right)\right]^2 + \lambda_x \int \left(\Delta f_x\right)^2 ds.$$

Then, step 3 follows in exactly the same way as the single regressor case.

Alternatively, one could account for additional regressors, first by smoothing the q regressors $(w_1, w_2, \ldots, w_q)$, and then using Sangalli et al. (2013) in step 3 including as additional regressors $\frac{w_1 - \widehat{w_1}}{x - \widehat{x}}, \frac{w_2 - \widehat{w_2}}{x - \widehat{x}}, \ldots, \frac{w_q - \widehat{w_q}}{x - \widehat{x}}$ with fixed coefficients. Which of these methods will work better is an issue that needs to be investigated. In our application in the following section, we used the second method, which is in direct alignment with our spatial hedonic model (5). However, there may be applications where the first method may be more useful, particularly when the surface of the additional regressors, w's, are not very smooth.[5]

5 Simulation study

Consistency of the proposed estimator derive from Ramsay (2002) and Sangalli et al. (2013), and hence we do not repeat the derivations here. In fact, our approach is simply a two-stage application of the smoothing splines in Ramsay (2002). In particular, they show consistency in the interior of the spatial domain and on the boundary. As described above, the method involves Delaunay triangulations that imply heterogenous sizes based on design density. However, this has no implications on consistency of the estimator,

[5] The intuitive idea is that in such cases, the second may be affected more explicitly by errors in smoothing the q regressors.

but greater precision in regions with higher design density; see also Ramsay (2002) and Sangalli et al. (2013). Efficiency can potentially be improved by applying a one-stage method, as in Basile et al. (2015); establishing such theoretical results are beyond the scope of the current paper. However, it is instructive to evaluate the performance of the estimator in finite samples.

Therefore, in this section, we compare finite sample performance of our proposed estimator for the functional surface of β with different sample sizes in a simulation study on a C-shaped domain (Figure 2a). The functional surface of β is the same as Wood et al. (2008), and a simple modification of the simulation presented in Ramsay (2002).

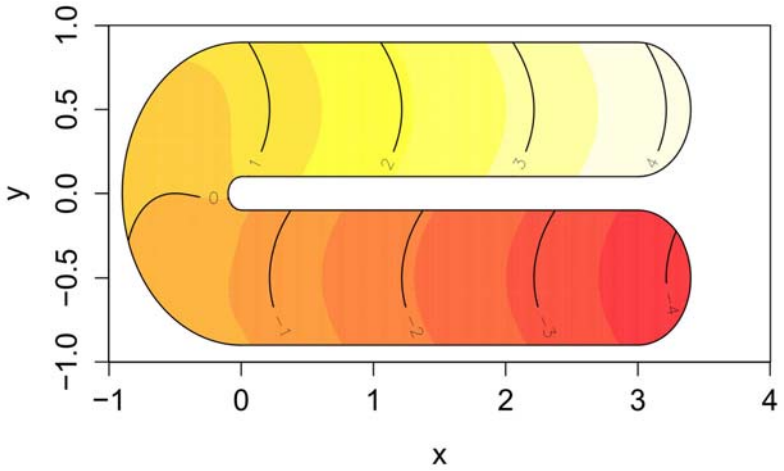
In the simulation, we randomly sample $n = 200$ locations, $s_1, s_2, \cdots, s_n$ from a mesh grid uniformly distributed within the C-shaped domain. For the i-th location sampled, we independently sample 3 covariates w_{i1}, w_{i2}, x_i from independent normal distributions $N(3, 2^2)$, $N(7, 2.5^2)$, and $N(6, 3^2)$, respectively. Further, we sample a random noise ε_i from $N(0, 0.5^2)$, and the response y_i at each location is obtained from

$$y_i = w_{i1}\gamma_1 + w_{i2}\gamma_2 + f(s_i) + x_i\beta(s_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

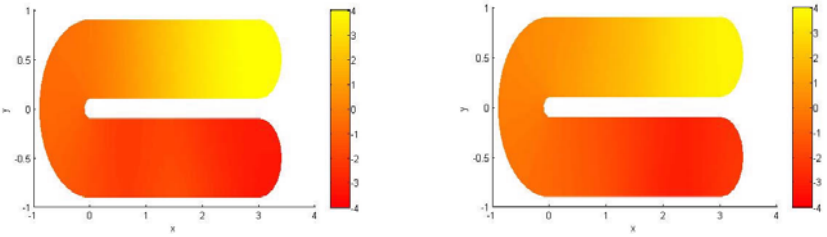
Since our focus is on the estimation of the spatially varying effect of x_i, we set the spatial fixed effect $f(s_i) = 3$ to be constant for simplicity. The other parameter values are set at $\gamma_1 = -1.5$ and $\gamma_2 = 1$. We estimate the functional surface $\beta(s_i)$ using the proposed method discussed above and the smoothing parameter is chosen based on 5-fold cross validation. The simulation is replicated $R = 30$ times, and the same procedure is repeated for a sample with sample size $n = 400$. The average estimated surface for $n = 200$ and $n = 400$ are shown in Figures 2b and 2c respectively.

The root-mean-square errors over the $R = 30$ simulation replicates are collected. The average RMSE when $n = 200$ is 0.317 while the average RMSE when $n = 400$ is 0.296. We can see that as the sample size becomes larger, the RMSE on average will gradually decrease, from which we infer the rate of convergence of our proposed estimator for the functional surface of β. The box-plots for RMSE are shown in Figure 2d. We conclude that the proposed estimator has satisfactory finite sample performance and expected asymptotic properties.

2a. C-shaped sampling domain and functional surface of $\beta(s)$



2b and 2c. Estimated average surface of $\beta(s)$, $n = 200, 400$ resp.



2d. Box-plots of RMSE: left ($n = 200$) and right ($n = 400$)

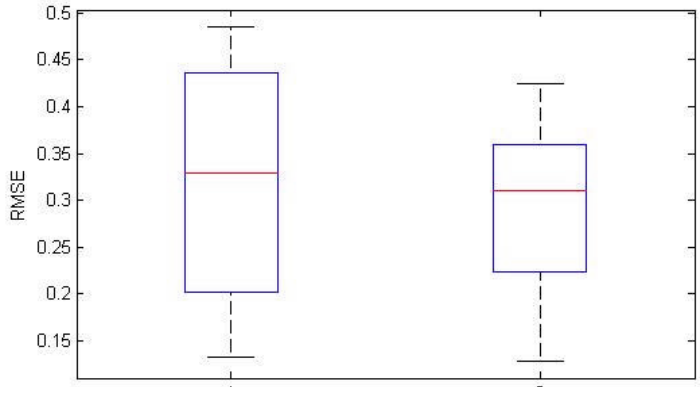


Figure 2: Results of the simulation study

6 An application to housing markets

The proposed framework and methods are applied to data on the housing market in the municipalities of Aveiro and Ílhavo and the adjoining peri-urban and rural area in central Portugal (Figure 1). As discussed above, the spatial domain is complex because it includes coastal areas that are separated from the centre by a large lagoon. There is limited connection of the beach areas to the mainland – specifically, a bridge connects to one area of the coast, and a waterway connects the other. Our interest lies in estimating the spatially varying surface of the implicit cost of living space (functional regression coefficient).

6.1 Data

The above urban housing market is located in the Centro Region of Portugal and includes two municipalities – Aveiro and Ílhavo. The municipality of Aveiro has a total area of 200 km^2 and had a total population of 78,454 in 2011 census; the municipality of Ílhavo has an area of 75km^2 and 38,317 inhabitants. Omitting the area of the lagoon, the population density is 600 inhabitants per km^2, which is typical for an urban agglomeration in Portugal.

The dataset used was provided by the firm Janela Digital S.A., which owns and manages the real estate portal database CASA SAPO. This portal is the largest website in Portugal for real estate advertisement. Data refers to the time period October 2000 to March 2010 and includes around 4 million records of properties available for transaction in Portugal, covering all the national territory. For the specific case of Aveiro and Ílhavo, 47,188 properties populated the database between 2000 and 2010. This empirical work uses 12,467 observations on completed transactions, after cleaning the data and removing all cases where data were incomplete. For further details and access to data, see Bhattacharjee et al. (2016).

Sale prices are not available, and therefore we consider listing price as our price variable, using time on the market to account for the wedge between listing and selling prices. Besides the price of the property, the dataset includes two main categories of variables for each house: i) the intrinsic physical attributes, and ii) the location and neighbourhood of the building. The first group includes number of rooms, state of restoration, age of construction and living area. A set of other physical housing characteristics were extracted from a free text field where real estate advertisers describe the property. The second group of attributes is related to the housing location and to the characteristics of the neighbourhood, data on which were constructed in previous work. For further details, refer to Bhattacharjee et al. (2012).

Our dependent variable is logarithm of house price per square meter of living area, and the regressor with spatial varying coefficients is the living space, measured as logarithm of square meters of living area. We include several other regressors in the estimated model, relating to proximity to the center (CBD) and local amenities, hedonic structural features of the house, and time on the market.

The hedonic characteristics (18 physical attributes of each house and 24 location attributes) are subjected to statistical factor analysis to extract 5 factors; see Bhattacharjee et al. (2012) for further details. Thus, apart from location fixed effects and (logarithm of) living space, the following six covariates are included and assumed to have spatially fixed coefficients: (logarithm of) time on the market (to account for the wedge between listing and sale prices); factor 1 (access to the centre or to central amenities); factor 2 (access to local services and amenities – health centres, parks/gardens, etc.); factor 3 (access to beaches, schools and local commerce); factor 4 (physical attributes of the house); and factor 5 (additional house facilities – garage, balcony, central heating, etc.). In principle, any of the additional covariates could also have spatially heterogenous slopes. Our methodology allows verification of the assumption of spatially fixed coefficients. In this specific application, the assumption appears to hold.

If the dependent variable were logarithm of house price and the regressor were logarithm of living space, the coefficient would be interpreted as the living space elasticity of price. However, in our case, the dependent variable is logarithm of price per square meter, and the relevant regressor is logarithm of living space (in squared meters). Hence the coefficient on the regressor may be interpreted as the corresponding elasticity, less one. Thus, in the following discussion, we add unity (one) to this coefficient to interpret it as the elasticity. There is an assumption of iid error terms underlying our methodology. This assumption can be strong. In our case, we have extensively modelled spatial heterogeneity and are therefore comfortable with this assumption.

Because only a reduced portion of houses are fully geo-referenced, they were located in the smallest homogeneous areas that the database can describe, and these centres were fully geo-referenced. Around 65 such regions (submarkets) have at least two observations, and are denoted by small circles in Figure 1. The boundary is irregular, and there is one big interior hole in the data – the lagoon region, and corresponding discontinuities between the beach areas and the mainland.

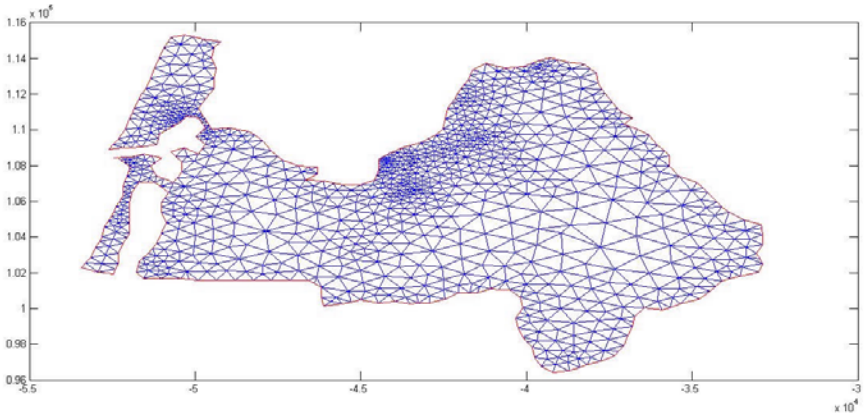


Figure 3: Delaunay Triangulation of the Aveiro-Ílhavo Housing Market

6.2 Application of the proposed methods

Using the method in section 3, first we obtain Delaunay triangulation of the spatial domain (Figure 3). The connections of the mainland to the beach areas, via the bridge and water network are clearly visible. Also, as expected, the triangles are dense in the centre (CBD) of Aveiro, as compared to sparsely populated rural areas in the periphery.

Next, we obtain smoothed surfaces of y (logarithm of price per square meter of living space) and x (logarithm of living space, in squared meters). For illustration, Figure 4 plots the smoothed surface of y. As might be expected, the highest prices are observed in the CBD of Aveiro, which is close to the employment centre, and the southern beach area, which is popular as holiday residences.

Following our methodology, the ratio of deviations of y and x from their locally smoothed values are then computed. Finally, the functional surface of the spatially varying regression coefficient is estimated (Figure 5), after adding the six spatially fixed covariates into the model. As discussed in section 3, we add additional regressors $\frac{w_1-\widehat{w_1}}{x-\widehat{x}}, \frac{w_2-\widehat{w_2}}{x-\widehat{x}}, \ldots, \frac{w_6-\widehat{w_6}}{x-\widehat{x}}$ in the place of w's in the Sangalli et al. (2013) model (3). This allows us to infer on spatial heterogeneity in the functional slope coefficient, after accounting for the effect of other regressors.

Figure 5 plots the estimated regression coefficient over the spatial domain.[6] The figure indicates the expected pattern of high implicit price of

[6]We need to add one to this coefficient to interpret it as the living space elasticity of price.

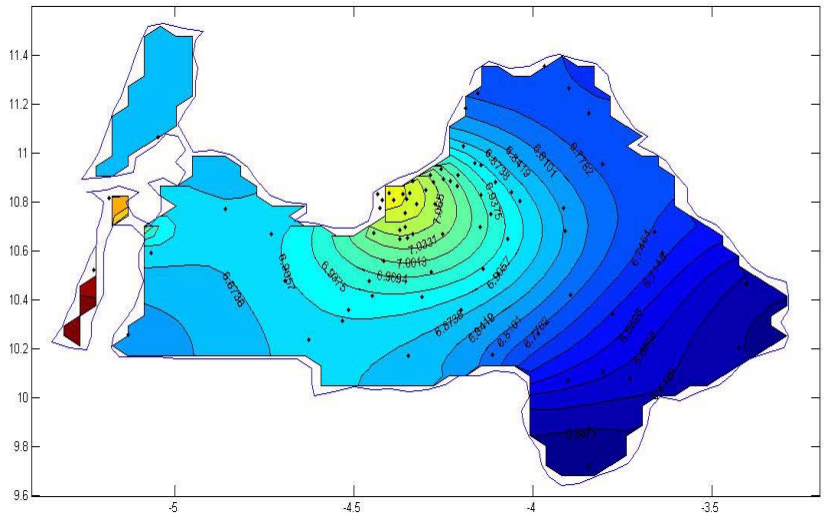


Figure 4: Smoothed surface of (log) price/ m^2

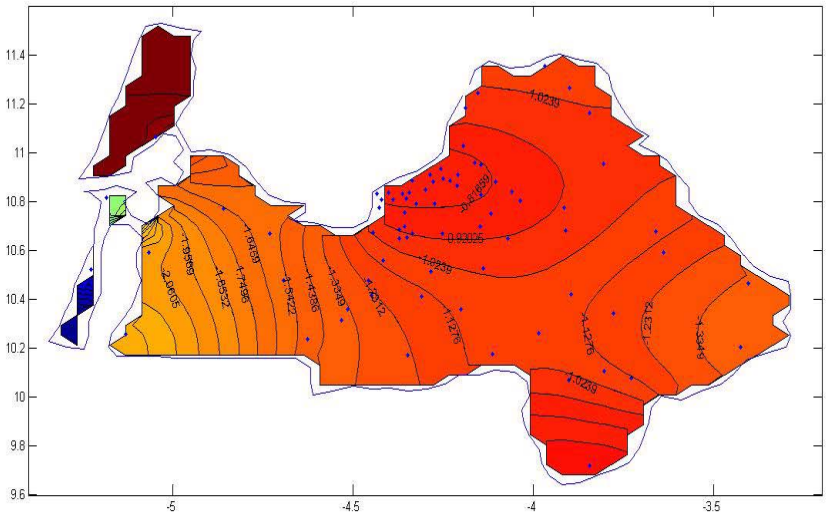


Figure 5: Implicit price (living space elasticity)
Spatially varying estimates (6 additional regressors)

living space in the centre of Aveiro, where population is dense and living space attracts a high premium. There, the estimated living space elasticity of price is in excess of 0.25, which implies that a 10% increase in living space would lead to a 2.5% rise in house price. The prices gradually decline towards the periphery along gradients that are determined by the major trunk roads converging on the centre. We can also see that the price of living space in the coastal area to the west has negative estimated effect on house prices; this issue is discussed in further detail below. The centre of the neighbouring municipality Ílhavo also has a relatively higher elasticity of around 0.10. Boundary spillages can be an issue, but its effect in this application appears to be small.

Our proposed estimator is similar to the ratio estimator in survey sampling (see, for example, Cochran, 1977), but applied to deviations from a locally smoothed surface. Hence, placing approximate Fieller confidence limits (Fieller, 1932; James et al., 1974) on our estimates is relatively straightforward. However, the computed confidence limits turn out to be very wide. In any case, the Fieller confidence limits are known to be not conservative enough (Cochran, 1977). Most importantly, while the Fieller limits are essentially local, better confidence limits can be obtained in our case by aggregating information over the spatial domain. The approach behind our estimation method is local. However, spatial dependence in the data can be used to obtain variogram-fitting (Opsomer et al., 1999) GLS-type pointwise standard errors, which are much smaller. Consistent standard error estimates for the finite element spline regression are also developed in Sangalli et al. (2013). In our case, these standard errors are only approximate, since they do not account for errors in the first stage of smoothing. Nevertheless, we also computed these standard error estimates for approximate comparison across the spatial domain. Referring to Figure 5, it turns out that the positive elasticity estimates for locations within the contour marked -0.92 (corresponding to elasticity of 0.08 or higher) are statistically significant at the 5 percent level (except one location), and the two locations close to the southern beaches (marked in blue and green) have significantly negative elesticities. However, full development of these standard error estimates remains in the domain of future work, and hence we do not report these.

Our observation of low or even negative estimated living space elasticities of price in locations around or with easy access to the beaches does not necessarily imply that housing in these localities is undesirable. It is possible that the locations themselves are desirable, but largely as second homes or holiday residences, or locations popular for holiday rentals. However, in the market for second residences or holiday homes, a larger house is not necessarily a desirable attribute because it may be more expensive to main-

tain or difficult to let. Our dataset does not contain information on second homes. Therefore, we included location fixed effects and a statistical factor representing distance to the beach, which would account for second house or holiday property effects. Then, the significant negative elasticties indicate that the benefits of larger living space are offset by higher maintenance costs, and perhaps lower demand for larger holiday rentals as well. Finally, this also highlights the important point that our proposed methodology appropriately takes into account the complexity of the spatial domain but not necessarily omitted variable bias. Hence, one would require extensive data on micro-geographic determinants and amenities.

The slope coefficients on the 6 covariates with spatially fixed slopes are consistent with *a priori* epectations (Table 1). The 5 statistical factors are orthogonal and standardised. A one standard deviation decrease in distance to the CBD or to central amenities (factor 1) corresponds to a 16 percent rise in prices, while the same for access to local services and amenities (factor 2) is 61 percent, and for access to beaches, schools and local commerce (factor 3) is 65 percent. These estimates are very large and reflect the high premium on good location. By contrast, a one standard deviation enhancement in physical attributes of the house (factor 4) leads to a 26 percent increase in the value of the house, as compared to additional house facilities (factor 5) which lead to 7 percent rise in the house price. Time on the market (in logarithms) is included as a regressor to account for the wedge between listing and sale prices. The coefficient reflects that each one additional month on the market reduces sale price by about 17 percent. Finally, the spatial fixed effects are accounted for in our estimation, but are not actually computed. Rather, they are removed by a local fixed effects transformation. In any case, there would be an incidental parameters problem with the fixed effects estimates, and hence we do not attempt to recover these.

Table 1: Estimates of spatially fixed coefficients

Covariate	**Coefficient** (t-ratio)
(Log) Time on market	0.0055 (4.77)
Access to CBD	−0.1631 (−2.18)
Access to local services	−0.6120 (−14.96)
Access to beaches, schools	−0.6468 (−24.04)
Physical house attributes	0.2611 (35.66)
Additional facilities	0.0698 (13.73)

In summary, applied to the above housing market, the methods produce estimates for the smoothed spatial surface of (logarithm of) price per square meter of living space (Figure 4) and the surface of shadow price of living space (Figure 5). This is after controlling for spatial fixed effects and 6 additional covariates with spatially fixed coefficients. The plots reflect the historical and spatial structure of the above housing market quite well, and the spatial variation in implicit prices (elasticties) provide useful implications for policy and planning (Cheshire and Sheppard, 1995).

7 Conclusion

Building upon recent literature on smoothing over irregular spatial domains and locally weighted semiparametric regressions, we develop a method for estimating the functional surface of a regression coefficient that varies over a complex spatial domain with irregular boundaries, peninsulas and interior holes. Applied to urban housing market data, the method produces useful estimates of the spatially varying effect of living space on house prices. The estimates adapt well to the irregular and complex spatial domain, and boundary spillages are small. Our work is related to Basile et al. (2015), who recently proposed an alternative spatial spline based method for nonparametric estimation in spatial autoregressive geoadditive models, but their splines are not designed specifically to account for complex spatial domains. Our work is complementary, where the spline functions are based on finite element analysis and easily adjust to any irregularities and discontinuities in the spatial domain.

In the literature, theoretical results have been developed in certain spatial setups. Lahiri (1996) developed asymptotic properties of estimators under the sampling structure of infill asymptotics. Specifically, least squares estimator and methods of moments variogram estimator for a random field process both converge in L^2 to non degenerate limiting random vectors. This is further extended to a more general case where the estimator sequence only need to satisfy some smoothness and symmetry conditions. Recently, in Bandyopadhyay et al. (2015), an empirical likelihood methodology for irregularly spaced spatial data in the spatial domain is developed. Under some regularity conditions, the spatial frequency domain empirical likelihood ratio statistics proposed have asymptotically chi-squared limit. Both the above papers are built on the simpler setup with one random field process $\{Z(x) : x \in \mathcal{R}\}$, without consideration for a regression setting.

A similar semi parametric modeling setup to our model can be found in the work of Sun et al. (2014), where they proposed a profile likelihood

estimator for a extended spatial autoregressive model with spatial varying regression coefficient for some regressors and fixed coefficient for others. The paper is also motivated by a housing price application. However, the asymptotic properties established in all of the above papers do not consider the irregularity of the spatial domain, such as the existence of irregular boundaries and interior holes. However, by allowing for nonparametric spatial fixed effects, we address issues of sampling in a relatively simple way, while retaining the simplicity of the spatial smoothing framework.

Several directions of further research emerge. First, the spatial smoothing literature pays little attention to estimates of precision. Thus, estimation of pointwise and global standard errors for the smoothing methods themselves will be useful. This will then lead to accurate and sufficiently conservative estimates of the precision of the functional surface of the slope,coefficient in our functional regression context. We have seen that the traditional Fieller (1932) confidence limits, using information only locally, do not work well in our setting. Hence construction of more appropriate limits using information aggregated over the spatial domain will be useful.

Second, functional regression methods have recently started to be applied to spatial data where the slope varies over a multidimensional domain, but observations are not replicated at different spatial data points, and where the spatial weights may be endogenous; see, for example, Bhattacharjee et al. (2016). However, some of these methods currently lack formal statistical foundations. Based on the framework inherent in this paper, more appropriate statistical foundations can be developed for these problems.

Third, Bhattacharjee et al. (2016) have recently developed an approach based on functional data analysis to delineate housing submarkets. Here, key emphasis lies on spatial clustering at two levels: (a) the functional surface of the regressor (logarithm of living space), and (b) the functional surface of its effect on house prices. Based on the spatial clustering methodologies in combination with the method developed here, specific emphasis can be placed on clustering that takes into explicit account the spatial positions of the locations. Finally, the proposed method suggests natural extensions to spatial autoregressive geoadditive models and locally weighted regressions that take into account the complexity of an irregular spatial domain. Further work is required to develop estimators based on these ideas. Development along the above lines constitutes a promising programme of future research.

References

[1] Anselin, L., Lozano-Gracia, N., Deichmann, U. and Lall, S. (2010). Valuing access to water: a spatial hedonic approach, with an application to Bangalore, India. *Spatial Economic Analysis* **5**(2), 161–179.

[2] Bandyopadhyay, S., Lahiri, S. N. and Nordman, D. J. (2015). A frequency domain empirical likelihood method for irregularly spaced spatial data. *The Annals of Statistics* **43**(2), 519–545.

[3] Banerjee, S., Gelfand, A.E., Knight, J.R. and Sirmans, C.F. (2004). Spatial modeling of house prices using normalize distance-weighted sums of stationary process. *Journal of Business and Economic Statistics* **22**(2), 206–213.

[4] Basile, R., Kayam, S., Minguez, R., Montero, J.M. and Mur, J. (2015). Semiparametric spatial autoregressive geoaddive models. In: Commendatore et al. (Eds.) *Complexity and Geographical Economics, Dynamic Modeling and Econometrics in Economics and Finance*, Springer-Verlag: Berlin.

[5] Bhattacharjee, A., Castro, E.A., Maiti, T. and Marques, J.L. (2016). Endogenous spatial structure and delineation of submarkets: a new framework with application to housing markets. *Journal of Applied Econometrics* **31**, 32–57.

[6] Bhattacharjee, A., Castro, E.A. and Marques, J.L. (2012). Understanding spatial diffusion with factor-based hedonic pricing models: the urban housing market of Aveiro, Portugal. *Spatial Economic Analysis* **7**(1), 133–167.

[7] Cheshire, P. and Sheppard, S. (1995). On the price of land and the value of amenities. *Economica* **62**, 247–267.

[8] Cochran, W.G. (1977). *Sampling Techniques*. 3rd ed. Wiley: New York.

[9] Eilers, P.H.C. (2006). P-spline smoothing on difficult domains. Seminar at Ludwig-Maximillians University Munich. http://www.statistik.lmu.de/sfb386/workshop/smcs2006/slides/eilers.pdf.

[10] Fieller, E.C. (1932). The distribution of the index in a bivariate Normal distribution. *Biometrika* **24**(3-4), 428–440.

[11] Fotheringham, A.S., Brunsdon C. and Charlton, M. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A* **30**, 1905–1927.

[12] Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley: Chichester (UK).

[13] Härdle, W., Hall, P. and Marron, J.S. (1988). How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum? (with discussion). *Journal of the American Statistical Association* **83**(401), 86–101.

[14] Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**(1), 153–161.

[15] Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press.

[16] James, A.T., Wilkinson, G.N., and Venables, W.N. (1974). Interval estimates for a ratio of means. *Sankhya Series A* **36**(2), 177–183.

[17] Lahiri, S. N. (1996). On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhya Series A* **58**(3), 403–417.

[18] Maclennan, D. (1977). Some thoughts on the nature and purpose of hedonic price functions. *Urban Studies* **14**, 59–71.

[19] McMillen, D.P. (2010). Issues in spatial data analysis. *Journal of Regional Science* **50**(1), 119–141.

[20] McMillen, D.P. and Redfearn, C.L. (2010). Estimation and hypothesis testing for nonparametric hedonic house price functions. *Journal of Regional Science* **50**(3), 712–733.

[21] Majumdar, A., Munneke, H.J., Gelfand, A.E., Banerjee, S. and Sirmans, C.F. (2006). Gradients in Spatial Response Surfaces With Application to Urban Land Values. *Journal of Business and Economic Statistics* **24**(1), 77–90.

[22] Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. Chapter 5, In: Gibb, K. and O'Sullivan, A. (Eds.), *Housing Economics and Public Policy: Essays in Honour of Duncan Maclennan*. Blackwell Science: Oxford (UK), 67-89.

[23] O'Donnell, D., Rushworth, A., Bowman, A.W., Scott, E.M. and Hallard, M. (2014). Flexible regression models over river networks. *Applied Statistics* **63**(1), 47–63.

[24] Opsomer, J.D., Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1999). Kriging with nonparametric variance function estimation. *Biometrics* **55**, 704-710.

[25] Pryce, G. (2013). Housing submarkets and the lattice of substitution. *Urban Studies* **50**, 2682–2699.

[26] Ramsay, J.O. and Silverman, B.W. (2005). *Applied Functional Data Analysis*. Springer-Verlag: New York.

[27] Ramsay, J.O. and Silverman, B.W. (2006). *Functional Data Analysis*, 2nd Ed. Springer-Verlag: New York.

[28] Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B* **64**(2), 307–319.

[29] Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* **82**(1), 34–55.

[30] Rothenberg, J., Galster, G., Butler, R.V. and Pitkin, J.K. (1991). *The Maze of Urban Housing Markets: Theory, Evidence and Policy*. University of Chicago Press.

[31] Sangalli, L.M., Ramsay, J.O. and Ramsay, T.O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B* **75**(4), 681–703.

[32] Sun, Y., Yan, H., Zhang, W., and Lu, Z. (2014). A semiparametric spatial dynamic model. *The Annals of Statistics* **42**(2), 700–727.

[33] Wang, H. and Ranalli, M.G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics* **63**(1), 209–217.

[34] Wood, S.N., Bravington, M.V. and Hedley, S.L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: Series B* **70**(5), 931–955.

[35] Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.