Online supplementary material for "Endogenous spatial regression and delineation of submarkets: A new framework with application to housing markets" [#]

Arnab Bhattacharjee [*]
Heriot-Watt University, UK
a.bhattacharjee@hw.ac.uk

Eduardo Castro
University of Aveiro, Portugal
ecastro@ua.pt

Taps Maiti
Michigan State University, USA
maiti@stt.msu.edu

João Marques
University of Aveiro, Portugal
jjmarques@ua.pt

Abstract

In the process of revision, the original paper was substantially improved, but had become too long for publication in printed form. Therefore, hard decisions had to be made in shortening the paper to the required length. This full version of the paper, including the omitted material is included in this online supplement. While this extended supplement is logically self-contained, it should ideally be read in conjunction with the printed version. In particular, the online material includes detailed discussion of the needs of submarket delineation, additional discussion placing our work within the context of the literature, more details on estimation and extended empirical results.

KEYWORDS: Spatial heterogeneity; Endogenous spatial dependence; Housing submarkets; Spatial lag model; Geographically weighted regression; Functional linear regression.

JEL CLASSIFICATION: C21; R31; C38; C51.

> *"(Social) space is a (social) product ... the space thus produced also serves as a tool of thought and of action; that in addition to being a means of production it is also a means of control, and hence of domination, of power. ... Change life! Change Society! These ideas lose completely their meaning without producing an appropriate space."* (Lefebvre, 1974 [1991], p.26, p.59).

[#] We thank the Editor and two anonymous reviewers for their encouraging, constructive and challenging comments that helped us improve the paper substantially. Some of our full response to these comments is omitted from the published version of the paper, but is included in this online supplement.

[*] Corresponding author: Arnab Bhattacharjee, Heriot-Watt University, Spatial Economics and Econometrics Centre (SEEC), Room 1.06, Mary Burton Building, Edinburgh EH14 4AS, Scotland, United Kingdom. E-mail: a.bhattacharjee@hw.ac.uk. Tel.: +44 (0)131 4513482. Fax: +44 (0)131 4513296.

1. Introduction

Endogenous evolution of urban space, as emphasized by Lefebvre (1974 [1991]), is central to spatial dynamics in urban housing markets. Definition of housing submarkets, at both conceptual and empirical levels, is important in this context. Understanding endogenous housing segmentation enables researchers to study spatial variation in housing prices, improving lenders' and investors' abilities to price the risk associated with financing homeownership; at the same time it reduces search costs to housing consumers (Malpezzi, 2003, Goodman and Thibodeau, 2007).

By its nature housing is a heterogeneous good, characterized by a diverse set of attributes (Lancaster, 1966; Rosen, 1974) and segmented and structured by complex spatial patterns. Different social groups, with specific tastes, preferences and economic capabilities tend to be organized into distinct territorial clusters (Galster, 2001), ranging from national or regional scale, through metropolitan areas, to below the metropolitan level (Follain and Malpezzi, 1980; Rothenberg *et al.*, 1991; Maclennan and Tu, 1996; Bourassa *et al.*, 1999). However the literature does not suggest an unequivocal and unique spatial approach to analyse this issue, encompassing different philosophies, techniques and criteria.

Housing markets are complex. Rather than being defined by a single combination of a quantity and a price, the market equilibrium for a heterogeneous good such as a house is given by the combination of a vector of hedonic characteristics with a vector of hedonic prices which, under an appropriate function specification, produce an overall price for the good (Lancaster, 1966; Rothenberg *et al.*, 1991). Thus, the existence of a unique vector of hedonic prices, combined with a distribution of houses with different hedonic characteristics, is a necessary condition for the existence of a single equilibrium and a unique market. However, what we generally observe is the co-existence of several submarkets, corresponding to different market equilibrium in each of these submarkets (Rothenberg *et al.*, 1991).

There are several reasons to explain this empirical evidence. First, houses are durable goods that cannot be continuously adjusted to changes in demand. New houses can be designed in order to meet the expected demand requirements. However, once they are built, any change to their characteristics is a costly and sometimes an impossible task, a rigidity which tends to create a permanent lag between supply and consumer tastes. Second, the existence of significant search costs and information asymmetries makes branding an important market feature (Williamson, 2000). Such branding corresponds to clusters of relatively homogeneous houses designed to meet the requirements of particular social groups. Because a house is simultaneously a consumer good, an asset and a status benchmark (Marques *et al.*, 2012), branding not only facilitates search but tends to endure in housing clusters, homogeneous in hedonic characteristics and prices as well as in the social composition of residents.

Supply rigidities and transaction costs are then the main drivers of heterogeneity, shaping the territory as landscapes of submarkets. Such landscapes can be either represented as sets of hedonic functions, each with one particular vector of hedonic prices, or as a continuous transition of vectors, represented by an hedonic functional. The application of a functional representation to the empirical study of housing hedonic price models is one of the main objectives of this paper.

Because of such inherent heterogeneity over space, understanding housing markets and the conduct of housing policy crucially depends on delineation of submarkets (Rothenberg *et al.*, 1991). Each such submarket is characterized by different supply and demand curves and a different equilibrium. A multitude of criteria have been proposed in the literature for defining housing markets and their constituent submarkets. In general, the delineation of submarkets can be based on *a priori* judgments such as pre-existing administrative boundaries or subjective knowledge, but equally, by using the structure of data to apply analytical methods such as hierarchical models or non-parametric spatial statistical models; see, for example, McMillen (1996) and Goodman and Thibodeau (2007).

These analytical methods are based on the theoretical assumptions underlying the definition of submarkets. As discussed in the next section, there are three main approaches, viewing submarkets as defined by three possible criteria: i) similarity in hedonic characteristics; ii) similarity in hedonic prices; or iii) close substitutability of housing units. We argue that spatial clustering based simultaneously on criteria i) and ii) is a sufficient condition for criterion iii) to hold. Since criterion i) is directly observable, we focus on ii). Thus, the central object of our inference is a regression model where the dependent variable is logarithm of house prices per square meter and housing features are regressors. The partial effect of these housing features varies over a two-dimensional territory. In this paper, we focus on a single regressor, logarithm of living area, so that the functional regression coefficient $\beta(s)$ can be interpreted as an elasticity which reflects a positive but decreasing marginal utility of living area. Generally, $-1 < \beta(s) < 0$; when the elasticity approaches zero consumers show a very low satiation of living space, while a value close to negative unity (-1) reflects a submarket with a rigid demand for living space.

Appropriate characterization of spatial structure is a key element of such analyses. Specifically, three distinct aspects of space – spatial heterogeneity, spatial dependence and spatial scale – are central to understanding the spatial organization of housing submarkets. Anselin (1988b:1) defines spatial heterogeneity as the *"heterogeneity inherent in the delineation of spatial units and from contextual variation over space."* In the context of a hedonic pricing model estimated using spatial panel data, this can modelled as cross sectional fixed effects and slope heterogeneity. In this paper, we consider a spatial cross section context, where spatial heterogeneity is modelled as variation across spatial submarkets in intercepts (spatial fixed effects) and spatially varying (heterogeneous) slopes of a regression model. By contrast, spatial dependence is associated with spatial spillover effects, contagion and diffusion; typically, this results in spatial autocorrelation between different spatial units (Anselin, 1988a,b). Additionally, choice of an appropriate spatial scale is important (Malpezzi, 2003). Spatial scale is not so much an econometric, but an important empirical issue; whether an urban scale is the most suitable, or whether the appropriate scale for analysis should be peri-urban (including an urban centre, adjoining suburbs and the countryside), regional or national, depends on both the spatial phenomenon under analysis and the specific spatial context. At the empirical level, the correct treatment of spatial heterogeneity increases the prediction accuracy of the estimated hedonic models and, in many cases, negates spatial strong dependence (Pesaran, 2006; Pesaran and Tosetti, 2011).[1]

We study how estimates of the above hedonic regression model can be used to identify submarkets, following the criterion of similar hedonic prices and characteristics by clustering jointly on the surface of the functional partial effect $\beta(s)$ and the regressor surface $x(s)$. Further, following the current literature (Bhattacharjee *et al.*, 2012), once

[1] Even though spatial heterogeneity and spatial dependence are theoretically distinct problems, adequate treatment of common factors with heterogeneous slopes is necessary for inference on structural spatial dependence (Bhattacharjee and Holly, 2013); see also McMillen (2003).

such submarkets have been delineated, spatial dependence can be examined by estimating cross- and within-submarket spatial weights.

Therefore, we propose a new framework to analyse housing markets, based on a synthesis of spatial econometrics, functional data analysis (FDA) and locally/ geographically weighted regression (GWR). We consider a simple spatial lag model, regressing logarithm of price per square meter of living space on logarithm of house area, allowing for spatial heterogeneity (spatial fixed effects and slope heterogeneity) and endogenous spatial dependence captured by a spatial weights matrix W. This, in turn, leads to a functional regression model where the response variable is scalar and the functional regressor is a spatially weighted version of the average functional surface of the regressor. When kernel weights are used, the model is very similar to GWR. This synthesis of GWR and FDA offers a spatial statistical model that is very rich and enable the full range of spatial analyses of housing markets.

The above model addresses two main limitations of previous approaches. First, the framework allows submarkets to evolve endogenously and abstracts from the requirement to delineate housing submarkets *a priori*. The submarkets can be delineated *ex post* by spatial clustering, or even simple hierarchical clustering, of the estimated functional regression slope and hedonic feature surfaces. Second, endogeneity in spatial structure and estimated spatial weights can be naturally incorporated into the model. Application to the housing market of the Aveiro-Ílhavo urban conglomeration in Portugal implies submarkets that emphasize the historical and endogenous evolution of the urban spatial structure.

The paper is organised as follows. Section 2 discusses some recent developments in the spatial econometrics literature applied to the hedonic pricing model, followed by delineation of submarkets in section 3. Section 4 highlights limitations of the spatial econometrics framework, discusses alternative approaches and proposes a new synthesis of several methods. Based on this synthesis, section 5 develops methodology for submarket delineation, followed by an application to the urban housing market of Aveiro and Ílhavo in Section 6. Finally, section 7 concludes.

2. Spatial Econometric Hedonic House Price Models

Smith *et al.* (1988) and Marques *et al.* (2012) discuss several spatial problems of current interest that are fundamental for understanding the housing market, covering supply and demand sides, price formation and policy. Of specific relevance in the current context is the use of hedonic models to study spatio-temporal dynamics and price formation.

Typically, hedonic and repeated sales models of local or regional house prices reflect not only geographically varying price effects, but also substantial clustering. Malpezzi (2003) argues that this is the outcome of supply rigidities, search costs and social segregation. Attempts have been made to explain such spatial clustering by neighbourhood characteristics such as crime rates, schooling, transport infrastructure and quality of public services, and social interaction and segregation; see, for example, Rothenberg *et al.* (1991). Therefore, empirical estimation of hedonic housing price models and the use of such estimates for evidence and policy have to take spatial effects explicitly into account.

2.1. Hedonic pricing model

Building on the early work of Lancaster (1966) and Rosen (1974), hedonic pricing models continue to be actively used in housing studies. In particular, valuation of housing attributes (including living space), neighbourhood features and access to central and local services, and construction of price indices based on single sales data, have been addressed through hedonic specifications; see Maclennan (1977) for a classic and critical discussion, and Malpezzi (2003) for an excellent review.

In hedonic pricing models, dwelling unit values (or proxies such as prices or rents) are regressed on a bundle of characteristics of the unit that determine the value:

$$Y = f(S, N, L, C, T), \tag{1}$$

where Y denotes the value of the house (typically logarithm of price, or logarithm of price per unit area), and S, N, L, C and T denote respectively, structural characteristics of the dwelling (living space, type of construction, tenure, etc.); neighbourhood characteristics (and local amenities); location within the market (or access to employment/ business centre); other characteristics (access to utilities and public services, such as clean water supply, electricity, central heating, etc.); and the time when the value is observed.

Estimating the hedonic price function using a collection of observed housing values and dwelling unit characteristics yields a set of implicit prices for housing characteristics that are essentially willingness-to-pay estimates. This allows analysis of various upgrading scenarios, targeted to specific subgroups, defined either by socio-economic characteristics or by location. Thus, the model facilitates understanding of residential location, and therefore urban structure, and provides valuable input towards urban planning and housing policy.

The two main limitations of traditional hedonic models are the frequent assumption that hedonic prices do not vary spatially and inadequate attention to spatial spillover effects. To overcome these problems, we consider a hedonic model incorporating both spatial variation in the relationship between housing price and living space, as well as spatial dependence. Following Bhattacharjee *et al.* (2012), we adopt a semi-log form, where logarithm of price per square meter of living space is regressed on logarithm of house area, conditioning on several other hedonic housing characteristics, used as control variables.

2.2. Spatial issues in hedonic pricing estimates

The recent literature has discussed the potential bias and loss of efficiency that can result when spatial effects are ignored in the estimation of hedonic models; see, for example, LeSage and Pace (2009), Anselin and Lozano-Gracia (2008) and Anselin *et al.* (2010). Specifically, substantial biases can result both from inadequate modelling of endogenous spatial effects and inadequate attention to spatial heterogeneity, while heteroscedasticity and spillovers in unobservable errors lead to large inefficiency.

As discussed above, spatial patterns in housing markets arise primarily from a combination of spatial heterogeneity and spatial dependence (Anselin, 1988a,b). Spatial dependence is associated with spatial spillover effects while spatial heterogeneity arises from contextual variation in space and results in spatially varying slopes and intercepts of a regression model (Anselin, 1988b). Additionally, choice of a spatial scale appropriate to a given application context is also important. We now turn to a discussion of spatial issues in the construction of hedonic pricing models, including all of the three above aspects of space.

2.2.1. Spatial scale and housing submarkets

Definition of submarkets is important at both conceptual and empirical levels. Housing markets are local and diverse, and hedonic price estimation requires careful delineation of these markets. The definition of submarkets in practice ranges from the national or regional scale, through metropolitan areas, to below the metropolitan level. Malpezzi

(2003) argues that one reason why the metropolitan area is appealing as the unit of analysis is that these areas are usually thought of as labour markets, which may therefore be approximately coincident with housing markets. On the other hand, submarkets below the metropolitan level can be segmented by location (central city/suburb), or by housing quality, or even by race or income levels. Such segmentation facilitates both understanding of residential neighbourhood choice and devising appropriate urban housing policy.

The definition of the most appropriate scale in the analysis of urban spatial patterns is a crucial aspect. A spatial configuration at a certain scale is not necessarily the same at another one, in other words, a specific urban pattern which appear to be structured at one scale, may appear to be disordered at other scales (Miller, 1978), leading to the so called "ecological fallacy" (Fujita and Thisse, 2002). According to Anas *et al.* (1998), one of the reasons for this uncertainty is the different effects of agglomeration economies that emerge at specific scales.

In the specific case the empirical application here, the urban area of Aveiro has the adequate size and variability to illustrate the issue of spatial heterogeneity in the shadow (hedonic) price of housing space, at the same time as allowing for spatial spillovers in house prices.

2.2.2. Spatial heterogeneity and spatial patterns

The model for spatial heterogeneity must, in principle, be based on a theoretical framework explaining why and how housing markets are segmented. As discussed earlier, the literature has defined submarkets either by similarity in hedonic housing characters (Rothenberg *et al.*, 1991; Adair *et al.*, 1996; Maclennan and Tu, 1996; Bourassa *et al.*, 1999; Watkins, 2001), by similarity in hedonic prices (Dale-Johnson, 1982; Rothenberg *et al.*, 1991), or by close substitutability of housing units (Goodman and Thibodeau 2007; Pryce, 2013).

In the first approach, a submarket is a collection of locations, or housing units located therein, that have similar bundle quality or, in other words, supplies a similar set of hedonic characteristics. The degree of similarity required to define a submarket is a debateable issue, particularly since a perfectly homogeneous location may be very small and therefore not useful for estimating hedonic models, which needs a minimum level of variety in order to enable reliable estimation (Bourassa *et al.*, 2003). In any case, the delineation of submarkets implied by this approach can be directly applied to the data, using a clustering methodology that may be spatial, or may not. This approach has a logic that stresses the role of branding and social segregation as the driver of submarkets.

The second approach defines submarkets as locations where hedonic (shadow) prices for different features are homogeneous. Submarkets can then be interpreted as clusters of houses with characteristics which are adjusted to a particular demand behaviour reflected in a set of equilibrium prices. The approach was proposed by Bourassa *et al.* (2003) as a means to improve the accuracy of price predictions. More importantly, the approach is intimately related to the basic philosophy of hedonic models stating that, within the same submarket, the implicit prices corresponding to each housing feature must be homogeneous. Thus, the delineation of submarkets based on hedonic prices depends on the capacity to encompass slope heterogeneity, across space, in the estimation of the hedonic model; this paper addresses this problem using methodology based on functional data analysis.

The third criterion for defining submarkets is based on the degree of substitutability (Grigsby, 1963). Pryce (2013) measures substitutability by the cross-price elasticities of price at different locations, estimated using data on spatial panels. A key assumption underlying this methodology is that one can estimate a regression model where the logarithm of house prices at one location is regressed on log-price at the same location at another time point together with a time trend. This time trend is, then, the sole latent factor that can contribute to spatial strong dependence. Inclusion of this regressor therefore ensures that the spatial structure contains only spatial weak dependence, in the sense of Pesaran and Tosetti (2011).[2] By contrast, we take the view that, in the context of a hedonic house price model based on cross-section data, a collection of suitably chosen housing characteristics constitutes a more natural set of latent factors. One may expect that inclusion of these factors account for any strong spatial dependence, which would then render the spatial model as containing only spatial weak dependence. Inference on substitutability using spatial cross-section data is in the domain of this paper.

Under what conditions are the above approaches equivalent? It is possible to envisage situations where homogeneity in hedonic characteristics does not imply close substitutability. If two locations with similar houses, similar provision of local services and amenities and similar accessibility to the centre are inhabited by two different social groups (for example, young highly educated professionals and middle-aged middle class), it is expected that different tastes and different responses to fashion will generate local branding effects which both mitigate against substitutability and create differences in hedonic prices. Nevertheless, two locations with both similar characteristics and hedonic prices must be good substitutes, as it will be very difficult to make a distinction between them. Therefore, we argue that simultaneous similarity in hedonic prices and characteristics is a sufficient condition for substitutability. However, this is not a necessary condition, because two types of houses with very different hedonic characteristics can be good substitutes. For example, a flat in a central location can be an alternative to a more peripheral detached house with a similar price; hence, nearby location is also required.

Therefore, when the delineation of submarkets using hedonic characteristics and hedonic prices criteria overlap, the outcome also corresponds to submarkets where the close substitutability criterion holds. However, when we extend the discussion to larger areas with some internal heterogeneity, the problem becomes more complex. Such heterogeneity makes it very unlikely that submarkets delineated by hedonic prices or hedonic characteristics overlap. It is more reasonable to expect partial overlap, implying that each criterion produces a specific set of submarkets which, although related to each other, represent distinct dimensions of preference. Therefore, the delineation of submarkets by combining physical characteristics and hedonic prices seems to be a good principle and we can assume that the trade-off between physical and price homogeneity still reflects a high degree of substitutability. In other words, two identical houses in terms of prices and characteristics located within the same spatial context, a submarket in our definition, are good substitutes.

The conceptual notion behind spatial submarkets discussed above implies that the price determining (hedonic) mechanism can be heterogeneous over space. This spatial heterogeneity, reflecting the absence of a single equilibrium in the housing market, can originate from demand and/ or supply side factors, institutional barriers or discrimination, each of which can cause differentials across neighbourhoods in the way

[2] Pryce (2013) does not explicitly state this assumption, but it is implied by the methodology. The methodology rests on computation of inflation in house prices at different locations, which assumes such an underlying spatial model, together with the assumption that inclusion of the time trend ensures spatial weak dependence; otherwise, such estimation would be biased.

housing attributes are valued by consumers and house prices determined (Anselin *et al.*, 2010). However, if spatial submarkets exist and are ignored, an average price across all the territory is estimated that ignores submarket heterogeneity.

As discussed above, heterogeneity is a key element of housing markets and its disregard seriously affects the understanding market diversification – a central element of the housing market. Worse still, average prices estimated by OLS are likely to be biased because the error term of the regression model may be correlated with the included regressors.

The standard urban model in the Alonso-Muth-Mills tradition predicts a generally declining pattern of prices with distance from the centre of the city, though there may be spatial variation in relative preference for centrality. Other models based on localised amenities or multiple centres imply a stronger impact of access to local amenities. Like distances, the implicit prices for dwelling characteristics and size may also vary spatially, reflecting either supply constraints or residential sorting. Follain and Malpezzi (1980) and Adair *et al.* (1996), among others, have examined intra-urban variation in the price of housing amenities using hedonic models.

There are two main methods proposed in the literature. In the first, one allows coefficients in the hedonic pricing model to vary across submarkets, and use the estimated variation to infer on residential neighbourhood choice and urban spatial structure. The second, and increasingly more popular approach, is geographically weighted regressions (GWR) (Fotheringham *et al.*, 1998), a specific form of local regression that we discuss later.

2.2.3. Spatial dependence and spatial weights matrix

In contrast to spatial heterogeneity, spatial dependence leads to spatial autocorrelation, implying that prices of nearby houses tend to be more similar than those of houses that are farther apart. Likewise, average price of houses in nearby or related submarkets may be correlated more strongly. A common explanation for spatial autocorrelation is spatial spillovers or contagion effects. However, incorrectly modelled spatial heterogeneity, measurement errors in explanatory variables and omitted variables can also lead to spatial autocorrelation (Anselin and Griffith, 1988). Perhaps most importantly, unmodelled spatial patterns in hedonic features lead to spatial dependence in the nature of the Spatial Durbin model; see, for example, LeSage and Pace (2009).

Spatial dependence is very common in housing markets, and a feature that we use in this paper to develop inferences for a functional regression model. Recent empirical literature has addressed issues of bias and loss of efficiency that can result when spatial effects are ignored in the estimation of hedonic models,[3] and the use of spatial econometric models to address spatial autocorrelation is becoming increasingly standard; see, for example, Anselin *et al.* (2010).

The usual approach to the representation of spatial interactions is to define a spatial weights matrix, denoted W, which represents a theoretical and *a priori* characterisation of the nature and strength of spatial interactions between different submarkets or dwellings.[4] These spatial weights represent patterns of diffusion of prices and unobservables over space, and thereby provide a meaningful and easily interpretable representation of spatial interaction (spatial autocorrelation).

[3] See, for example, LeSage and Pace (2009) and Anselin and Lozano-Gracia (2008).

[4] For a setting with n spatial units, W is an $n \times n$ matrix with zero diagonal elements. The off-diagonal elements are typically either dummy variables for contiguity or inversely proportional to the distance between a pair of units, so that spillovers between a pair of units that are farther apart is lower.

Given a particular choice of the spatial weights matrix, there are two important and distinct ways in which spatial dependence is modelled in spatial regression analysis – the spatial lag model and the spatial error model. In the former, the hedonic regression includes as an additional regressor the spatial lag of the dependent variable *y* (which in this case is price), represented by *Wy*:

$$\underline{y} = \rho W \underline{y} + X\beta + \underline{\varepsilon}, \tag{2}$$

where *X* denote the combination of hedonic characteristics (*S*, *N*, *L*, *C* and *T*) and the regression errors (ε) are completely idiosyncratic. By contrast, in the spatial error model, the regression errors are spatially dependent on their spatial lag, $W\varepsilon$:

$$\underline{y} = X\beta + \underline{\varepsilon},\ \underline{\varepsilon} = \lambda W \underline{\varepsilon} + \underline{\eta}. \tag{3}$$

The implications of spatial interaction on estimation of these two models are different. In the spatial lag model, the endogenous spatial lag implies that OLS estimates not accounting for spatial interaction would be biased, while in the spatial error model, they will be unbiased but inefficient.

A common approach is to first estimate the hedonic pricing model under the spatial error assumption. Next, to judge whether endogenous spatial lags are relevant, one can perform a test for spatial lag dependence by nesting the spatial error model within a hybrid model incorporating both spatial lag and spatial error dependence; for more discussion on sequential model selection in the spatial context, see Bhattacharjee *et al.* (2012). Some recent literature (LeSage and Pace, 2009) argues that the choice of appropriate spatial weights not so crucial for the accuracy of a spatial model's estimates. Yet, the correct specification of spatial weights is important when the main goal is to analyse spatial dynamics and how they determine spatial structure in general equilibrium (Arbia, 2014; Bhattacharjee *et al.*, 2014).[5]

The spatial weights are typically modelled either as spatial contiguity, or as functions of geographic or economic distance. The distance between two spatial units reflects their proximity with respect to prices or unobservables, so that the spatial interaction between a set of units (dwellings) can be represented as a function of the economic distances between them. However, spatial data may be anisotropic, where spatial autocorrelation is a function of both distance and the direction separating points in space (Gillen *et al.*, 2001). Further, spatial interactions may be driven by other factors, such as trade weights, transport cost, travel time, and socio-cultural distances. The choice typically differs widely across applications, depending not only on the specific economic context but also on availability of data. The problem of choosing spatial weights is a key issue in many applications; see, for example, Harris *et al.* (2011) and Bhattacharjee and Jensen-Butler (2013).

2.3. Interconnection between the three aspects

Recently, Bhattacharjee *et al.* (2012) developed a framework that emphasizes the connection between urban spaces and housing markets and places focus all the three distinct but interconnected features of space – spatial heterogeneity, spatial dependence and spatial scale. Further, while the traditional literature assumed an *a priori* known structure of spatial dependence in terms of a pre-specified spatial weights matrix, and then examined spatial dependence and spatial heterogeneity within a spatial context implied by the pattern of spatial weights, a branch of the current literature treats these

[5] The literature suggests that the accuracy of spatial weights affects profoundly the estimation of spatial dependence models (Anselin, 2002; Harris *et al.*, 2011). In particular, estimate of the spatial persistence (the spatial autoregressive parameter) is sensitive to the choice of *W* (Harris *et al.*, 2011) and likewise inferences on spatial general equilibrium effects (Arbia, 2014).

weights as unknown and an object of econometric inference. Based on a given definition of urban submarkets (or a fixed set of spatial locations) and panel data on these spatial units, Bhattacharjee and Holly (2013) and Bhattacharjee and Jensen-Butler (2013) developed several methods to estimate the spatial weights matrix between the submarkets.

Bhattacharjee *et al.* (2012) extended the panel estimation methodology in Bhattacharjee and Jensen-Butler (2013) under the structural assumption of symmetric spatial weights to a purely cross-section setting. Their methodology combines spatial hedonic analysis based on orthogonal factors with a method for inferences on unknown spatial weights matrix under the structural constraint of symmetric spatial weights. First a suitable spatial scale is fixed. Next, at the above chosen scale, the housing market is segmented into submarkets, based on a combination of several criteria: administrative boundaries, hedonic prices and socio-cultural segmentation. Given the above segmentation into submarkets, spatial dependence relates to inferences on spatial weights representing spillovers across different submarkets, and those between houses within the same submarket. Since, spatial strong dependence in this model arises from the underlying factor structure, estimation of spatial weights is based on matching residuals across submarkets by closeness across a vector of estimated statistical factors. Finally, spatial heterogeneity is used to inform spatially varying coefficients, spatial structural change and heteroscedasticity.

The resulting spatial model is useful for understanding relative importance of various elements – housing characteristics and access to central and local amenities, as well as interactions within and between housing submarkets – and provides useful inferences on residential location, urban planning and policy. Substantial gains are also obtained with regard to house price prediction. However, whereas Bhattacharjee *et al.* (2012) focussed on estimation and inferences on an unknown W in a setting where the delineation of submarkets was assumed known *a priori*, the current paper focuses on identifying submarkets in a setting where spatial structure (represented by W) may be estimated and potentially endogenous.

3. Delineation of Housing Submarkets

As discussed above, defining and delineating submarkets is an area of considerable debate and multiple alternate approaches. The task of dividing a large market into submarkets raises numerous theoretical and methodological questions (Rothenberg *et al.*, 1991). One problem is the definition of submarket. Theoretically, a submarket corresponds to a local equilibrium between supply and demand. However, the way this concept translates into measurement and modelling leads to difficult questions about the levels of aggregation and about the methods which can be used to cluster basic spatial units in order to define submarkets. In practice, these questions are often answered in an ad-hoc manner, using as the basis predefined or otherwise convenient geographical boundaries. In some cases, statistical tests are used to determine whether the *a priori* submarkets are indeed distinct (Bourassa *et al.*, 1999).

3.1. Submarkets based on similarity in hedonic characteristics and prices

Difficulties of implementation and interpretation with such ad hoc procedures have led to recent attempts to use more systematic methods for defining submarkets. Typically, such analyses proceed first by conducting principal component analysis or factor analysis on a large number of hedonic characteristics of houses to combine these into a small number of meaningful factors. Next, clustering methods are used to obtain a set of submarkets that maximise the degree of internal homogeneity (within each submarket) and external heterogeneity (across different submarkets); see Bourassa *et al.* (1999) for further discussion. While the philosophical underpinnings of the above methods are not clearly expressed in the literature, they can be viewed as being closely related to the definition of submarkets by similarity in hedonic housing characteristics (Rothenberg *et al.*, 1991).

An alternative approach is to apply the criterion of homogeneous hedonic prices, using as a measure of homogeneity small residuals from a hedonic pricing model estimated separately for each submarket; see, for example, Bourassa *et al.* (1999, 2003). The objective for such an approach is to use the notion of submarkets to optimise the accuracy of hedonic predictions for mass appraisal purposes. However, as we argued above, homogeneity in hedonic prices is deeply rooted in the basic concepts which underlie hedonic models.

It can be argued that the two approaches are not entirely satisfactory from a housing economics point of view. This is because they do not pay explicit attention to the demand side of the housing market, which is where individual households make neighbourhood and housing choice decisions. Similarity in hedonic housing characteristics relate explicitly to the supply side, and similarity in hedonic prices relate to market equilibria in submarkets. While the supply side is endogenously related to the demand side through the market equilibria, and prices are therefore also determined by both supply and demand through the equilibrating process, there is no direct way to understand the demand side of the market. For this purpose, the concept of substitutability is useful to the extent that it can be interpreted as reflecting the synthetic valuation of houses by buyers (Pryce, 2013).

3.2. Submarkets based on substitutability

Grigsby *et al.* (1987) define submarkets as a set of dwellings that are reasonably close substitutes for one another, but relatively poor substitutes for dwellings in other submarkets. An approach proposed recently by Pryce (2013) attempts to get to the heart of the above key issue, by taking house prices as the sole determinant of housing choice[6] and by evaluating the cross-price elasticity of price for each pair of housing properties.

Then, two houses are deemed to lie within the same submarket if this cross-price elasticity is close to unity, implying therefore that the two houses are substitutable. Pryce (2013) uses house price inflation for computation of the elasticities, and illustrates the methodology for delineating submarkets using data on Glasgow, Scotland, UK. While the above methodology was not provided any specific structural interpretation, placing it within the context of a structural spatial econometric model is useful for our discussion.

The underlying structural model in Pryce (2013) is a spatial error model:

$$\underline{y}_t = \underline{y}_{t-1} + \underline{\varepsilon}_t,\ \underline{\varepsilon}_t = W\underline{\varepsilon}_t + \underline{\eta}_t, \qquad (4)$$

where $\underline{y}_t$ denotes the vector of prices (in logarithms) across all houses, and $\underline{y}_{t-1}$ its lagged value, so that $\underline{\varepsilon}_t$ denotes the growth rate in prices. Then, the elements of W are the cross-price elasticities for each pair of houses.[7] The model does not include any

[6] Taking house prices, rather than hedonic characteristics, as the sole basis for evaluation of substitutability is not an innocuous modelling assumption. See Pryce (2013) for discussion on the motivation behind this assumption, which is sharply distinct from most of the literature.

[7] Since $\underline{\varepsilon}_t$ denotes the growth rate in prices, the spatial error part of (5) models how growth rate in an index location is related linearly to the growth rates at all the other locations, the corresponding coefficients being elements of W. Hence, these elements are essentially cross-price elasticities. In Pryce

explanatory variables except for lagged logarithm of prices with unit coefficient, which is assumed to ensure that the regression errors (growth rates) are stationary in the temporal domain. Nevertheless, the model itself is structural because it assumes that the process of diffusion of shocks (idiosyncratic errors, η_t) is driven by an underlying spatial structure represented in W. The underlying spatial structure is inferred from the estimates, and this in turn determines the delineation of submarkets.

From a spatial econometric point of view, this approach deals with potential temporal nonstationarity, by inclusion of the lag on the right hand side. This also suggests an interpretation of elasticity as the measure of a cause-effect relationship which, by nature, is time lagged. However, spatial nonstationarity is a potential problem. This would be evident if some elements of W are close to unity (or even larger), which would imply violation of the spatial granularity condition (Pesaran and Tosetti, 2011).

Further, violation of this stationarity condition is expected in this setting because cross-price elasticities are by definition close to unity for houses within the same submarket. By such delineation of submarkets, one can ensure that cross submarket spatial diffusion is bounded, and therefore the spillover of house price shocks across the submarkets is spatially stationary. However, spatial weights of houses within the same submarket will be large. Therefore, without suitable modifications, the above model (4) cannot be cast into the framework of spatial econometrics.

3.3. Submarkets based on a structural spatial lag model

This indicates that the above model should be extended in two ways. First, violation of the spatial granularity condition points towards spatial strong dependence, which is caused by ignoring the effect of common factors (Pesaran, 2006; Pesaran and Tosetti, 2011). The solution is to include regressors that will take strong dependence out of the model; see Bhattacharjee and Holly (2013) for further discussion. In the current context, hedonic characteristics can be added to the model, allowing the cross-price elasticities to be measured more robustly.

Second, assumption of a spatial error model is somewhat simplistic. If we really believe that house prices are spatially endogenously determined by the interaction between housing choices of economic agents, the spatial lag model (2) will be more appropriate and permit stronger structural interpretations.

In a setting where W is unknown (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013), W and ρ are not separately identifiable. Hence, we assume without loss of generality that $\rho = 1$. Further, in this case, an unknown W is not in general identified. Further structural assumptions are required for identification. Following Bhattacharjee and Jensen-Butler (2013), we make the assumption that W is symmetric. In applications, spatial weights matrix W is often based on distances, which are symmetric by definition.

Thus, we make the following assumption.

Assumption 1. ***Spatial lag model.*** *The dependent variable y follows a spatial lag model*

$$y = Wy + X\beta + \varepsilon \Rightarrow y = (I - W)^{-1} X\beta + (I - W)^{-1}\varepsilon. \qquad (5)$$

with full spatial heterogeneity in both the slope and intercept (heterogeneity in β across the territory, plus location fixed effects). W is unknown but symmetric, and satisfies the

(2013), two houses are viewed as being substitutable if the cross-price elasticity is close to unity, which in Equation (5) implies the corresponding spatial weights are close to unity.

spatial granularity condition $\rho(W) < 1$, *where* $\rho(W) = \max\left\{\|W\|_1, \|W\|_\infty\right\}$ *is the norm of W,* $\|W\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |w_{ij}|$ *the column norm of W, and* $\|W\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |w_{ij}|$ *the row norm of W. The regression errors,* ε*, have mean zero, and X are potentially stochastic regressors with positive definite covariance matrix that are uncorrelated with* ε*.*

The spatial granularity condition implies that there is no spatial strong dependence (Pesaran and Tosetti, 2011). If there were latent factors causing violation of the spatial granularity condition, these factors are included as regressors in the model (5).

As an illustration, consider a simple spatial lag model regressing logarithm of price per square meter (y) on logarithm of living space (x), allowing for spatial heterogeneity and endogenous spatial dependence. Further, to fix ideas, let us first consider a sample of only two locations with potentially different slopes, with only one house in each location. Then:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = W\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \quad W = \begin{bmatrix} 0 & w_{12} \\ w_{21} & 0 \end{bmatrix}$$

$$\Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = [I - W]^{-1}\begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + [I - W]^{-1}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \tag{6}$$

$$\approx [I + W]\begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + [I + W]\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

where the final step follows under the assumption that the spatial weights are very small compared to unity, so that $[I - W]^{-1} \approx [I + W]$. This assumption is valid if the inclusion of X as a regressor takes out spatial strong dependence from the model, in which case the spatial granularity condition in Assumption 1 holds.

Equation (6) emphasizes that both y_1 and y_2 are endogenously determined by each other, and in addition are functions of x_1, x_2, β_1 and β_2. Endogenous spatial lags have the following additional implication under the spatial lag model (5). The response at location i, y_i, is a function of the regressor at the same location (x_i) with slope β_i, but also the regressor at every other location j, x_j, but with a different slope ($w_{ij}\beta_j$).

Further, conceptually, y_1, y_2, x_1, x_2, β_1 and β_2 can all be thought of as functions of space (S), where $s \in S$ is a representative point in the spatial domain S. We assume that these are sufficiently smooth functions so that all partial derivatives are well defined.

Assumption 2. ***Smoothness.*** *The functional regression coefficient, β(s) is smoothly varying over the spatial domain S, which is a convex set. That is, β(s) has derivatives at every s ϵ S. Likewise, the functional random variables x(s) and y(s) have mean functions,* $\overline{X}(s)$ *and* $\overline{Y}(s)$ *respectively, that are smoothly varying over S.*

By Assumption 2, the partial derivatives $\partial\beta/\partial s$, $\partial x/\partial s$ and $\partial y/\partial s$ are well defined. Then, we have the following result, where all partial derivatives are interpreted with respect to space, $s \in S$.

Theorem 1: *Under Assumptions 1 and 2, similarity in hedonic characteristics, prices and location together imply that two different houses are substitutable, that is the cross-price elasticity of price is close to unity.*

Proof: *By the granularity condition in Assumption 1, the elements of W are small, and hence up to first order Taylor expansion,* $(I - W)^{-1} \approx (I + W)$*. Also, the idiosyncratic errors can be ignored in computation of cross-price elasticities. Then, for any two distinct houses in locations i and j:*

$$\frac{\partial y_i}{\partial y_j} \approx \frac{1+w_{ji}}{1+w_{ij}} \cdot \frac{\partial\left(x_i\beta_i\right)}{\partial\left(x_j\beta_j\right)} = \frac{1+w_{ji}}{1+w_{ij}} \cdot \frac{x_i d\beta_i + \beta_i dx_i}{x_j d\beta_j + \beta_j dx_j} = \frac{x_i d\beta_i + \beta_i dx_i}{x_j d\beta_j + \beta_j dx_j}, \tag{7}$$

where $\left(1+w_{ji}\right)/\left(1+w_{ij}\right)=1$ *since the spatial weights are symmetric (Assumption 1). Thus, sufficient conditions for houses i and j to be (approximately) substitutable are that: (i)* $x_i \approx x_j$*; (ii)* $\beta_i \approx \beta_j$*; and (iii) the locations i and j are in each other's neighbourhood, so that by Assumption 2,* $d\beta_i \approx d\beta_j$ *and* $dx_i \approx dx_j$*. The proof in the general case follows by noting that computation of elasticities involves only a pairwise comparison between 2 properties i and j, and other houses can be ignored because elements of W are small.*

The above result has very important implications for delineation of submarkets. First, a sufficient condition for (houses in) locations *i* and *j* to be substitutable is that the spatially varying means of the hedonic characteristic *x* and the *β*'s in the two locations match, and their slopes match as well. This implies spatial clustering of the *x*'s and the *β*'s, together with smoothness assumption (Assumption 2). Thus, two locations are in the same submarket if their *x*'s and *β*'s match, and the locations are close to each other, so that the derivatives also match. In other words, based on the close substitutability definition, submarkets may be delineated by spatial clustering on both these two dimensions at the same time.

Second, the insights can be easily extended to the case of multiple regressors (or hedonic factors). In this case, clustering should include all the included hedonic factors as well as their spatially varying slopes. Third, the methodology in Pryce (2013) is only appropriate if there are no regressors. In this case, the cross-price elasticities will be solely determined by elements of *W*. However, because of spatial nonstationarity, the model will then not offer any useful structural interpretation, and the estimates of elasticities are also likely to be biased.

Then, how can one implement such a procedure for delineating submarkets, in a way that is computationally feasible? For this, we develop a synthesis of several empirical approaches rather than a purely spatial econometric framework. Next, we turn to a discussion of some related methods and our proposed synthesis.

4. A Synthesis of Empirical Approaches

Spatial econometrics aids unique understanding of spatial housing markets, in terms of neighbourhood choice, housing preferences, and the evolution of urban spatial structure. The recent literature has considered unknown and endogenous spatial weights (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013; Bailey *et al.*, 2014; Kelejian and Piras, 2014) as well as interconnections between spatial heterogeneity, spatial dependence and spatial scale (Bhattacharjee *et al.*, 2012). Thus, there are important implications for place based urban planning and housing policy, informed by a clear understanding of the links between space and housing.

4.1. The limits of spatial econometrics?

There are, however, two leading aspects where the framework needs to be extended and enhanced. First, while the above framework uniquely combines spatial heterogeneity and spatial dependence, the way spatial dependence is modelled is somewhat unsatisfactory. Specifically, in restricting spatial spillovers to a spatial error model, adequate attention is not paid to endogenous evolution of space itself. At the same time, it is perhaps inevitable that housing markets are endogenously related over space. Location choices and consequently prices are not only spatially contingent, but also potentially directly connected, which implies that spatial dependence through a spatial lag model is more appropriate. While traditional research has paid elaborate attention to spatial lag dependence, for example, Anselin and Lozano-Gracia (2008) and Anselin *et al.* (2010), this has been in a context where the spatial weights are known *a priori*, and there is no spatial heterogeneity. When these spatial interactions are unknown and are themselves objects of inference, endogeneity issues become quite complex. Here, a new framework for modelling and analysis is required.

Second, the above framework assumes a segmentation into housing submarkets which is given *a priori*. A better alternative method is the delineation of submarkets based on hedonic characteristics and prices that are spatially heterogeneous within a spatial context where spatial dependence is endogenous. Once again, this requires a new framework.

4.2. Some alternate approaches

We now turn to alternative perspectives from the geography and statistics literatures, specifically local regressions (for example, geographically weighted regressions, GWR), functional data analysis (FDA) and spatial statistics.

4.2.1. Geographically (or locally) weighted regression

In the literature, spatial heterogeneity is typically modelled using locally weighted regressions (McMillen, 1996), of which the Geographically Weighted Regression (GWR) approach (Fotheringham *et al.*, 1998) is perhaps the most popular.[8] GWR replaces the single regression coefficient in a linear model with a series of (geographically weighted) estimates for a number of spatial data points. This provides a range of location-specific parameter estimates that can be mapped. This methodology arguably provides the best practice in understanding relationships that vary over space. GWR is a kernel based nonparametric regression method where

$$E\left[\int_S Y(s) f_{h,i}(s)ds\right] = \alpha_i + \beta_i \int_S X(s) f_{h,i}(s)ds, \tag{8}$$

where Y and X are both defined over a territory S determined by a medium or large urban housing market, i is a location within the spatial domain S, $f_{h,i}\ (s)$ is a kernel density with bandwidth h and centred on location i, the regression slope β_i varies over space, and α_i can be interpreted as a location specific fixed effect. In effect, this method provides pointwise estimates β_i of the regression effect of a kernel weighted local average of Y on a similarly kernel weighted local average of X.

4.2.2. Functional data analysis

Functional data analysis (FDA) is a framework and collection of tools for statistical analysis of functional data, which refers to curves, surfaces or anything else that varies over a continuum; see Ramsay and Silverman (2005, 2006) for extensive book-length discussions. The continuum is often taken as time, but may also be spatial location, wavelength, probability, etc.

The main challenge in FDA is that functional data (curves or surfaces) are intrinsically infinite dimensional, even when sample sizes are limited. Hence these data have to be

[8] Two other alternatives are not considered here, First, spatial expansion method (EM) is a popular method where regression coefficients are estimated as a function of other locational attributes (such as longitude and latitude) (Cassetti, 1972). This is somewhat restrictive compared to GWR which produces more complex variability over space (Fotheringham *et al.*, 1998). Second, LeSage (2004) proposed Bayesian GWR inference. However, our work here is in a classical domain. We have neither suitable priors nor adequate Bayesian methods applicable to the functional regression model in a spatial context.

projected on the span of a limited basis, assuming that the data are intrinsically smooth, while observed data are potentially bumpy because of measurement error. For such smoothing, FDA often makes use of the information in the slopes and curvatures of curves, as reflected in their derivatives. Plots of first and second derivatives as functions of the continuous domain, or plots of second derivative values as functions of first derivative values, may reveal important aspects of the data generating processes. Hence, curve estimation methods designed to yield good derivative estimates often play a critical role in functional data analysis.

In the typical application where the functional domain is time, inferences are based on projection to a basis space. Typically, a Fourier basis is used for periodic data or smoothing splines for data that are not periodic (Ramsay and Silverman, 2005, 2006). In our spatial context, the functional linear regression model takes the form:

$$E[y_i] = \alpha + \int_S \beta(s) x_i(s), \qquad (9)$$

where the response (*y*) is scalar, and the regressor (*x*) and slope (*β*) are functional.

The domain of our functional data is not time, but two-dimensional space. This makes application of FDA in our context more challenging. Spatial data, unlike time series, has no well-defined ordering of observations, and neither a sense of progression. This makes construction of a suitable basis function difficult. A Fourier basis may not be suitable, and there is no well-defined extension of the spline basis to two dimensional space. Therefore, we consider the basis space given by functional principal components, and adapt to our spatial context an intuitive and powerful estimator proposed by Cai and Hall (2006) and Hall and Horowitz (2007).

4.2.3. Spatial statistics

Guillas and Lai (2010) have also proposed a methodology for functional linear regression based on bivariate splines over triangulations which we intend to explore in future work. However, the bivariate spline approach is not entirely satisfactory because it does not take into explicit consideration the spatial context of the housing market application, in terms of the geography of the region under study and spatial dependence. This is the explicit domain of spatial statistics, which is a collection of methods and tools for quantitative analysis of spatial data and the statistical modelling of spatial variability and uncertainty.

The literature of spatial statistics is large and has substantial intersection with spatial econometrics. Our specific focus lies on multivariate spatial models, which has in recent years proven an effective tool for analysing spatially related multidimensional data arising from a common underlying spatial process. Many spatial problems, including the housing market application here, are inherently multivariate, meaning that two or more variables are recorded at each spatial location simultaneously. With rapid enhancements in geographic information systems (GIS) technology that enables us to analyse and display such data at varying spatial resolutions, multivariate spatial analysis is becoming more relevant and popular.

Sain and Cressie (2007) viewed the developments of spatial analysis in two main categories: models for geostatistical data (that is, the indices of data points belong to a continuous set) and models for lattice data or spatial grids (data with indices in a discrete or countable set), while specifically mentioning that research in the latter direction is not equally well developed. Which category our domain would lie in partly depends on the data generating process. However, most spatial housing market data are aggregated into specific prespecified spatial grids, however fine the scale of these grids may be. In this sense, our data belongs to the latter category.

There is a substantial and rapidly expanding literature on spatial grids. Most of this literature assumes a multivariate generalized linear mixed model with the spatial effects modelled by a multivariate Gaussian conditional autoregressive (CAR) model (Besag, 1974; Mardia, 1988). These models are quite similar to those used in spatial econometrics, but with a spatial weights matrix that is often even more rigidly defined – typically as a row normalized version of an adjacency matrix.

Bayesian inferences on this model have considered spatial clustering; see, for example, Knorr-Held and Raßer (2000) and Booth *et al.* (2008). Given the implications of Theorem 1, this literature is of potential interest in our context. However, suitable classical inferences on spatial clustering are yet to be developed, beyond the obvious step of adding location co-ordinates to the list of variables to be clustered (in this case, x and β). Improved small area methods may also be useful in this context; see, for example, Hall and Maiti (2006).

4.3. A Proposed Synthesis of Different Perspectives

Here, we propose a new framework, based on a synthesis of spatial econometrics, functional data analysis and spatial statistics for the analyses of housing markets. This framework addresses some of the limitations of the previous approaches.

Intuition suggests that such a synthesis may be promising. For illustration, consider again the simple spatial lag model in (6) specific to two housing properties, but incorporating heterogeneity in slopes. As discussed above, the reduced form

$$\begin{pmatrix} y_i \\ y_j \end{pmatrix} \approx [I+W] \begin{pmatrix} x_i\beta_i \\ x_j\beta_j \end{pmatrix} + [I+W] \begin{pmatrix} \varepsilon_i \\ \varepsilon_j \end{pmatrix}. \qquad (10)$$

implies a regression model where, in addition to $x_i\beta_i$, the right hand side also includes $x_j\beta_j$, but with a much smaller weight, since $w_{ji} << 1$. This in turn suggests the functional regression model where the response variable is scalar and the functional regressor is $x_i(s) = x_i\, f_{h,i}(s)$ with kernel weights $f_{h,i}(s)$ proportional to the elements of $[I-W]^{-1} \approx [I+W]$. This intuition easily generalises to multiple locations and houses.

Thus, the spatial lag model is a special case of the functional regression model, corresponding to a particular definition of the functional regressor. The above representation is also very similar to GWR, in the sense that, as the bandwidth h goes close to zero, GWR and functional regression becomes very similar. This suggests that a synthesis of perspectives from spatial econometrics, GWR and FDA offers a spatial statistical model that is likely to be very rich and enable the full range of spatial analyses of housing markets. Importantly, the model offers efficiency and robustness by using information from neighbours through the spatial weights matrix.

The above model addresses both the limitations of the previous approaches. First, the proposed framework takes the regressor (x_i) and (kernel) spatial weights, $f_{h,i}(s)$, together, and combines these into a functional regressor, $x_i(s)$. This allows inference on a functional slope to proceed beyond the limitations of exogenously specified submarkets or spatial weights. Now, endogeneity in the weights can in principle be addressed in standard ways. One would need either a dynamic model for how these weights evolve over time, or use suitable instruments for $x_i(s)$.

Second, the framework allows submarkets to evolve endogenously and abstracts from the requirement to delineate housing submarkets *a priori*. As discussed before, one view posits a submarket is characterised as a collection of locations that have a similar bundle "quality", that is, close hedonic substitutability (Rothenberg *et al.*, 1991;

Bourassa *et al.*, 2003). Alternatively, submarkets may be defined by housing units that are closely substitutable (Grigsby, 1963; Pryce, 2013).

In the context of a hedonic model with homogenous slopes, the two definitions are equivalent. However, this is not true when there is heterogeneity across submarkets. This heterogeneity is not only in hedonic characters, but equally importantly in the shadow prices assigned to such features. As our Theorem 1 suggests, in the presence of such heterogeneity, housing submarkets should be delineated by spatial clustering jointly of the functional partial effect *β(s)* and the functional surface of the hedonic characteristics *x(s)*.

5. Methodology

The precursor to all the above analyses, beginning with delineation of submarkets is the estimating of a functional regression model

$$\begin{aligned} E[y_i] &= \alpha(s) + \int_S \beta(s) x_i(s), \\ x_i(s) &= x(s) f_{h,i}(s), \end{aligned} \tag{11}$$

where y_i is a scalar response at location *i*, *x(s)* is defined over a spatial domain *S* corresponding to a medium or large urban housing market (and, unlike the typical functional regression model, not a subset of the positive real line $\mathbf{R}^+$), and $f_{h,i}(s)$ is a kernel density with bandwidth *h* and centred on location *i*.

5.1. Estimating the Functional Regression Model

The functional linear regression model (11) is based on a large (and potentially infinite) dimensional functional regressor that needs to be regularised. Hence, estimation proceeds by projection to a suitable basis space. We choose a method based on functional principal components (FPC) developed in Cai and Hall (2006) and Hall and Horowitz (2007), which we find intuitively appealing in the current context.[9]

However, estimating the model using functional principal components presents some challenges in our setting. For a specific location *i*, the functional surface of $x_i(s)$ is a weighted form of x_i, with the weights given by a kernel $f_{h,i}(s)$. This kernel places a large weight in the neighbourhood of location *i*, but relatively small weights elsewhere. This implies that the functional surface has very sparse information which in turn requires a large number of principal components and also produces a poor approximation. Econometrically, this is a problem of regularisation. Here, we develop a variant of functional principal components that works in this situation.

Our data generating process is as follows. The data constitute a collection of dependent pairs $(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)$ indexed on *n* locations on a compact space $S \subset \mathbf{R}^2$. For a specific location $i \in S$, both *Y* and *X* are scalar random variables. The response variables Y_i are generated by a functional linear regression model

$$\begin{aligned} Y_i &= \alpha + \int_S \beta X_i^* + \varepsilon_i, \qquad i = 1,\ldots .n, \\ X_i^*(u) &= \begin{cases} X_i & \text{if} \qquad u = i \\ X_j f_{ij} & \text{if} \quad u = j \in \{1,\ldots,n\}, i \neq j, f_{ij} = f_{h,i}(j) \\ 0 & \qquad \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

[9] There are other popular choices of basis space, such as a Fourier basis, spline smoothing and, in the specific case of a spatial domain, the bivariate spline.

The errors ε_i are potentially spatially dependent, but are identically distributed with finite variance and zero mean. The errors are also independent of the explanatory variables, but there no distributional assumptions are required.

The main problem with model (12) is that the functional regressor surface of X_i^* is very irregular. Even if the mean of the underlying regressor, $\overline{X}(u)$, has smooth variation over the set *S*, the combination of a large (unit) weight at location *i* with a kernel function elsewhere renders X_i^* very irregular and spiky. Hence, usual regularisation by principal components as in Cai and Hall (2006) and Hall and Horowitz (2007) is almost impossible. The tuning parameter in Hall and Horowitz (2007) required here will be very large, and spacings between eigenvalues will be very small, so that the results in Hall and Horowitz (2007) are not directly applicable.

Hence, our approach focuses on directly regularising the surface of $\overline{X}(u)$ using functional principal components. To motivate the approach, consider the surface of the functional regressor X_i^* for the specific observation *i*. The challenge here is the spiky nature of X_i^*, due to a very large (unit) weight at the location of observation *i*, together with much lower weights (f_{ij}) at other locations. Our object of inference here is the functional surface of the regression coefficient (β) which we have assumed to be relatively smooth (Assumption 2). Hence, the regressor at this location can be potentially combined with values in its neighbourhood. By averaging, the irregularity of the functional regressor surface can therefore be reduced. This suggests that partitioning *S* into several (*K*) regions (say, $\{P_1, P_2, \ldots, P_K\}$) may be a good starting point. This approach may also be viewed as a first stage of regularisation, where the basis function is a histogram sieve.

The above approach is in line with functional data where the regressor is observed at a (large) number of fixed time points. However, since the number of such partitions is typically very large, and can in principle even exceed the sample size (*n*), a second stage of regularisation is required on the averaged regressor process. In our case, we use functional principal components on the averaged regressor process across the partitions, that is on $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$, where $\bar{x}_k = E(X_i \mid i \in P_k), k = 1, \ldots, K.$ The procedure poses two major challenges: (a) by averaging, we would lose variability across observations, and therefore implementation of functional principal components is challenging; and (b) if we were to implement principal components, we need to develop a method similar to Hall and Horowitz (2007) to then use these principal components to estimate the functional surface of the regression coefficient (β)?

For (a), the same spike that was a problem earlier now helps once a histogram sieve (partition) has been placed. To see this, consider the compact space *S* partitioned into *K* regions $P_1, P_2, \ldots, P_K$, with corresponding sample sizes $n_1, n_2, \ldots, n_K$, with $\sum n_k = n.$ With a small abuse of notation to simplify expressions, we denote by *k(i)=k* the partition that observation *i* belongs to, that is *i* ϵ P_k. Then, the sieve functional regressor for observation *i* is

$$[n_1 f_{i1} \bar{x}_1, n_2 f_{i2} \bar{x}_2, \ldots, n_{k-1} f_{i,k-1} \bar{x}_{k-1}, [(1 - f_{ii}) X_i + n_i f_{ii} \bar{x}_k], n_{k+1} f_{i,k+1} \bar{x}_{k+1}, \ldots, n_K f_{iK} \bar{x}_K]. \quad (13)$$

Dividing the *j*-th element of the functional regressor vector (13) by the scalar exogenous weight $n_j f_{ij}$, we have

$$X_i^{**} = \left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{i-1}, \bar{x}_i + \frac{1 - f_{ii}}{n_i f_{ii}} X_i, \bar{x}_{i+1}, \ldots, \bar{x}_K \right]. \quad (14)$$

Because there is now variation in the functional regressor surface across observations within each partition as well, functional principal components can be implemented. At the same time, $\frac{X_i - f_{ii}\bar{x}_i}{n_i f_{ii}} \to 0$ as $n \to \infty$, so that in large samples, (14) approximates the average process. Thus it is expected to be smooth over space since, by Assumption 2, the functional surface of the average $\overline{X}(s)$ is smooth. In large samples, when variation in X_i does not matter for the construction of the functional regressor, $X^* \approx \overline{X}Z$, where $\overline{X} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K]$ and Z is a vector that takes value 1 at location *i* and 0 otherwise.

With (b) note that, for the data within a specific partition P_k, the functional regression coefficient for the partition is β_k times $n_k f_{kk}$, where the coefficient itself corresponds to the *k*-th element of X^{**}. Note also that, within this same partition, there is no cross-section variation in the other elements of X^{**}, and hence their effects are encompassed within a fixed effect for the partition. Hence, the entire functional surface of the regression coefficient can be estimated by a functional regression model where the dependent variable is measured in deviations from the local (within partition) mean, and the functional regressor is given by equation (14).

In our application, we have a finite but large-dimensional setting where the number of partitions (*K*) is large. Below, we assume that the spatial design, given by *Z*, is held fixed in repeated sampling. We obtain a final estimator $\hat{\beta}$ by dividing the *k*-th element of the functional regression estimator by the known deterministic scalar $n_k f_{kk}$.

Thus, consider the modified linear functional regression model:

$$Y_i = a + \int_S b(u) X_i^{**}(u)\,du + \varepsilon_i, \qquad i = 1, \ldots n,$$

$$X_i^{**}(u) = \begin{cases} \overline{X}(u) + \dfrac{1 - f_{ii}}{n_i f_{ii}} X_i & \text{if} \qquad u = i \\ \\ \overline{X}(u) & \text{otherwise.} \end{cases} \tag{15}$$

The X_i^{**}'s are random functions, $S \subset \mathbf{R}^2$ denotes a compact set on which each such function is defined, the intercept *a* and the errors ε_i are scalars and the slope *b*, the main object of inference, is a function.

Let (X^{**}, *Y*, ε) denote a generic (X_i^{**} , Y_i, ε_i). Define

$$K(u,v) = \operatorname{cov}\{X^{**}(u), X^{**}(v)\}, \qquad u, v \in S,$$

$$\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{X_i^{**}(u) - \overline{X}^{**}(u)\}\{X_i^{**}(v) - \overline{X}^{**}(v)\},$$

where $\overline{X}^{**}(.) = n^{-1}\sum_i X_i^{**}(.)$. Write the spectral expansions of *K* and $\hat{K}$ as:

$$K(u,v) = \sum_{j=1}^{\infty} \kappa_j \phi_j(u)\phi_j(v), \qquad \hat{K}(u,v) = \sum_{j=1}^{\infty} \hat{\kappa}_j \hat{\phi}_j(u)\hat{\phi}_j(v),$$

where $\kappa_1 > \kappa_2 > \ldots > 0$ and $\phi_1, \phi_2, \ldots$ are the eigenvalue and corresponding orthonormal eigenvector sequences of the linear operator with kernel *K*, and similarly $\hat{\kappa}_1 \geq \hat{\kappa}_2 \geq \ldots \geq 0$ and $\hat{\phi}_1, \hat{\phi}_2, \ldots$ for the kernel $\hat{K}$. The sequences $(\hat{\kappa}_j, \hat{\phi}_j)$ of eigenvalues and eigenvectors of the empirical covariance matrix of X^{**} constitute an estimator of (κ_j, ϕ_j). Then, the functional principal components estimator (Cai and Hall, 2006; Hall and Horowitz, 2007) of the regression slope the slope *b(.)* is given by

$$\hat{b}(u)=\sum_{j=1}^{m}\hat{b}_j\hat{\phi}_j(u), \tag{16}$$

where the spectral cutoff *m* is a tuning parameter, $\hat{b}_j=\hat{\kappa}_j^{-1}\hat{g}_j$, $\hat{g}_j=\int\hat{g}\hat{\phi}_j$, and

$$\hat{g}(u)=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\bar{Y}\}\{X_i^{**}(u)-\bar{X}^{**}(u)\}.$$

Note that the functional regression estimator at (16) is a least squares estimator, depending only on the sample covariance function of the functional regressor X^{**}, truncated at a finite cutoff for the spectral expansion, and the covariance function of *Y* and X^{**}. Thus, it is essentially a method of moments estimator that requires neither independent errors nor a specified error distribution. Instead, it is based on mean zero errors and orthogonality of the regressor and the errors. This is useful in our context.

Note that, the errors in our reduced form spatial model (7) are correlated. Further, the functional regressor in our spatial setting can be endogenous, either because of endogenous (or estimated) spatial weights, or because the underlying regressor *X* is itself endogenous. In such cases, an instrumental variables estimator can be constructed along the above lines. Finally, note that, given the nature of our problem, *Y* is measured in terms of deviation from the local (within partition) mean, thus allowing for spatial fixed effects.

Next, we make assumptions required for consistency and convergence rates of the above functional principal components estimator.

Assumption 3: ***Technical assumptions for functional regression inference.***

(a) *The data are generated by fixed spatial design, so that Z is not stochastic.*

(b) *All other technical assumptions in Hall and Horowitz (2007) hold. Specifically, conditions on the distribution of X^{**} (the spatial functional regressor), distribution of ε, eigenvalues and Fourier coefficients hold.*

i) *X has finite fourth moments, and hence so does X^{**}. The error ε_i are identically distributed with zero mean and finite variance not exceeding some constant C.*

ii) *Consider the Karhunen-Loève expansion of the random function X^{**}: $X^{**}-E(X^{**})=\sum_{j=1}^{\infty}\xi_j\phi_j$, where the ξ_j are pairwise uncorrelated random variables that have zero means and variances κ_j that are eigenvalues of the expansion. The κ_j satisfy the spacing condition $\kappa_j-\kappa_{j+1}\geq C^{-1}j^{-\alpha-1}$ for all j and some exponent α > 1.*

iii) *Let $b_j=\kappa_j^{-1}g_j$ where $g(u)=E[(Y-EY)(X(u)-EX(u))]$ and $g_j=\int g\phi_j$. The b_j's satisfy $|b_j|\leq Cj^{-\delta},\delta>\frac{1}{2}\alpha+1$.*

iv) *The tuning parameter m increases with n such that $m/n^{1/(\alpha+2\delta)}$ is bounded away from zero and infinity.*

Then, the functional surface *b(s)* can be estimated by the functional principal components estimator in Hall and Horowitz (2007). However, ultimately our object of inference is the functional surface of *β(s)* in (12) and not the *b(s)* in (15). Assumption 3(a) provides a simple way to go from the $\hat{b}(s)$, estimated by functional principal components as in (16), to the $\hat{\beta}(s)$. In the finite but large dimensional setting, or equivalently with a histogram sieve placed on the spatial domain, one simply has $\hat{\beta}(s) = \hat{b}(s)/(n_k f_{kk}), s \in P_k$.

Assumptions 3(b) are explicitly stated in Hall and Horowitz (2007). Condition 3(b)i) is standard. The condition (b)ii) ensures that all eigenvalues have unit multiplicity, and their spacing decreases exponentially, so that the estimator can be obtained with a small smoothing spectral cutoff. Assumption (b)iii) ensures that the Fourier coefficients are bounded below and above. Condition (b)iv) ensures that the number of basis function terms used in the smoothing process of *b* is much lower than *n*.

Then, we have the following result.

Theorem 2 (Hall and Horowitz, 2007): *Let $\Im(C, \alpha, \delta)$ denote the set of distributions F of (X^{**}, Y) that satisfy Assumption 3 for given values of C, α and δ. Let B denote a class of measurable functions $\overline{b}$ of the data $(X_1^{**}, Y_1), \ldots, (X_n^{**}, Y_n)$ generated by (15). Then, $\hat{b}(s) \overset{P}{\rightarrow} b(s)$. Specifically,*

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{F\in\Im} P_F \left\{ \int_S \left(\hat{b} - b\right)^2 > Dn^{-(2\delta-1)/(\alpha+2\delta)} \right\} = 0 \quad \text{as} \quad n \to \infty$$

and

$$\liminf_{n\to\infty} n^{(2\delta-1)/(\alpha+2\delta)} \inf_{\overline{b}\in \mathrm{B}} \sup_{F\in\Im} \int_S E_F \left(\overline{b} - b\right)^2 > 0,$$

which then imply that for each $F \in \Im$,

$$\int_S \left(\hat{b} - b\right)^2 = O_p\left(n^{-(2\delta-1)/(\alpha+2\delta)}\right).$$

For the technical details of the proof and associated discussion, refer to Hall and Horowitz (2007). The functional principal components estimator is a method of moments estimator. The identical distribution condition for the errors is stated in Assumption 3(b)i), but is not required in the proof of Theorem 2 beyond the zero covariance between X^{**} and *ε*. The technique of proof is somewhat nonstandard, in showing that the supremum and infimum have the same rate of convergence, and moreover in using probability (P_F) rather than expectation (E_F) in the supremum statement. The rate of convergence $n^{-(2\delta-1)/(\alpha+2\delta)}$ is generic to noisy inverse problems.

The main result was shown in Hall and Horowitz (2007). Our main innovation here is to adapt the above general result to a spiky functional regressor surface. This we achieve by using a histogram sieve.

Corollary 1: *Under Assumption 3, $\hat{\beta}(s) \overset{P}{\rightarrow} \beta(s)$ and for each $F \in \Im$,*

$$\int_S \left(\hat{\beta} - \beta\right)^2 = O_p\left(n^{-(2\delta-1)/(\alpha+2\delta)}\right).$$

Proof: *The proof follows directly from Theorem 1, noting that by Assumption 3(a), $n_k f_{kk}$ is a fixed scalar. Since $\hat{\beta}(s) = \hat{b}(s)/(n_k f_{kk}), s \in P_k$, the result follows.*

Intuitively, the above result may hold and estimator would be well-defined even if we had random (but independent) sampling over space. In this case, *Z* would have a multinomial distribution over spatial partitions (or sieves), and spatial stationarity would then imply that *plim* $n_k f_{jk} = w_{jk}$, for some inter-partition spatial weights matrix $W_{(K\times K)}$. In particular, *plim* $n_k f_{kk} = w_{kk}$, where w_{kk} is the intra-partition spatial weight in partition P_k.[10] However, relaxing the potentially strong Assumption 3(a) appears to be technically difficult, and is retained for future work.

As discussed above, the functional principal components estimator is essentially a method of moments estimator of the functional regression coefficient *b(s)* based on orthogonality of the error ε with both *X* and *W*. This estimator is consistent only when *W* is exogenous. If *W* is endogenous, then an instrument is required. Such a functional instrument *V* has to be strictly exogenous but correlated with the functional regressor X_i^{**}. The instrument *V* may be based, for example, on a weights matrix where the elements are functions of geographic distances, which are exogenous by construction.

Kelejian and Piras (2014) consider an application to demand for cigarettes in the USA, where consumers living close to the border of a state can travel some distance into the neighbouring state to buy their tobacco. However, they would do so only if the travel distance is small and the prices in the neighbouring state are lower. This implies a weights matrix that is a combination of geographic distances and prices, and is endogenous because prices are endogenous. Endogenous spatial weights can also arise if the weights matrix is estimated using the same data. For example, in the context of our application here, a natural choice is the estimator of a symmetric spatial weights matrix proposed in Bhattacharjee *et al.* (2012).

The natural extension of the above estimation to the endogenous functional regressor case is based on the covariance function of a suitable functional instrument *V*. Note that, in the case of simple linear regression where the OLS is given by $\hat{b}_{OLS} = Y'X / X'X$, the corresponding IV estimator is $\hat{b}_{IV} = Y'V / X'V$.

As before, define

$$\hat{K}_V(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{V_i(s) - \bar{V}(s)\}\{V_i(t) - \bar{V}(t)\} = \sum_{j=1}^{\infty}\hat{\kappa}_j^{(V)}\hat{\phi}_j^{(V)}(s)\hat{\phi}_j^{(V)}(t);$$

$$\hat{g}_{Y,V}(s) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \bar{Y}\}\{V_i(s) - \bar{V}(s)\}; \text{ and}$$

$$\hat{G}_{V,X^{**}}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{X_i^{**}(s) - \bar{X}^{**}(s)\}\{V_i(t) - \bar{V}(t)\}.$$

Then, the following functional principal components IV estimator of the regression slope *b(.)* can be proposed:

$$\hat{b}_{IV}(s) = \frac{\sum_{j=1}^{m}\hat{b}_j\hat{\phi}_j(s)}{\sum_{j=1}^{m}\hat{B}_j\hat{\phi}_j(s)}, \tag{17}$$

where *m* is the spectral tuning parameter, $\hat{b}_j = \hat{\kappa}_j^{-1}\hat{g}_j$, $\hat{g}_j = \int \hat{g}_{Y,V}\hat{\phi}_j$, and

$$\hat{B}_j = \hat{\kappa}_j^{-1}\hat{G}_j, \quad \hat{G}_j = \int\int \hat{G}_{V,X^{**}}(s,t)\hat{\phi}_j(s)\hat{\phi}_j(t).$$

The numerator in (17) is the principal components estimator (Hall and Horowitz, 2007) of the functional regression of *Y* on the functional instrument *V*, and the denominator an estimator for the regression of X^{**} on *V*. Thus, under a suitable instrument validity condition, the estimator in (17) will be consistent for *b(s)*.

[10] The spatial weights matrix *W* has zero diagonal elements. Note however that, in a partitioned spatial domain (or sieve) context, there will be non-zero intra-partition spatial weights reflecting the interaction between different units (houses) within the same partition; see Bhattacharjee *et al.* (2012).

Assumption 4: ***Technical assumptions for functional IV regression inference.***

(a) *Instrument validity:*

$$G(s,t) = E[X^{**}(s) - EX^{**}(s)]\,[V(t) - EV(t)] \neq 0 \text{ for all } (s,t) \in S \times S.$$

(b) *Technical assumptions in Hall and Horowitz (2007) hold for both the functional regressions: Y on V, and X^{**} on V.*

i) *V has finite fourth moments. The errors in the above two regressions are identically distributed with zero mean and finite variance not exceeding some constant C.*

ii) *Consider the Karhunen-Loève expansion of the random function V: $V - E(V) = \sum_{j=1}^{\infty} \xi_j \phi_j$, where the ξ_j are pairwise uncorrelated random variables that have zero means and variances κ_j that are eigenvalues of the expansion. The κ_j satisfy the spacing condition $\kappa_j - \kappa_{j+1} \geq C^{-1} j^{-\alpha-1}$ for all j and some exponent α>1.*

iii) *Let $b_j = \kappa_j^{-1} g_j$ and $B_j = \kappa_j^{-1} G_j$, where $g_j = \int g\phi_j$,* $g(u) = E[(Y - EY)(V(u) - EV(u))]$ and $G_j = \int\int G(s,t)\phi_j(s)\phi_j(t)$. *The b_j's and B_j's satisfy* $\max\{|b_j|, |B_j|\} \leq Cj^{-\delta}, \delta > \frac{1}{2}\alpha + 1$.

iv) *The tuning parameter m increases with n such that $m / n^{1/(\alpha+2\delta)}$ is bounded away from zero and infinity.*

Assumption 4(b) are very similar to Assumption 3. Assumption 4(a) imposes non-zero covariance between the functional regressor and the instrument everywhere over the spatial domain *S*. This assumption can be relaxed at the cost of analytical complexity.

Corollary 2: *Under Assumption 4, $\hat{\beta}_{IV}(s) \xrightarrow{P} \beta(s)$ where $\hat{\beta}_{IV}(s) = \hat{b}_{IV}(s)/(n_k f_{kk}), s \in P_k$.*

Proof: *By Theorem 1, the numerator and denominator of $\hat{b}_{IV}(s)$ converge to the respective functional regression coefficients, and hence the ratio converges in probability. That is, $\hat{b}_{IV}(s) \xrightarrow{P} b(s)$. Then the proof follows as in Corollary 1, noting that $n_k f_{kk}$ is a fixed number.*

Optimal choice of instruments possible in this context, and weak instrument robust inference is a potential area of future research. Also, there can be alternate estimators. For example, we can define an estimator

$$\hat{b}(s) = \sum_{j=1}^{m} I\left\{\left|\hat{B}_j\right| > \Delta\right\} \frac{\hat{b}_j}{\hat{B}_j} \hat{\phi}_j(s),$$

where *m* is the spectral tuning parameter and $\Delta > 0$ is a cutoff for the estimated covariance between X^{**} and *V*. Technical details relating to statistical properties of this estimator need further work, and is left outside the purview of the current paper. The main challenge lies in dealing with sample covariances $\hat{B}_j$ that are close to zero.

In the remainder of this paper, including the empirical application, we focus on an exogenous weights matrix. However, as discussed above the proposed framework based on the functional regression model allows us to construct estimators that allow for potential endogeneity of spatial structure. This extends the literature substantially.

The traditional spatial econometrics literature has focussed either on spatial dependence or on spatial heterogeneity, but not both of these aspects together. The recent literature has developed methods for estimating the spatial weights (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013; Bailey *et al.*, 2014), as well as inferences where the spatial weights are endogenous and known *a priori* (Kelejian and Piras, 2014). By contrast, this paper presents inferences for endogenous (and potentially unknown) spatial structure together with spatial heterogeneity.

5.2. Implementation of the Functional Principal Components Estimator

Our objective is to estimate a spatial lag model with spatial heterogeneity in the slope of a specific regressor, with spatial weights defined exogenously. In our empirical application, we define spatial weights by a kernel function, as in (11). Inferences are conducted by expressing the spatial lag model in reduced form as a functional regression model (6), where the functional regressor has the form given by (12).

First, the spatial domain *S* is partitioned into a large number (*K*) of small areas, denoted $\{I_1, I_2, \ldots, I_K\}$. Next, we obtain average values of the hedonic characteristic in each of these *k* locations, combined into a spatial vector $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$. Finally, we conduct functional principal components on this vector of spatial averages. However, the above vector does not have any cross section variation. This is because the cross section variation in x_i is sacrificed in the process of aggregation by averaging. To recover this information, we replace $\bar{x}_k$, for observation $i \in I_k$, with

$$X_k^{**} = \bar{x}_k + x_i \frac{1-f_{0i}}{n_k f_{0i}} = \frac{1}{n_k f_{0i}} \left[x_i (1 - f_{0i}) + n_k f_{0i} \bar{x}_k \right],$$

where $f_{0i} = f_{h,i}(I_k)$ is the modal kernel density centred on the location of *i*, and n_k denotes the sample size in partition I_k.

Correspondingly, we transform the response variable (*y*) into local mean deviations: $y_i^* = y_i - \bar{y}_k,\ i \in I_k$. Then, functional regression proceeds by obtaining a small number of functional principal components and regressing the transformed response variable (y^*) on these functional principal components.

The steps of the estimation method are as follows:

1. Partition the territory into *k* potential submarkets, denoted $\{I_1, I_2, \ldots, I_K\}$. For each house *i*, identify the partition *j* to which it belongs: $i \in I_k$.
2. Construct functional average surface as

$$\begin{aligned} X_i^{**} &= \left(x_1^{**}, x_2^{**}, \ldots, x_K^{**}\right) \\ &= \left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k + x_i \frac{1-f_{0i}}{n_k f_{0i}}, \ldots, \bar{x}_K \right) \end{aligned}$$

   and the response variable as $y_i^* = y_i - \bar{y}_j,\ i \in I_k$.

3. Conduct functional principal components on X_i^{**}, estimate the *m* principal component factors $(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_m), m << K$, with corresponding eigenvalues $\hat{\kappa}_1 > \hat{\kappa}_2 > \ldots > \hat{\kappa}_m >> 0$.
4. Obtain the functional principal components estimator as (16):

$$\hat{b}(I_k) = \sum_{j=1}^{m} \hat{b}_j \hat{\phi}_j (I_k), \quad k = 1, \ldots, K,$$

where $\hat{b}_j = \hat{\kappa}_j^{-1}\hat{g}_j$, $\hat{g}_j = \int \hat{g}\hat{\phi}_j$, and

$$\hat{g}(I_k) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i^* - \bar{Y}^*\}\{X_i^{**}(I_k) - \bar{X}^{**}(I_k)\}.$$

5. Finally obtain the estimated FPC surface of the functional regression coefficient as $\hat{\beta}(I_k) = \hat{b}(I_k)/(n_k f_{0i})$.

The estimation methodology can now be applied to the data.

5.3. Submarket Delineation by Spatial Clustering

Once the functional regression slope surface is estimated, Theorem 1 suggests using the surface $\hat{\beta}(s)$ and $\bar{x}(s)$ to identify housing submarkets by spatial clustering. The notion of clustering here is also closely related to projections on the effective dimension reduction (EDR) space; see Li and Hsing (2010). Based on the importance accorded to spatiality, there are several ways such clustering can be undertaken: spatial clustering (Knorr-Held and Raßer, 2000); clustering based only on similarities in functional variables (Booth *et al.*, 2008); or clustering based on a combination of spatial proximity and similarity in characteristic space, either by an intersection between both criteria (Feng *et al.*, 2012), or by penalties on smoothness over the spatial domain.

We estimate submarkets using a two-stage procedure. First we estimate the functional surfaces $\hat{\beta}(s)$ and $\bar{x}(s)$, by spatial functional regression and spatial local averaging, respectively. Then, in the second stage, we estimate submarkets by applying Ward's aggregative clustering jointly to $\hat{\beta}(s)$ and $\bar{x}(s)$; see Everitt (1993). Theorem 1 shows that submarket delineation in our context should be conducted by spatial clustering. However, the full development of spatial clustering methods in a spatial functional setting lies outside the domain of the current research, and is retained for future work. Importantly, the clusters estimated in our application are observed to have a strong spatial orientation, which relaxes the necessity to conduct spatial clustering in our case. Hence, we identify submarkets simply by Ward's aggregative method, which at each stage joins the two subclusters that result in the minimum increase in the degree of within-cluster heterogeneity (sum of squares).

Several other lines of methodological development follow from the framework and methods developed here. First, in future research, we plan to conduct inference on spatial structure, that is on an unknown spatial weights matrix *W*. Specifically, the error term of the reduced form spatial functional regression model (7) is $(I - W)^{-1}\varepsilon$. Thus, the error term is spatially correlated, and such spatial correlation can potentially be used to learn about the spatial structure (that is, *W*). Specifically, if the elements of ε were initially uncorrelated, the covariance structure of estimated residuals from the functional regression model can be used to infer on the spatial filtering necessary to reduce the residual vector to white noise, which has a 1–1 correspondence with the unknown spatial weights matrix *W*. Note that, since the functional principal components regression estimator is based on moments, spatial correlation in the reduced form errors is allowed in this setting.

There are other promising alternative approaches to inferences on an unknown *W*. First, the submarkets identified in the previous step can be used to estimate within and between submarket spatial weights using methods similar to Bhattacharjee *et al.* (2012). Second, the methodology in Bhattacharjee and Holly (2013) can be extended to estimate *W* using instruments (and corresponding moment conditions) obtained by using information from finer spatial scales.

Finally, estimation and inferences on spatial structure based on an unknown spatial weights matrix W has another important advantage. In principle, functional analyses on the partial effects and other variables can borrow strength over the network (defined by W) using ideas and concepts from small area statistics; see, for example, Hall and Maiti (2006, 2012). Perhaps most importantly, the proposed framework offers the possibility of studying the endogenous evolution of urban spatial structure. All these lines of future research are exciting.

6. Application to the Aveiro-Ílhavo Urban Housing Market in Portugal

Now, we apply the methodology described in previous sections to analyse housing submarkets in a specific urban housing market – the neighbouring municipalities of Aveiro and Ílhavo located in the Centro (central) Region of Portugal (Figure 1). The municipality of Aveiro has a total area of 200 km² and a total population of 78,454; the municipality of Ílhavo has an area of 75 km^2 and 38,317 inhabitants (Census of Portugal, 2011). Leaving aside the area of the lagoon (the shaded area in Figure 1), the population density is 600 inhabitants per km^2, which is typical for an urban agglomeration in Portugal.

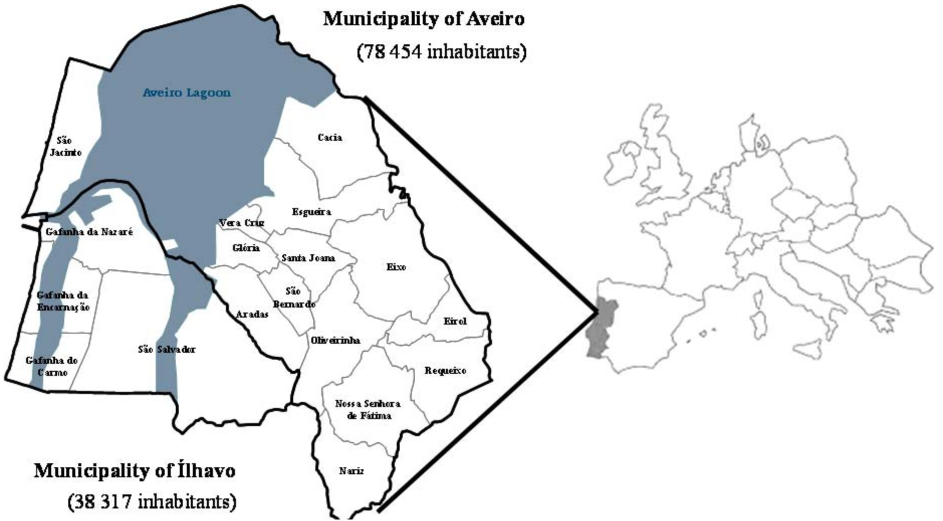


Figure 1 – Location of the study area: Municipalities of Aveiro and Ílhavo

The above spatial domain is divided into the following main zones, each representing aggregation of smaller administrative areas with relatively homogeneous neighbourhoods and house prices (see Figure 2):

i) The inner city of Aveiro, with a population of 32,000 inhabitants, which is the core of the urban municipality.

ii) The smaller city of Ílhavo, with a population of 5,000 inhabitants, which is the second urban centre of the agglomeration.

iii) A semi-rural area with 30,000 inhabitants, where a significant part of the land is used for agriculture, but almost the entire population works in the manufacturing and service sectors, spread all over the urban agglomeration. Housing constitutes a mixture of new urban settlements with blocks of flats and clusters of detached houses and old rural settlements, following a typical local pattern of strings of houses extended along the roads.

iv) A suburban area with 33,500 inhabitants spread around the city of Aveiro, with a settlement and employment pattern similar to the above semi-rural area but with a higher proportion of new urban settlements.

v) Gafanha da Nazaré, with a population of 13,000 inhabitants, is where the port of Aveiro is located. This zone is characterized by a mix of industrial and residential areas.

vi) The seaside resorts Barra and Costa Nova, with a permanent population of 3,000 inhabitants and where secondary residences and holiday rental properties dominate.

As the above description shows, the area, with approximately 116,500 inhabitants, has enough variation over space to enable use of the methods and framework proposed here to delineate the Aveiro-Ílhavo urban housing market into different submarkets.

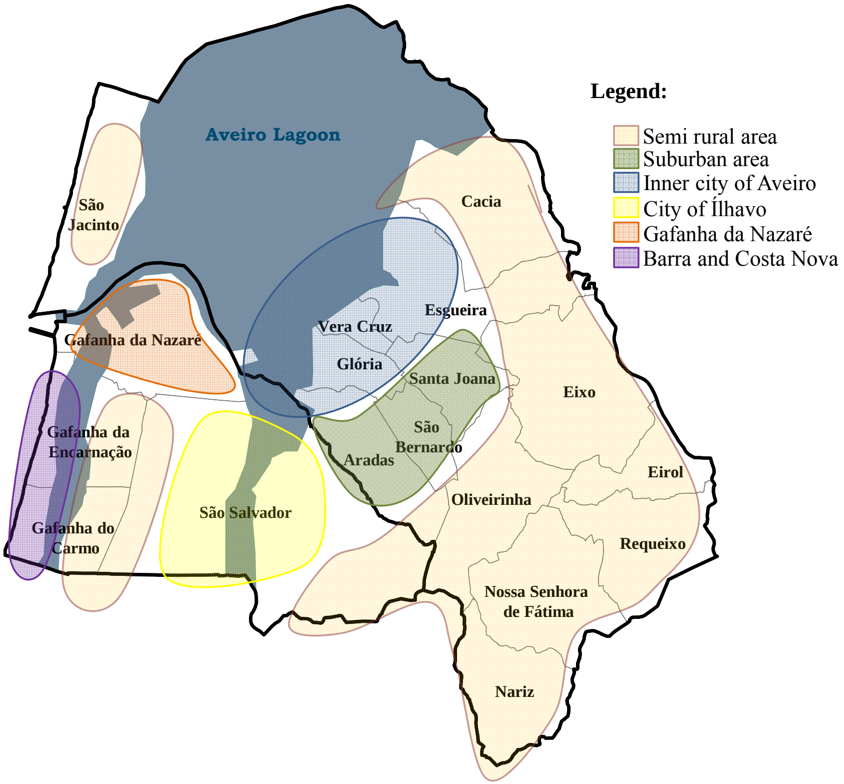


Figure 2 – Major zones of the municipalities of Aveiro and Ílhavo

The database used for this empirical work is provided by the firm Janela Digital S.A., which owns and manages the real estate portal database Casa Sapo. This portal is the largest site in Portugal of real estate advertisement. The data pertain to the time period October 2000 and March 2010 and include around 4 million records of properties available for transaction in Portugal, covering the entire national territory. For the specific case of Aveiro and Ílhavo, the database included 47,188 different properties. This empirical work used 12,467 observations on completed transactions; cases where data were incomplete or inconsistent were removed after careful consideration.[11]

In addition to the price of each property, the database includes two main categories of variables to describe each dwelling: i) the intrinsic physical attributes, and ii) the location and neighbourhood of the building; see Bhattacharjee *et al.* (2012) for a full discussion. The first group includes number of rooms, state of preservation (restoration), age of construction and area (living space, built area, etc.). A set of other physical housing characteristics, obtained from a free text field where real estate

[11] For a detailed description of the main challenges in cleaning the data and construction of housing attribute variables (intrinsic, location and neighbourhood), see Bhattacharjee *et al.* (2012).

advertisers describe the property, was also used. The second group of attributes is related to the housing location and to the characteristics of the neighbourhood, aggregated into a set of distances from different urban, local utility, recreation and transport facilities (Marques *et al.*, 2012).

Since only a small proportion of houses were fully geo-referenced, the houses were placed within into the smallest homogeneous areas that the database can describe, and the centres of such areas were geo-referenced; Figure 3 provides the locations of these 76 areas (which we call zones), which constitute our partition of the spatial domain.

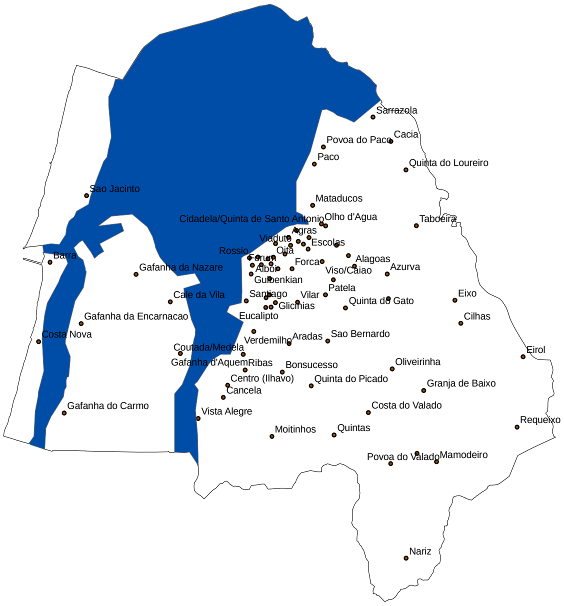


Figure 3 - Housing location by zones

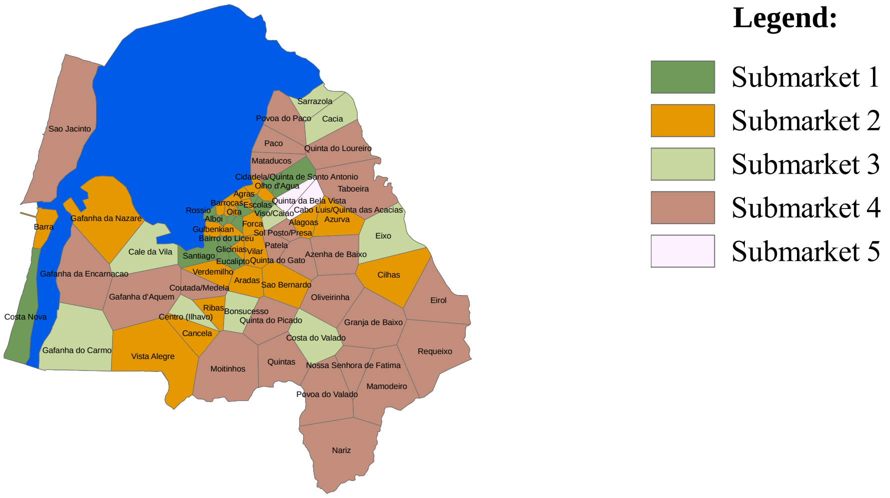
The data reflect large variation in housing characteristics (Bhattacharjee at al., 2012). For example, the average price (in euros per square meter) is 1,126 and its variation ranges from 178 up to 5,714 across the 76 zones. The average dwelling dimension across the 76 zones is 149 m^2, varying between 20 m^2 and 600 m^2.

Of the 12,467 sample dwellings, 28.4% are single houses, 71.6% are flats and 12.3% are duplex (flats with two floors); 39.3% have a balcony, 18.2% have a terrace, 16.1% have garage space, of which 63.8% have a garage; 43.3% have central heating while 28.9% have a fireplace. Location attributes show large spatial variation as well. On average, houses are located at 3.2 km from the CBD, while the maximum distance to the CBD is 16 km.

In order to capture the main dimensions of the housing characteristics, maximum likelihood factor analysis with orthogonal varimax rotation was applied to the hedonic housing attributes. It is important to verify that the extracted factors do not reflect solely statistical properties but behavioural collections of housing characteristics (Maclennan, 1977; Malpezzi, 2003). The factors thus extracted nicely align into housing features. The hedonic features were organised into 5 factors, which together explain 54% of the total variation in 43 hedonic characteristics: 19 internal physical characteristics of a house and 24 location attributes. The factors provide clear interpretation in terms of behavioural collections of housing characteristics: of the 5 factors, 3 relate to location attributes (factor 1 - accessibility to the centre or central amenities; factor 2 - accessibility to local amenities; factor 3 - accessibility to beaches) and the other two represent the intrinsic attributes of dwellings (factor 4 - housing dimension; and factor 5 - additional desirable features).[12]

We estimate a hedonic model using the above data, modeling house prices as a function of living area, the above 5 factors, and time on the market.[13] In this paper, specific attention is focused on living area, the shadow price for which is expected to vary over the spatial domain, and may be considered a good candidate to analyse housing spatial segmentation in the Aveiro-Ílhavo area. Therefore, our functional regression slope, $\beta(s)$, corresponds to living area, and the remaining attributes are assumed to have spatially fixed coefficients and used as control variables.

Our central inference is reported in three maps based on clustering along three different characters that represent the spatial housing segmentation for Aveiro and Ílhavo. For this purpose, cluster analysis (Everitt, 1993) was applied to: i) housing living area (measured in square meters) averaged across all houses within each zone, $\bar{x}(s)$ (Figure 4); ii) the estimated functional regression coefficient $\hat{\beta}(s)$, representing the shadow price of living space, using the methods developed in section 5 (Figure 5); and finally, iii) cluster analysis based on a combination of both $\bar{x}(s)$ and $\hat{\beta}(s)$ (Figure 7). Theorem 1 (section 3) emphasizes spatial clustering, which we do not explicitly apply here. However, the clusters reflect clear spatial concentration.

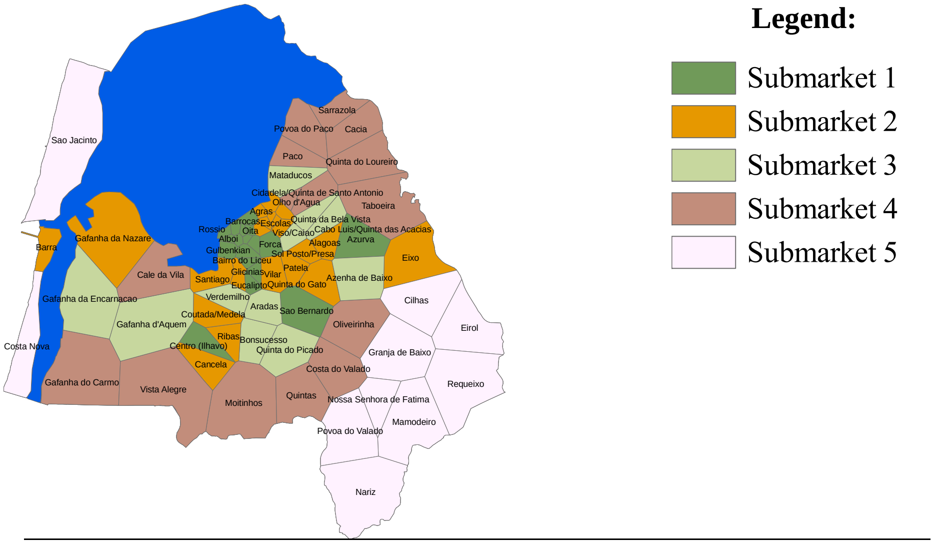


	Number	Mean	Std. Deviation	Mean
	of zones	(ln Area m^2)	(ln Area m^2)	(Area m^2)
Submarket 1	17	4.588	0.090	98.298
Submarket 2	24	4.869	0.085	130.191
Submarket 3	9	5.164	0.047	174.863
Submarket 4	24	5.434	0.113	229.064
Submarket 5	2	6.144	0.010	465.914

Figure 4 - Submarkets based on Ward's linkage clusters: Housing characteristics– living space of house (m^2)

[12] See Bhattacharjee *et al.* (2012) for details on how these factors were constructed and defined. Table 11 in Bhattacharjee *et al.* (2012) reports a detailed description of the factors.

[13] The prices reported in the dataset are asking prices and not final transaction prices. Time on the market is included to capture the wedge between asking and final prices (Bhattacharjee *et al.*, 2012).



	Number of zones	Mean (FDA elasticity)	Std. Deviation (FDA elasticity)
Submarket 1	18	0.845	0.056
Submarket 2	22	0.713	0.026
Submarket 3	11	0.614	0.020
Submarket 4	15	0.498	0.026
Submarket 5	10	0.381	0.050

Figure 5 - Submarkets based on Ward's linkage clusters: Functional surface – Living area elasticity of house price

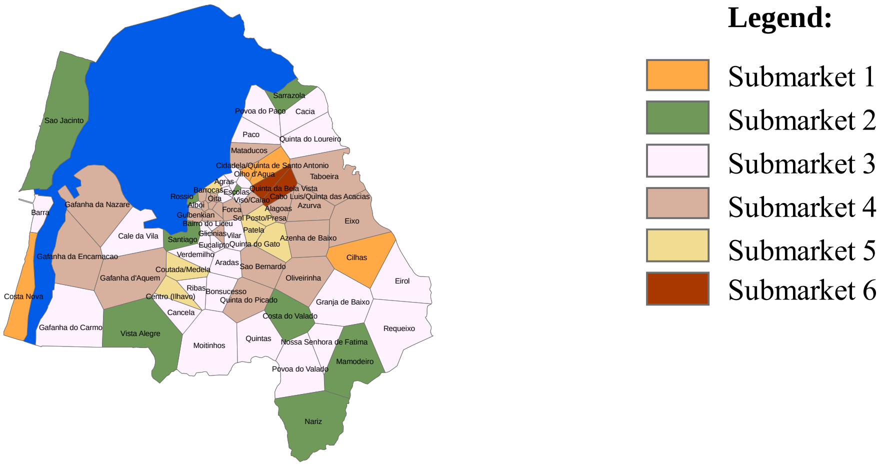
Figure 4 represents the territorial distribution of housing living area, and marks a clear distinction between the smaller available space in inner urban areas and the increasing availability of space as we move towards the periphery. The beach areas, secondary urban centres and main roads distort somewhat this regular concentric pattern. Inside the inner city there is a distinction between areas with old traditional buildings and social houses (with the lowest living space) and more modern and affluent residential areas; a similar contrast between Barra and Costa Nova beaches is also clear. The smooth spatial variation in average living area indicates that functional principal components may be useful in this application. We construct X^{**} and conduct spectral decomposition.

Next, we construct our dependent variable controlling for additional regressors and spatial fixed effects. We conduct fixed effects regression for the logarithm of price per square meter (y) on the 5 factors, plus time-on-the-market, allowing for zone-level fixed effects. The regressor slopes are assumed fixed, not spatially varying. The residuals constitute our modified dependent variable, y^*, for functional regression.

Using exogenous distance-based spatial weights using a bivariate Gaussian kernel, and the spectral decomposition of the covariance function of X^{**}, we obtain our functional regression estimates, first of $\hat{b}(I_k)$, and then $\hat{\beta}(I_k), k = 1,\ldots,76$. From these estimates, we infer the estimated spatially varying living area elasticity of price. Note that, the response variable here is logarithm of price per unit living area, and not the logarithm of price in itself. Hence the estimated elasticity for zone k is given by $1+\hat{\beta}(I_k)$. Finally, we conduct cluster analysis on these shadow prices, and report the spatial pattern in Figure 5.

In Figures 4-7, the zone-boundaries are demarcated by Voronoi tessellations (Okabe *et al.*, 2000), as convex polygons from the intersection of half-spaces between centres of neighbouring zones; likewise for Figures 6 and 7. From Figure 5, a concentric pattern of the shadow prices is evident. The elasticities (shadow prices of living space) are highest at the city centre and decrease as we move towards the periphery, meaning that the premium for a larger house is higher in the more urban areas.

The above regular concentric pattern is punctuated by four exceptions: i) areas corresponding to urban expansion along the main axial roads; ii) the urban centre of Ílhavo; iii) the urban centre of Gafanha; and iv) the Barra seaside resort, where the predominant new flats have a relatively strong premium for a larger apartment. Conversely, Costa Nova, a resort to the south of Barra, has mainly traditional small houses with rigid dimensions, attracting a very low premium for extra size (people are interested in a nice location and style of houses, and not so much in a larger living space).



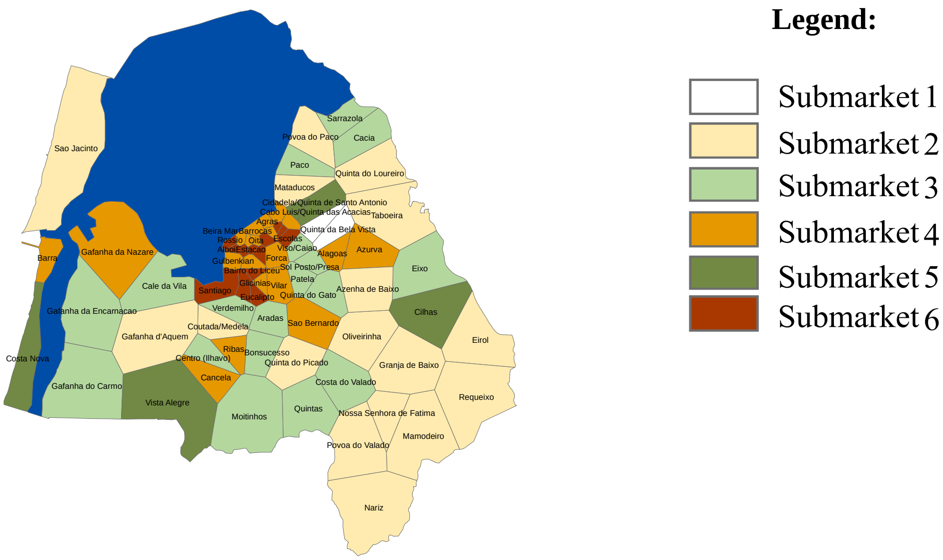
	Number of zones	Mean	Std. Deviation
Submarket 1	4	-2.346	0.228
Submarket 2	9	-0.957	0.220
Submarket 3	32	-0.231	0.209
Submarket 4	23	0.544	0.183
Submarket 5	6	1.256	0.089
Submarket 6	2	2.683	0.000

Figure 6 - Submarkets based on Ward's linkage clusters: Housing characteristics compared with marginal returns to living space

Figures 4 and 5 confirm substantial spatial heterogeneity, both in terms of living area and corresponding implicit (shadow) prices. The pattern of overlap between the two mappings is also interesting. To investigate this issue, we compare the two figures by standardizing $\bar{x}(s)$ and $\hat{\beta}(s)$ and mapping their difference (Figure 6).

The dominant pattern is one of inverse relationship between $\bar{x}(s)$ and $\hat{\beta}(s)$; the lower the average living space, the higher the shadow price. The signs of the standardized results were selected in line with the above idea. Therefore, negative values in Figure 6 imply that the returns to living space are lower than what would be expected based on the available space, while positive values represent the opposite. There is a pervasive dominance of low positive or negative differences, with few exceptions such as Costa Nova and other small areas. As a consequence, and following the Theorem 1 in section 3.3, we can argue that the submarkets presented in Figure 6 are robust to the two delineation principles – similarity in hedonic characteristics and similarity in hedonic prices.[14]

Thus, there is a clear case of partial overlap between submarkets delineated either by marginal utility of living area or by living space; the visual comparison of Figures 4 and 5 gives a similar impression. Therefore, we conclude that in line with Theorem 1, the combination of both criteria gives the best delineation of submarkets, in terms of two similar houses inside each submarket being good substitutes.



	Number of zones	FDA elasticity (standardized values)	Ln Area m^2 (standardized values)
Submarket 1	2	-0.156	2.839
Submarket 2	17	-1.050	1.124
Submarket 3	18	-0.268	0.402
Submarket 4	20	0.782	-0.524
Submarket 5	5	-1.137	-0.948
Submarket 6	14	0.931	-1.201

Figure 7 - Submarkets based on Ward's linkage clusters: Housing characteristics combined with marginal returns to living space

Finally, we report results of cluster analysis (Everitt, 1993) jointly on the two dimensions: $\bar{x}(s)$ and $\hat{\beta}(s)$. The corresponding delineation into submarkets is shown in Figure 7. As before, a strong spatial context persists. The submarkets, ordered by decreasing value of average living space, still show a concentric pattern, with some interesting features. The urban core of Aveiro corresponds to the submarket 6, with the smallest living area and the highest premium for additional space; submarket 4 corresponds to the outer ring of Aveiro, with extensions along the main roads, but also to some inner city areas (Gulbenkian and Bairro do Liceu) with relatively large high quality houses; submarket 5 corresponds to the previously discussed case of Costa Nova and three other areas with limited residential use, where the reduced living space is coupled with very low marginal returns to space; the remaining submarkets reflect the expected pattern of peripheral areas.

[14] This is also confirmed by simple regressions used only as a descriptive tool. Removing locations with less than 10 properties each, the R^2 between hedonic prices and characteristics is 0.31; retaining these low density zones, R^2 is 0.75.

In summary, the submarkets obtained from the above analysis, based on clustering jointly along two dimensions (Everitt, 1993), $\bar{x}(s)$ and $\hat{\beta}(s)$, produces submarkets that have a clear spatial context and approximately concentric pattern around the CBD of Aveiro. However, this concentric pattern is punctuated by processes of urban development – beach areas, secondary urban centres and main axial roads – that in turn reflect historical processes of development of the urban area.

The above delineation is based on a new functional regression framework and methodology accounting both for spatial dependence and spatial heterogeneity. In our empirical analysis, the spatial weights matrix is defined exogenously by a geographical-distance based independent bivariate Gaussian kernel. However, one can equally use estimated spatial weights, where a natural choice may be the estimator proposed in Bhattacharjee *et al.* (2012). However, in this case, the spatial weights and functional regressor will be endogenous. The IV estimator proposed here provides very similar submarket delineation, and the results are not reported separately.

7. Conclusion

The main topic of this paper was the definition of housing submarkets, both in terms of its conceptualization and empirical delineation. A new framework and methods based on functional data analysis were developed, integrating ideas and approaches from functional data analysis, spatial econometrics and locally weighted regressions. This allows for spatial dependence and spatial heterogeneity, and can also accommodate applications where the spatial structure is potentially endogenous. In allowing for endogenously determined submarkets and endogenous spatial regression, our work addresses two important limitations of existing methods. Both these aspects are key features of spatial dynamics in housing markets, as highlighted by Lefebvre (1974 [1991]), and an important point of our departure from the literature.

In the literature, analysis of housing segmentation has been conducted in several ways: i) by the similarity of hedonic housing characteristics, ii) by the similarity of hedonic prices, or iii) by the degree of substitutability of housing units. We apply our methodology to delineate submarkets in the Aveiro-Ílhavo urban housing market in Portugal. The results show that housing characteristic and prices produce submarkets that partially overlap, suggesting that spatial clustering based on a combination of the former two criteria provides a more reasonable approach towards defining submarkets, and one that also satisfies the condition that houses within the same submarket are highly substitutable.

The proposed synthesis and corresponding methods extend the literature along several directions. First, and most importantly, the framework can allow spatial structure and submarkets to evolve endogenously. This is in line with economic intuition, as well as empirical evidence. Second, our framework extends FDA tools and methods to the spatial domain in a way that is consistent with structural spatial econometric models of the housing market, and specifically the spatial lag model with spatial heterogeneity in slopes and spatial fixed effects. Third, once such submarkets have been delineated, spatial dependence can be examined by estimating cross- and within-submarket spatial weights (Bhattacharjee *et al.*, 2012).

Several further research problems and areas of development emerge from our work. First, while the framework enables analyses of endogenously produced submarkets, finding the asymptotic convergence rates for instrumental variables estimator within the functional regression model is retained for future work. Inference robust to weak instruments in this setting may also be useful. Further, relaxing the fixed design assumption would enhance applicability of the proposed methods. Combining the proposed approach with estimated spatial weights is a topic for further research. Second, in the above setting one can conduct inference on spatial structure, that is on an unknown spatial weights matrix *W*, exploiting the fact that the error term of the reduced form model has the structure $(I - W)^{-1}\,\varepsilon$, so that the spatial autocovariance matrix of these errors is a 1–1 function of *W* under the assumption of symmetric spatial weights (Bhattacharjee and Jensen-Butler, 2013). In particular, the submarkets identified in the previous step can also be used to estimate within and between submarket spatial weights (Bhattacharjee *et al.*, 2012; Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013). These potential methodological developments are exciting.

References

Adair, A., Berry, J. and McGreal, W. (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* **13**(1), 67-83.

Anas, A., Arnott, R. and Small, K. (1998). Urban spatial structure. *Journal of Economic Literature* **36**(3), 1426-1464.

Anselin, L. (1988a). *Spatial Econometrics: Methods and Models*. Kluwer: Boston.

Anselin, L. (1988b). Lagrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis* **20**(1), 1-17.

Anselin, L. (2002). Under the hood: issues in the specification and interpretation of spatial regression models. *Agricultural Economics* **27**, 247-267.

Anselin, L. and Griffith, D. (1988). Do spatial effects really matter in regression analysis?. *Papers of the Regional Science Association* **65**, 11-34.

Anselin, L. and Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics* **34**, 5-34.

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.

Arbia, G. (2014). "Dirty" spatial econometrics. Paper presented at the *VIII World Conference of The Spatial Econometrics Association*, Zurich, Switzerland, June 2014.

Bailey, N., Holly, S. and Pesaran, M.H. (2014). A two stage approach to spatiotemporal analysis with strong and weak cross-sectional dependence. CESifo Working Paper No. **4592**, Center for Economic Studies & Ifo Institute.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**, 192-236.

Bhattacharjee, A., Castro, E.A. and Marques, J.L. (2012). Spatial interactions in hedonic pricing models: the urban housing market of Aveiro, Portugal. *Spatial Economic Analysis*, **7**(1), 133-167.

Bhattacharjee, A. and Holly, S. (2013). Understanding interactions in social networks and committees. *Spatial Economic Analysis* **8**(1), 23-53.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics* **43**, 617-634.

Bhattacharjee, A., Maiti, T. and Petrie, D.J. (2014). General equilibrium effects of spatial structure: health outcomes and health behaviours in Scotland. *Regional Science and Urban Economics* **49**, 286-297.

Booth, J.G., Casella, G. and Hobert, J.P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B* **70**, 119-140.

Bourassa, S.C., Hamelink, F., Hoesli, M. and MacGregor, B.D. (1999). Defining housing submarkets. *Journal of Housing Economics* **8**, 160-183.

Bourassa, S.C., Hoesli, M. and Peng, V.C. (2003). Do housing submarkets really matter?. *Journal of Housing Economics* **12**(1), 12-28.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**(5), 2159-2179.

Cassetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical Analysis* **4**, 81-91.

Census of Portugal (2011). Recenseamento da população e habitação, Instituto nacional de estatística (http://www.ine.pt/).

Dale-Johnson, D. (1982). An alternative approach to housing market segmentation using hedonic price data. *Journal of Urban Economics* **11**(3), 311-332.

Everitt, B. S. (1993). *Cluster Analysis*. 3rd edition. Arnold: London.

Feng, W, Lim, C., Maiti, T. and Zhang, Z. (2012) Simultaneous estimation of disease risks and spatial clustering: a hierarchical Bayes approach. Technical Report **RM697**, Department of Statistics and Probability, Michigan State University.

Follain, J.R. and Malpezzi, S. (1980). *Dissecting Housing Value and Rent.* Urban Institute: Washington DC.

Fotheringham, A., 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.

Fujita, M. and Thisse, J. (2002). *Economic of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press.

Galster, G. (2001). On the nature of neighbourhood. *Urban Studies* **38**(12): 2111-2124.

Gillen, K., Thibodeau, T. and Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics* **23**(1), 5-30.

Goodman, A. C. and Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics* **35**(2), 209-232.

Grigsby, W.G. (1963). *Housing Markets and Public Policy*. University of Pennsylvania Press: Philadelphia.

Grigsby, W., Baratz, M., Galster, G., and Maclennan, D. (1987). The dynamics of neighborhood change and decline. *Progress in Planning* **28**(1), 1–76.

Guillas, S. and Lai, M.J. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics* **22**, 477-497.

Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35**(1), 70-91.

Hall, P. and Maiti, T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Annals of Statistics* **34**, 1733-1750.

Hall, P. and Maiti, T. (2012). Choosing Trajectory and Data Type when Classifying Functional Data. *Biometrika* **99**, 799-811.

Harris, R., Moffat, J. and Kravtsova, V. (2011). In search of 'W'. *Spatial Economic Analysis* **6**, 249-270.

Kelejian, H.H. and Piras, G. (2014). Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics* **46**, 140-149.

Knorr-Held, L. and G. Raßer (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**(1), 13-21.

Lancaster, K.J. (1966). A new approach to consumer theory. *Journal of Political Economy* **74**(2), 132-157.

Lefebvre, H. (1974 [1991]). *The Production of Space*. (Trans.) Nicholson-Smith, D., Blackwell: Oxford.

LeSage, J.P. (2004). A family of geographically weighted regression models. In: L. Anselin, R.J.G.M. Florax and S.J. Rey (Eds.), *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Springer: Berlin, 241-264.

LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*. Chapman & Hall/CRC.

Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Annals of Statistics* **38**(5), 3028-3062.

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

Maclennan, D. and Tu, Y. (1996); Economic perspectives on the structure of local housing systems. *Housing Studies* **11**(3), 387-406.

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.

Mardia, K.S. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis* **24**, 265-284.

Marques, J.L., Castro, E.A. and Bhattacharjee, A. (2012). Methods and models for analysis of the urban housing market. In: Capello, R. and Dentinho, T.P. (Eds.), *Networks, Space and Competitiveness: Evolving Challenges for Sustainable Growth*, Edward Elgar: Cheltenham UK, Chapter 7, 149-180.

McMillen, D.P. (1996). One hundred and fifty years of land values in Chicago: a nonparametric approach. *Journal of Urban Economics* **40**, 100-124.

McMillen, D.P. (2003). Spatial autocorrelation or model mis-specification?. *International Regional Science Review* **26**(2), 208-217.

Miller, D. (1978). The factor of scale: ecosystem, landscape mosaic and region. In: Hammond, K.A., Macinko, G. and Fairchild, W.B. (Eds.), *Sourcebook on the Environment: A Guide to the Literature*, University of Chicago Press: Chicago, 63-88.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. 2nd edition. John Wiley.

Pesaran, M.H. (2006). Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica* **74**, 967-1012.

Pesaran, M.H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics* **161**, 182-202.

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

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

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

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

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.

Sain, S. and Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics* **140**, 226-259.

Smith, L.B., Rosen, K.T. and Fallis, G. (1988). Recent developments in economic models of housing markets. *Journal of Economic Literature* **26**(1), 29-64.

Watkins, C. (2001). The definition and identification of housing submarkets. *Environment and Planning A* **33**(12), 2235-2253.

Williamson, O. (2000). The new institutional economics: taking stock, looking ahead. *Journal of Economic Literature* **38**(3), 595-613.