Endogenous spatial structure and delineation of submarkets: 
A new framework with application to housing markets

Arnab Bhattacharjee, Eduardo Castro, 
Taps Maiti and João Marques
Endogenous spatial structure 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
cecastro@ua.pt

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

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

KEYWORDS: Spatial heterogeneity; Submarkets; Spatial lag model; Geographically weighted regression; Functional data analysis.

JEL CLASSIFICATION: C21; R31; C51.

Abstract
Definition of housing submarkets is important at both conceptual and empirical levels. In the housing studies literature, submarkets have been defined according to three different criteria: i) similarity in hedonic housing characters, ii) similarity in hedonic prices; iii) substitutability of housing units. We argue that the simultaneous fulfilment of criteria i) and ii) is a sufficient condition for criteria iii) to be fulfilled. Criterion i) is directly observable, while criterion ii) can be checked by a model able to detect and analyse spatial heterogeneity in the shadow prices. Here, we propose a new framework, based on a synthesis of spatial econometrics, functional data analysis (FDA) and geographically weighted regression (GWR). The framework is applied to a hedonic regression model where the dependent variable is logarithm of house prices per square meter and housing features are regressors. Thus, we delineate submarkets by clustering (jointly) on the surfaces of the estimated functional partial effects and housing features. The above model addresses two main limitations of previous approaches. First, endogeneity in spatial structure can be incorporated in the model. Second, the framework does not require delineation of housing submarkets a priori. 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.

* Corresponding author: Arnab Bhattacharjee, Heriot-Watt University, Spatial Economics & Econometrics Centre (SEECE), Room 1.06, Mary Burton Building, Edinburgh EH14 4AS, Scotland, United Kingdom. E-mail: a.bhattacharjee@hw.ac.uk

# The paper has benefited from comments by participants at the VI World Conference of the Spatial Econometrics Association (Salvador, Brazil, July 2012) and the Workshop on Housing Economics organised by the European Network for Housing Research (Vienna, March 2012), and seminars at London School of Economics and University of Illinois at Urbana-Champaign. We are grateful to Anil Bera, Edwin Deutsch, Duncan Maclennan, Daniel McMillen, Peter Robinson and Jacques-Francois Thisse for many helpful comments and suggestions. The usual disclaimer applies.
"(S)pace is a social product ... the space thus produced also serves as a tool of thought and of action ... 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]

1. Introduction

Definition of housing submarkets is important at both conceptual and empirical levels. The correct understanding of housing segmentation enables researchers to better understand the 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 et al., 2007).

By its nature housing is a heterogeneous good, characterized by a diverse set of attributes (Lancaster, 1966 and Rosen, 1976) and segmented and structured by complex spatial patterns, where 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 (Linneman, 1981; Mills and Simenauer, 1996), through metropolitan areas (Follain and Malpezzi, 1980), to below the metropolitan level (Straszheim, 1975; Gabriel, 1984; Grigsby et al., 1985; Rothenberg et al., 1991; Maclellan and Tu, 1996; and 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.

In conceptual terms, the key issue is to understand why a housing market is divided into several submarkets, corresponding, in theory, to an equal number of market equilibria. In turn, a market equilibrium in an heterogeneous good such as a house, rather than being defined by a single combination of a quantity and a price, is given by the combination of a vector of hedonic characteristics with a vector of hedonic prices which, under an appropriate functional specification, produce an overall price for the good (Lancaster, 1966). Thus, the existence of a unique vector of hedonic prices, combined with a homogeneous distribution in space of houses with different hedonic characteristics, is a necessary condition for the existence of a single equilibrium and a unique market. However, there are several reasons why such a unique combination is in general not observed. 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 but, once they are built, any changes 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, 1998, 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.
At the empirical level, the correct treatment of spatial heterogeneity increases the prediction accuracy of the estimated hedonic models and, in many cases, could negate spatial strong dependence (Pesaran, 2006; Pesaran and Tosetti, 2011).\footnote{Even though spatial heterogeneity and spatial dependence are theoretically distinct problems (Can and Megbolugbe, 1997), adequate treatment of common factors with heterogeneous slopes is necessary for inference on structural spatial dependence (Bhattacharjee and Holly, 2013); see also McMillen (2003).} A large diversity of approaches can be found in the literature for defining housing markets. In general, the delineation of submarkets can be done based on \textit{a priori} judgments such as pre-existing geographic or administrative boundaries or subjective knowledge and expertise (Straszheim, 1975; Goodman, 1981), or using the structure of data to apply more sophisticated analytical methods such as hierarchical models (Goodman and Thibodeau, 1998, 2003) and non-parametric spatial statistical models (Clapp et al. 2006; Bhattacharjee et al., 2013).

Housing economics literature has defined submarkets either by similarity in hedonic housing characters (Rothenberg et al., 1991), by similarity in hedonic prices (Bourassa et al., 2003), or by substitutability of housing units (Pryce, 2013). In the first approach, a submarket is characterised as a collection of locations, or housing units located therein, that have a similar bundle quality or, in other words, supply a similar set of hedonic characteristics. While how much similarity is required to define a submarket is a debateable issue, it is important to stress that a perfectly homogeneous location is necessarily very small and is not useful for hedonic models, which need a minimum level of variety in order to enable reliable estimation of hedonic prices (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 technique 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 prices are homogeneous; submarkets can then be interpreted as clusters of houses with characteristics which are adjusted to a particular demand behaviour, which is reflected in a particular 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 deeply related to the basic philosophy of hedonic models, 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 a technique based on functional data analysis.

The third criterion for defining submarkets is based on substitutability. Under this criterion, submarkets are defined by similarity in house prices, with variation in hedonic characters already capitalised in these prices. Pryce (2013) shows that substitutability can be measured by the cross-price elasticities of price at different locations estimated using data on spatial panels. A key assumption underlying the methodology in Pryce (2013) 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 solely spatial weak
dependence, in the sense of Pesaran and Tosetti (2011). 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 can then expect that inclusion of these factors would account for any strong spatial dependence, which would then render the spatial model as containing only spatial weak dependence. Estimation of substitutability using such cross-section data is discussed below in this paper.

Under what conditions are these approaches equivalent? Though a comprehensive discussion of the question is beyond the scope of this paper, 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 services and amenities and similar accessibilities 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).

The above sufficient condition shows that when the delineation of submarkets using hedonic characteristics and hedonic prices criteria overlap, the outcome also corresponds to submarkets where the substitutability criterion holds. However, when we extend the discussion to larger areas with some internal heterogeneity, the problem becomes more complex. Such heterogeneity makes very unlikely the perfect overlapping of sub-markets delineated by hedonic prices or hedonic characteristics. It is more reasonable to expect partial overlapping, implying that each criterion produces a specific set of sub-markets which, although related to each other, represent distinct dimensions of reality. Therefore, the delineation of sub-markets by combining physical characteristics and hedonic prices seems to be a good procedure and we can assume that the trade-off between physical and price homogeneity still reflects substitutability. In other words, two identical houses in terms of prices and characteristics located inside the same sub-market defined as above are good substitutes.

In this paper, 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. Initially, our focus lies 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 (0) 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.

\[ \beta(s) \]

\[ -1 < \beta(s) < 0; \] when the elasticity approaches zero \( (0) \) 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.

---

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 (the strong factor contributing to spatial strong dependence) is sufficient to ensure spatial weak dependence.
We study how estimates of this regression model can be used to identify submarkets, following the criterion of similar hedonic prices 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 such submarkets have been delineated, spatial dependence can be examined by estimating cross- and within-submarket spatial weights.

Towards this end, we propose a new framework, based on a synthesis of spatial econometrics, functional data analysis (FDA) and geographically weighted regression (GWR) for the analyses of housing markets. 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 (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, endogeneity in spatial structure (spatial weights) can be incorporated in the model. Second, 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 surface. 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

Marques et al. (2012) discuss several spatial problems of current interest, covering supply and demand sides, price formation and policy, that are fundamental for understanding the housing market; see also Smith et al. (1988) for an excellent but somewhat dated review. 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 prices reflect not only geographically varying price effects, but also substantial clustering; Rosenthal (1999) and Malpezzi (2003) address this issue, which was discussed in the previous section as being 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). In this paper, we discuss these issues in the context of a spatial hedonic pricing model.
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 access to central and local services and other housing attributes, 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 Chattopadhya (1999), Malpezzi (2003) and Palmquist (2005) for recent reviews.

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:

\[ Y = f(S, N, L, C, T), \]

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 (date, month) 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. Following Bhattacharjee et al. (2012), we consider the hedonic model with a small modification to the 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.

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, Pace and LeSage (2004), Anselin and Lozano-Gracia (2008) and Anselin et al. (2010). Spatial patterns in housing markets arise from a combination of spatial heterogeneity and spatial dependence (Anselin, 1988a). Additionally, as discussed before, choice of an appropriate spatial scale is important (Malpezzi, 2003). 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 (Malpezzi, 2003). 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.

However, the empirical literature does not suggest an unambiguous definition of a unique spatial scale. Analysis of spatial hedonic models for the city of Aveiro (Portugal) in Bhattacharjee et al. (2012) reflects some advantages of using a flexible spatial scale, since processes of agglomeration and dispersion operate differently at different scales; see also Arbia et al. (2010a,b).

2.2.2. Spatial heterogeneity and neighbourhood effects

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 supply factors, institutional barriers or discrimination, each of which can cause differentials across neighbourhoods in the way housing attributes are valued by consumers and house prices determined (Anselin et al., 2010). However, if spatial submarkets are present and ignored, an average price across all the sample is estimated that ignores submarket heterogeneity. As discussed above, heterogeneity is a key element of housing markets and its disregard seriously affects the understanding of how market diversification is a central element of the housing sector. Worse still, estimated average prices are most certainly wrong because the error term of the regression can then be correlated with the included regressors and ordinary least squares (OLS) will produce biased estimates.

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), Mozolin (1994), Adair et al. (2000) and Soderberg and Janssen (2001), among others, have examined intra-urban variation in the price of housing 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., 2002); we discuss this method 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 other forms of contagion effects. However, incorrectly modelled spatial heterogeneity, measurement problems in explanatory variables, omitted variables, and unmodelled features having a spatial pattern can also lead to spatial autocorrelation (Anselin and Griffith, 1988). 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, and the use of

---

3 See, for example, Pace and LeSage (2004) and Anselin and Lozano-Gracia (2009).
spatial econometric models to address spatial autocorrelation is becoming increasingly standard.\(^4\)

The usual approach to the representation of spatial interactions is to define a spatial weights matrix, denoted \(W\), which typically represents a theoretical and \textit{a priori} characterisation of the nature and strength of spatial interactions between different submarkets or dwellings.\(^5\) 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). The spatial weights are typically modelled 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.

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 \(Wy\):

\[
\hat{y} = \rho Wy + X\beta + \varepsilon,
\]

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

\[
\hat{y} = X\beta + \varepsilon, \quad \varepsilon = \lambda Wy + \eta.
\]

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. However, though different in interpretation, the above two models are very difficult to distinguish empirically (Anselin, 1999, 2002). A popular approach in the area of spatial econometrics 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 Born and Breitung (2011).

The choice of appropriate spatial weights is a central component of spatial models as it imposes \textit{a priori} a structure of spatial dependence, which may or may not closely correspond to reality. Further, the accuracy of these measures affects severely the estimation of spatial dependence models (Anselin, 2002; Fingleton, 2003). Spatial contiguity or suitable functions of geographic distances are frequent choices. However, spatial data may be anisotropic, where spatial autocorrelation is a function of both distance and the direction separating points in space (Simon, 1997; Gillen et al., 2001). Further, spatial interactions may be driven by other factors, such as trade

---

\(^4\) For representative applications using hedonic models in a spatial econometric setting, see Can (1992), Pace and Gilley (1997), Basu and Thibodeau (1998) and Anselin et al. (2010).

\(^5\) For a setting with \(n\) spatial units under study, \(W\) 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.
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.

2.3. Interconnection between the three aspects

More recently, attention has begun to get focused on the interconnection between different spatial aspects. Bhattacharjee et al. (2012) develop a framework and applications that pay special attention to the three related but distinct aspects of spatial analyses relating to housing markets – 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 well-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 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 substitutability 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 framework in Bhattacharjee et al. (2012) emphasizes the connection between urban spaces and housing markets and places focus on understanding urban housing markets in terms of three distinct but interconnected features of space – spatial heterogeneity, spatial dependence and spatial scale. 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 $W$) is unknown and potentially endogenous.
3. Delineation of Submarkets

Defining and delineating housing 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 (Palm, 1978). One problem is the definition of submarket. Theoretically, a submarket correspond to a local equilibrium between supply and demand; however, the way to translate this concept into measurable variables and meaningful models leads to difficult questions about the levels of aggregation (or disaggregation) and about the methods which can be used to cluster (or to divide) basic spatial units in order to define submarkets. In practice, these questions are often answered in an ad hoc manner, using predefined or otherwise convenient geographical boundaries as the basis for defining submarkets. In some cases, statistical tests are applied to determine whether the a priori submarkets are, in fact, 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, 2007). The stated 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 equilibriums in submarkets. While the supply side is endogenously related to the demand side through the market equilibrium, and prices are therefore also determined by both supply and demand through the equilibrating process, there is no direct way of understanding 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.

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 (2010) 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 (2010) uses house price inflation for computation of the elasticities, and illustrates the methodology for delineating submarkets using data on Glasgow. 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 (2010) is a spatial error model:

\[
y_t = y_{t-1} + \varepsilon_t + W\varepsilon_t + \eta_t,
\]

where \(y_t\) denotes the vector of prices (in logarithms) across all houses in the sample, and \(y_{t-1}\) its lagged value, so that \(\varepsilon_t\) denotes the growth rate in prices. Then, the elements of \(W\) are the cross price elasticities of price for each pair of houses. The above model of house prices does not include any explanatory variables, except for lagged logarithm of prices with unit coefficient, which in turn ensures that the regression errors (growth rates) are stationary in the temporal domain. Nevertheless, despite the lack of explanatory variables, the model itself is structural because it assumes that the process of diffusion of shocks (idiosyncratic errors, \(\eta\)) is driven by an underlying spatial structure represented in \(W\). The objective in Pryce (2010) was to infer from the data the underlying spatial structure, which 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 clearly constitute a 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 unity for houses within the same submarket. By such delineation of submarkets, one can of course 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 (5) cannot be cast into the framework of spatial econometries.

### 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.

---

\(^6\) Taking house prices, rather than hedonic characteristics, as the sole basis for evaluation 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.
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, perhaps a more appropriate model, and one with stronger structural interpretations is the spatial lag model (2). In a setting where $W$ is unknown (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013), $W$ and $\rho$ are not separately identifiable. Hence, we assume without loss of generality that $\rho = 1$. Thus, we make the following assumption.

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

$$y = \rho Wy + X\beta + \varepsilon \Rightarrow y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon. \quad (6)$$

with full spatial heterogeneity in both the slope and intercept (heterogeneity in $\beta$ across the territory, plus location fixed effects). $W$ is unknown and potentially endogenous, and satisfies the spatial granularity condition $\rho(W) < 1$, where $\rho(W) = \max \|W\|_2, \|W\|_w$ is the norm of $W$. $\|W\|_w = \max \sum_{I \leq j \leq n} |w_{ij}|$ the column norm of $W$, and $\|W\|_r = \max \sum_{j = 1}^n |w_{ij}|$ the row norm of $W$.

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 (6).

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

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

$$\Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (I - W)^{-1} \begin{pmatrix} x_1 \beta_1 \\ x_2 \beta_2 \end{pmatrix} + (I - W)^{-1} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \quad (7)$$

$$\approx \begin{pmatrix} I + W \end{pmatrix}^{-1} \begin{pmatrix} x_1 \beta_1 \\ x_2 \beta_2 \end{pmatrix} + \begin{pmatrix} I + W \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

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

**Theorem 1:** Under Assumption 1, similarity in hedonic characteristics and prices imply house’s substitutability (cross price elasticities of price close to unity).

**Proof:** Since elements of $W$ are small, these can be ignored in calculating cross price elasticities up to first order Taylor expansion. Also, the idiosyncratic errors can be ignored. Then, for any two similar housing properties $i$ and $j$:

$$\frac{\partial \bar{\varepsilon}_i}{\partial y_j} \approx \frac{x_i}{x_j} \cdot \frac{\partial \beta_i}{\partial \beta_j} \approx 1. \quad (8)$$

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, the result shows that typically 2 housing properties will be substitutable \( (\hat{\partial} y_1 / \hat{\partial} y_2 = 1) \) if two conditions simultaneously hold: (a) the shadow prices at the two locations are the same \( (\beta_1 = \beta_2) \) and (b) the value of the hedonic characteristic is also the same at these two locations \( (x_1 = x_2) \). In other words, submarkets, whatever the adopted definition, should be delineated by spatial clustering on both these two aspects 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 appropriate if there are no regressors. In this case, the cross price elasticities will be solely determined by elements of \( W \). However, the model will then not offer any useful structural interpretation, because of spatial nonstationarity.

The question that follows is, how can one implement such a procedure for delineating submarkets, in a way that is computationally feasible? For this, we find a synthesis of several empirical approaches more useful than a purely spatial econometric framework. We turn next to a discussion of some related methods and the proposed synthesis.

### 4. A Synthesis of Empirical Approaches

The framework and methods in Bhattacharjee et al. (2012) highlights the interconnections between spatial heterogeneity, spatial dependence and spatial scale. This leads to a unique understanding of spatial aspects of the housing market, in terms of neighbourhood choice, housing preferences, and the evolution of urban spatial structure. Therefore, this has 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 spatial weights have been assumed. When these spatial interactions are unknown and themselves objects of inference, endogeneity issues become quite complex. 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 geographically weighted regressions (GWR), functional data analysis (FDA) and spatial statistics.

4.2.1. Geographically (or locally) weighted regression

In the recent literature, spatial heterogeneity has been modelled using locally weighted regressions (McMillen, 1996), of which the Geographically Weighted Regression (GWR) approach (Fotheringham et al., 2002) is perhaps the most popular. In particular, GWR replaces the single regression coefficient in a standard linear model with a series of (geographically weighted) estimates for a number of regression points estimated from the underlying data. The end result is 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. It allows analysis to move beyond a single global partial effect and consider local and contingent explanations. GWR is a kernel based nonparametric regression method where

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

(9)

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 space \( S \), \( f_{h,i}(s) \) is a kernel density with bandwidth \( h \) and centred on location \( i \), the regression slope \( \beta_i \) varies over space, and \( \alpha_i \) can be interpreted as a location specific fixed effect. In effect, this method provides pointwise estimates \( \beta_i \) 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, or 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 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 processes generating the data. As a consequence, curve estimation methods designed to yield good derivative estimates often play a critical role in functional data analysis.

In the typical application where the domain is time, projection to a basis space is based on a Fourier basis for periodic data (Hall and Hosseini-Nasab, 2006), or smoothing splines for functional data that are not periodic (Crambes, Kneip and Sarda, 2009). The functional regression model can then be estimated by functional principal components (Cai and Hall, 2006; Hall and Horowitz, 2007). In our spatial context, such a functional regression model would take the form:

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

(10)
where the dependent variable ($y$) is scalar, and the regressor ($x$) and slope ($\beta$) are functional; alternatively, the dependent variable can also be specified as a functional variable in this framework.

The main challenge in the application of FDA to our context lies in the fact that the domain of our functional data is not time, but a two-dimensional space. Spatial data, unlike time series, has no well-defined ordering of observations, and neither a sense of progression. This leads to considerable challenges in construction of a suitable basis function. A Fourier basis does not appear to be suitable, and neither is there a well-defined extension of the spline basis to two dimensional space. Guillas and Lai (2010) have recently proposed a methodology for functional linear regression based on bivariate splines over triangulations which we intend to explore.

4.2.3. Spatial statistics

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. However, the specific area that we are interested in is multivariate spatial models, which has in recent years proven an effective tool for analyzing 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 such variables are recorded at each spatial location simultaneously. With rapid enhancements in geographic information systems (GIS) technology that enables us to analyze 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 (data with indices in a discrete or countable set), while specifically mentioning that the latter is not as 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 modeled 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 typically even more rigidly defined – typically as a row normalized version of an adjacency matrix. Bayesian inferences on this model are well developed; see, for example, Gelfand and Vounatsou (2003), Banerjee et al. (2004), Jin et al. (2007), and Dass et al. (2009, 2010). Further, each square grid can be treated as a small area, and then a suitable small area model can be developed. Such small area models are very useful for prediction; see Rao (2003). For flexible models, Hall and Maiti (2006a,b) can also be potentially useful.
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 potentially 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 (7) specific to two housing properties, but incorporating heterogeneity in slopes. As we showed before, in this case:

\[
\begin{bmatrix}
  y_1 \\
  y_2
\end{bmatrix}
\approx
\begin{bmatrix}
  x_1\beta_1 \\
  x_1\beta_2
\end{bmatrix}
+ \begin{bmatrix}
  I + W \\
  I + W
\end{bmatrix}
\begin{bmatrix}
  e_1 \\
  e_2
\end{bmatrix}.
\]

This implies a regression model where, in addition to \(x_1\beta_1\), the right hand side also includes \(x_2\beta_2\), but with a much smaller weight. 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 \([I - W]^{-1}\approx [I + W]\).

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.

Potentially, 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 be incorporated in fairly standard ways. One would need either a dynamic model for how these weights evolve over time, or use suitable instruments.

Second, the framework allows submarkets to evolve endogenously and abstracts from the requirement to delineate housing submarkets a priori. As discussed before, Rothenberg et al. (1991) support the view that a submarket is characterised as a collection of locations that have a similar bundle "quality" (that is, close hedonic substitutability), and these sets of close substitute locations (or housing units located therein) may or may not have any spatial content. Under this view, spatial clustering will be potentially an important methodology to delineate housing submarkets.

The main alternative to defining submarkets by similarity in hedonic characters is to base the definition on substitutability of housing units (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 necessarily in hedonic characters, but more importantly in the shadow prices assigned to such features. As our Theorem 1 shows, in the absence of such heterogeneity, housing submarkets should be delineated by clustering on both the surface of the functional partial effect \(\beta(s)\) 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

\[ E[y_i] = \alpha(s) + \int_S \beta(s) \gamma_i(s), \]

\[ x_i(s) = x(s) f_{h,i}(s), \]  

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

5.1. Estimating the Functional Regression Model

Potentially, the model (11) can be estimated either by functional principal components (FPC) (Hall and Hosseini-Nasab, 2006), or potentially also smoothing splines (Crambes, Kneip and Sarda, 2009). However, estimating the model using functional principal components presents some challenges, mainly because the functional surface is two dimensional space, rather than \( \mathbb{R}^+ \). Use of the B-spline or a bivariate spline as in Guillas and Lai (2010) are alternate options. In this paper, we work with a variant of functional principal components.

The main issue with functional principal components in the current setting is the following. 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 typically 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, so sparse that approximating such information requires a large number of principal components and also produces a poor approximation. Econometrically, this is a problem of regularisation.

Our data generating process is the following. 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 \( S \subset \mathbb{R}^2 \). For a specific location \( i \in S \), both \( Y \) and \( X \) are scalar random variables. However, the response variables \( Y_i \) are generated by a functional linear regression model

\[ Y_i = \alpha + \int_S \beta X^*_i + \epsilon_i, \quad i = 1,\ldots,n, \]

\[ X^*_i(u) = \begin{cases} X_i & \text{if } u = i \\ X_j f_{ij} & \text{if } u = j \in \{1,\ldots,n\}, i \neq j, f_{ij} = f_{h,i}(j) \\ 0 & \text{otherwise.} \]  

The choice of the kernel function is made such that the errors \( \epsilon_i \) are independent and identically distributed with finite variance and zero mean and that the errors are also independent of the explanatory variables.

The main problem here is that the functional regressor surface of \( X^*_i \) is very irregular. Even if the underlying regressor \( X \) has smooth variation over the set \( S \), the combination of Dirac delta at location \( i \) with a kernel function elsewhere renders \( X^*_i \) very irregular and spiky. Hence, usual regularisation by principal components as in Ramsay and Silverman (1997) 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 here.
Our approach focuses on directly regularising the surface of $X$ 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 weight (unity, 1) at the location of observation $i$, compared to much lower weights ($f_{ij}$) at other locations. Our object of inference here is the functional surface of the regression coefficient ($\beta$) which may be assumed to be relatively smooth. 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 usual functional data where the regressor is observed at a (large) number of fixed time points. However, since the number of such partitions would typically be large compared to sample size ($n$), a second stage of regularisation would be 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 $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_K)$, where $\tilde{x}_i = E(X_i | i \in P_k), k = 1, \ldots, K$. 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, is there a way similar to Hall and Horowitz (2007) to then go from these principal components to the estimation of the functional surface of the regression coefficient ($\beta$)?

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 some abuse of notation, denote the partition that observation $i$ belongs to also by the index $i$, that is $i \in P_i$. Then, the sieve functional regressor for observation $i$ is

$$X_i^* = \left[ n_{i1}f_{i1}\tilde{x}_1, n_{i2}f_{i2}\tilde{x}_2, \ldots, n_{ik}f_{ik}\tilde{x}_k, \left(1 - f_{ik}\right)X_i + n_{ik}f_{ik}\tilde{x}_k \right]_{i1, i2, \ldots, ik, k}.$$  \hfill (13)

Dividing the functional regressor (13) by the scalar exogenous weights $n_k f_{iks}$, we have

$$\frac{X_i^* - f_{ik}\tilde{x}_i}{n_k f_{ik}} \to 0 \text{ as } n \to \infty,$$

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, $\frac{X_i - f_{ik}\tilde{x}_i}{n_k f_{ik}} \to 0$ as $n \to \infty$, so that in large samples, this is really the average process. Thus it is expected to be smooth over space if the functional surface of $X$ is itself smooth.

With (b) challenges remain. However, note that, in large samples, when variation in $X_i$ does not matter for the construction of the functional regressor, $X^* \approx \bar{X}Z$, where $\bar{X} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K]$ and $Z$ is a vector that takes value 1 at location $i$ and 0 otherwise.

Now let us make the following assumptions.
Assumption 3:

(a) The data are generated by fixed spatial design, so that \( Z \) is nonstochastic.

(b) All other technical assumptions in Hall and Horowitz (2007) hold. Specifically, conditions on the distribution of \( \bar{X} \) (the spatial average process of \( X \)), distribution of \( \varepsilon \), eigenvalues and Fourier coefficients hold.

Under assumption 3(a), \( \text{Cov}(Y, \bar{X}Z) = \bar{Z}\text{Cov}(Y, \bar{X}) \). Then, following Hall and Horowitz (2007), the covariance can be estimated by functional principal components. Then, estimation of the functional surface \( \beta(s) \) by the functional principal components estimator in Hall and Horowitz (2007), \( \hat{\beta}(s) \), follows. Then, we have the following result.

**Theorem 2 (Hall and Horowitz, 2007):** Under Assumption 3, \( \hat{\beta}(s) \xrightarrow{p} \beta(s) \). The rate of convergence is generic to noisy inverse problems.

Theorem 2 follows from a straightforward application of Hall and Horowitz (2007). The only innovation here is adapting this result to a spiky functional regressor surface. This we achieve by making a fixed design assumption, together with imposition of a histogram sieve. Relaxing this potentially strong assumption appears to be difficult, and is retained for future work.

### 5.2. Implementation of the Functional Principal Components Estimator

Along the above lines, we partition the territory 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 across these \( k \) locations, combined into a spatial vector \( \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k\} \). We aim to conduct functional principal components on this vector of spatial averages. However, the above vector in itself does not have any cross section variation. This is because the cross section variation in \( x_i \) has been lost in the process of aggregation by averaging. To recover this information, we therefore replace \( \bar{x}_j \), for observation \( i \) belonging to partition \( I_j \), with

\[
\bar{x}_j^* = \bar{x}_j + x_i \frac{1-n_j f_{0i}}{n_j f_{0i}} \left[ f_i \left( 1 - f_{0i} \right) + n_j f_{0i} \bar{x}_j \right],
\]

where \( f_{0i} = f_{0i}(I_i) \) is the modal kernel density centred on the location of \( i \), and \( n_j \) denotes the sample size in partition \( I_j \). Then, functional regression proceeds by obtaining a small number of functional principal components and regressing the 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: \( i \in I_j \).

2. Construct functional average surface as

\[
X_j^* = \left( x_1^*, x_2^*, \ldots, x_k^* \right)
= \left( \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_j + x_i \frac{1-n_j f_{0i}}{n_j f_{0i}} , \ldots, \bar{x}_k \right).
\]
3. Conduct functional principal components on $X_i^*$, estimate the $l$ principal component factors $\left\{ \hat{f}_1, \hat{f}_2, \ldots, \hat{f}_k \right\}$, and obtain predicted average aggregate regressor vector $\hat{X}_{i}^*$ based on these factors.

4. Construct the FPC functional regressor surface by multiplying each element of $\hat{X}_{i}^*$ by $n_d f_{h,d}(I_d)$, $d = 1, 2, \ldots, k$, to obtain:

$$\hat{X}_{i}^{**} = (n_1 f_{h,1}(I_1), n_2 f_{h,2}(I_2), \ldots, n_k f_{h,k}(I_k)).$$

5. Finally, regress $y$ on the above $k$ regressors, as well as any other exogenous explanatory variables (with regression coefficients fixed over space) in the spatial hedonic model, using all the data. The object of inference, $\beta(s)$, together with standard errors is recovered as the slope coefficients for elements of $\hat{X}_{i}^{**}$.

The estimation methodology can now be easily applied to the data.

5.2. Submarket Delineation by Functional Clustering

Once the functional regression slope surface is estimated, the next step of further analysis constitutes using the surface $\beta(s)$ and $x(s)$ to identify housing submarkets by functional clustering. The notion of spatial 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.

The full development of these methods lies within the domain of future research. In this paper, we identify submarkets 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); see Everitt (1993).

Several other lines of methodological development directly follow from the methodology developed here. Most importantly, in future research, we plan to conduct inference on spatial structure, that is on an unknown spatial weights matrix $W$. Specifically, the error term is of the reduced form spatial functional regression model (7) is of the nature $(I+W)e$. 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 $e$ 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 $W$.

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 drilling down on the functional spatial domain.
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. 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 extremely exciting.

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

In this section it is applied the methodology described in previous sections to analyse housing submarkets in an urban specific context. The analysed area is located in the Centro Region of Portugal and includes two municipalities: Aveiro and Ílhavo (see Figure 1).

![Figure 1 – Location of the study area: Municipalities of Aveiro and Ílhavo](image)

The municipality of Aveiro has a total area of 200 km$^2$ and a total population of 78454; the municipality of Ílhavo has an area of 75km$^2$ and 38317 inhabitants (Census, 2011). If the area of the lagoon is removed, the population density is 600 inhabitants per km$^2$, a typical value for an urban agglomeration. The following main zones can be defined (see Figure 2):

i) A semi-rural area with 30000 inhabitants, where a significant part of the territory is dedicated to agriculture but where almost all population work in the manufacturing and service sectors, spread all over the urban agglomeration; new urban settlements with blocks of flats and clusters of detached houses are mixed with old rural settlements, following the typical local pattern of strings of houses extended along the roads.

ii) A suburban area with 33500 inhabitants, which spreads around the city of Aveiro, with a settlement and employment pattern equivalent to that of the previous area but with a higher weight of new urban settlements.

iii) The inner city of Aveiro, with a population of 32000 inhabitants, which is the core of the urban area.

iv) The smaller city of Ílhavo, with a population of 5000 inhabitants, which is the second urban centre of the agglomeration.
v) Gafanha da Nazaré, with a population of 13,000 inhabitants, is where the port of Aveiro is located. This zone 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 predominate.

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 different types of 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. Data refers to the time interval between October 2000 and March 2010 and includes around 4 million records of properties available for transaction in Portugal, covering all the national territory. For the specific case of Aveiro and Ílhavo, between 2000 and 2010, the database was populated with 47,188 different properties. This empirical work used 12,467 observations, after removing all cases where data were incomplete or inconsistent.7

Besides price of each property, the database includes two main categories of variables to describe each dwelling (see Bhattacharjee et al., 2012): i) the intrinsic physical attributes, and ii) the location and neighbourhood of the building. The first group includes number of rooms, preservation, age of construction and area. A set of other physical housing characteristics, obtained from a free text field where real estate advertisers describe the property, was used. The second group of attributes is related

---

7 For a detailed description of the main challenges in cleaning the data and procedures used for the construction of new housing attributes (intrinsic and location and neighbourhood), see Marques (2012).
to the housing location and to the characteristics of the neighbourhood, obtained by potential and distances measures (see Marques 2012).

Because only a reduced portion of houses are geo-referenced, they were located in the smallest homogeneous areas that the database can describe, being the centres of such areas geo-referenced (see Figure 3 where the areas are represented).

![Figure 3 - Housing location by zones](image)

The data reflect large variation in housing characteristics (see Bhattacharjee et al. 2012). The average price (in euros per square meter) is 1126 and ranges from 178 up to 5714; the dwelling dimension is 149 m², varying between 20 m² and 600 m². Regarding other housing characteristics, the distribution is the following: 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 and 63.8% have a garage; 43.3% have central heating and 28.9% have a fireplace. Regarding the location attributes, on average, houses are located at 3.2 km from the CBD, being the maximum distance 16 km.

In order to capture the main dimensions of the housing characteristics, a factorial analysis with orthogonal varimax rotation was applied to the dataset. The explanatory variables considered in the analysis were organised in 5 factors: 3 are related 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).⁸

The following three maps represent the spatial housing segmentation for Aveiro and Ílhavo considering three different strategies. A spatial clustering analysis was applied to: i) the values of the functional regression coefficient $\beta(s)$ (resulting from the application of FDA, as described in the section 5) which, as explained above, represent the decreasing marginal utility of living area (Figure 4); ii) the values of the

---

⁸ See Bhattacharjee et al. (2012) and Marques (2012) for additional details on how these factors were applied and defined.
housing living area (measured in square meters) (Figure 5); and finally, iii) a combination of both (Figure 7).

From Figure 4, a concentric pattern of the housing submarkets is evident, where the decreasing returns of living space are lower in the city centre and grow higher as we move towards the periphery, meaning that the premium of a larger house increase in the more urban areas.

<table>
<thead>
<tr>
<th>Submarket</th>
<th>Number of zones</th>
<th>Mean (FDA_coef)</th>
<th>Std. Deviation (FDA_coef)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submarket 1</td>
<td>18</td>
<td>-0.155</td>
<td>0.056</td>
</tr>
<tr>
<td>Submarket 2</td>
<td>22</td>
<td>-0.287</td>
<td>0.026</td>
</tr>
<tr>
<td>Submarket 3</td>
<td>11</td>
<td>-0.386</td>
<td>0.020</td>
</tr>
<tr>
<td>Submarket 4</td>
<td>15</td>
<td>-0.502</td>
<td>0.026</td>
</tr>
<tr>
<td>Submarket 5</td>
<td>10</td>
<td>-0.619</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Figure 4 - Submarkets based on Ward’s linkage clusters: Functional surface – Decreasing marginal utility of living area

The regular concentric pattern of submarkets is punctuated by four exceptions: i) areas corresponding to urban expansion along the main axial roads; ii) the urban centre of Ilhavo; iii) the urban centre of Gafanha; iv) 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, conferring a very low premium for extra size (people are interested in a nice location and style of houses, and not so much in living space).

Figure 5 represents the territorial distribution of housing living area, marking a clear distinction between the smaller available space in inner urban areas and the increasing availability of space as we move towards the periphery. Once more, the beaches, secondary urban centres and main roads disturb the 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; the above mentioned contrast between Barra and Costa Nova beaches is also clear.
Figure 5 - Submarkets based on Ward’s linkage clusters: Housing characteristics– living space of house (m²)

Legend:
- Submarket 1
- Submarket 2
- Submarket 3
- Submarket 4
- Submarket 5
- Submarket 6

<table>
<thead>
<tr>
<th>Number of zones</th>
<th>Mean (Ln Area m²)</th>
<th>Std. Deviation (Ln Area m²)</th>
<th>Mean (Area m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submarket 1</td>
<td>17</td>
<td>4.588</td>
<td>98.298</td>
</tr>
<tr>
<td>Submarket 2</td>
<td>24</td>
<td>4.869</td>
<td>130.191</td>
</tr>
<tr>
<td>Submarket 3</td>
<td>9</td>
<td>5.164</td>
<td>174.863</td>
</tr>
<tr>
<td>Submarket 4</td>
<td>24</td>
<td>5.434</td>
<td>229.064</td>
</tr>
<tr>
<td>Submarket 5</td>
<td>2</td>
<td>6.144</td>
<td>465.914</td>
</tr>
</tbody>
</table>

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

Legend:
- Submarket 1
- Submarket 2
- Submarket 3
- Submarket 4
- Submarket 5
- Submarket 6

<table>
<thead>
<tr>
<th>Number of zones</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submarket 1</td>
<td>-2.346</td>
<td>0.228</td>
</tr>
<tr>
<td>Submarket 2</td>
<td>-0.957</td>
<td>0.220</td>
</tr>
<tr>
<td>Submarket 3</td>
<td>-0.231</td>
<td>0.209</td>
</tr>
<tr>
<td>Submarket 4</td>
<td>0.544</td>
<td>0.183</td>
</tr>
<tr>
<td>Submarket 5</td>
<td>1.256</td>
<td>0.089</td>
</tr>
<tr>
<td>Submarket 6</td>
<td>2.683</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figures 4 and 5 confirm a strong pattern of heterogeneity, over space, both in terms of living area corresponding implicit prices. However, the more interesting result is the pattern of overlapping between the two mappings. To illustrate this fact, we made a comparison by standardizing the values of living space and the results obtained using and taking their difference. The pattern that emerges is that, the lower the living space, the lower the returns. The signs of the standardized results were selected in line with the above idea; therefore, negative values in figure 6 imply that decreasing returns to living space are stronger than what would be expected, considering the available space, while positive values represent the opposite. There is a pervasive dominance of low positive and negative differences, with few exceptions such as Costa Nova and small areas which may be considered as outliers. As a consequence, and following the theorem in section 3.3, we can argue that the submarkets presented in Figure 6 are robust to all the three delineation methodologies; similarity in hedonic characteristics, similarities in hedonic prices and substitutability.

This is a clear case of partial overlapping between sub-markets delineated by either by marginal utility of living area or by the area (the visual comparison, of figure 4 and 5 gives a similar impression). Therefore, we conclude that the combination of both criteria gives the best delineation of sub-markets, supporting the assumption that two similar houses inside each sub-market are good substitutes.

Legend:

<table>
<thead>
<tr>
<th>Number of zones</th>
<th>FDA_coefficients (standardized values)</th>
<th>Ln Area m² (standardized values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submarket 1</td>
<td>2</td>
<td>-0,156</td>
</tr>
<tr>
<td>Submarket 2</td>
<td>17</td>
<td>-1,050</td>
</tr>
<tr>
<td>Submarket 3</td>
<td>18</td>
<td>-0,268</td>
</tr>
<tr>
<td>Submarket 4</td>
<td>20</td>
<td>0,782</td>
</tr>
<tr>
<td>Submarket 5</td>
<td>5</td>
<td>-1,137</td>
</tr>
<tr>
<td>Submarket 6</td>
<td>14</td>
<td>0,931</td>
</tr>
</tbody>
</table>

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

---

This analysis is also confirmed by simple regressions used only as a descriptive tool. Removing locations with less than 10 properties each, the $R^2$ value between hedonic prices and characteristics is 0.31; retaining these outliers, $R^2$ is 0.75.
Such delineation is shown in figure 7, where sub-markets, 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 sub-market 6, with the smallest living area and the highest premium for extra-size; sub-market 4 corresponds to the outer ring of Aveiro, with extensions along the main roads, but also to an inner city area (Gulbenkian and Bairoo do Liceu) with relatively big high quality houses; submarket 5 corresponds to the already mentioned case of Costa Nova and three other areas with few houses (which can be considered as outliers), where the reduce living space is coupled with abnormal values for marginal returns to space; the remaining submarkets reflect the expected pattern of peripheral areas.

7. Conclusion

The main objective of this paper was the definition of housing submarkets, both in terms of its conceptualization and in terms of empirical delineation. A new framework and methods based on functional data analysis is developed, integrating ideas and approaches from functional data analysis, spatial econometrics and locally weighted regressions.

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 substitutability of housing units. Applied to a specific urban area and submarkets were delineated using the proposed methodology, the results show that housing characteristic and prices produce sub-markets that partially overlap, suggesting that the combination of the former two criteria provides a more reasonable approach towards defining sub-markets, which also satisfies the condition of internal substitutability of similar houses.

The proposed synthesis and corresponding methods extend the literature along several directions. First, and most importantly, the framework allows 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. Third, once such submarkets have been delineated, spatial dependence can be examined by estimating cross- and within-submarket spatial weights, along the lines of Bhattacharjee and Jensen-Butler (2013) and Bhattacharjee et al. (2012).

A large number of further research problems and areas of development emerge from our work. While the framework enables analyses of endogenously produced submarkets, allowing for such endogeneity within the functional regression model is retained for future work. 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.

References


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