Day, Brett; Bateman, Ian; Lake, Iain

Working Paper

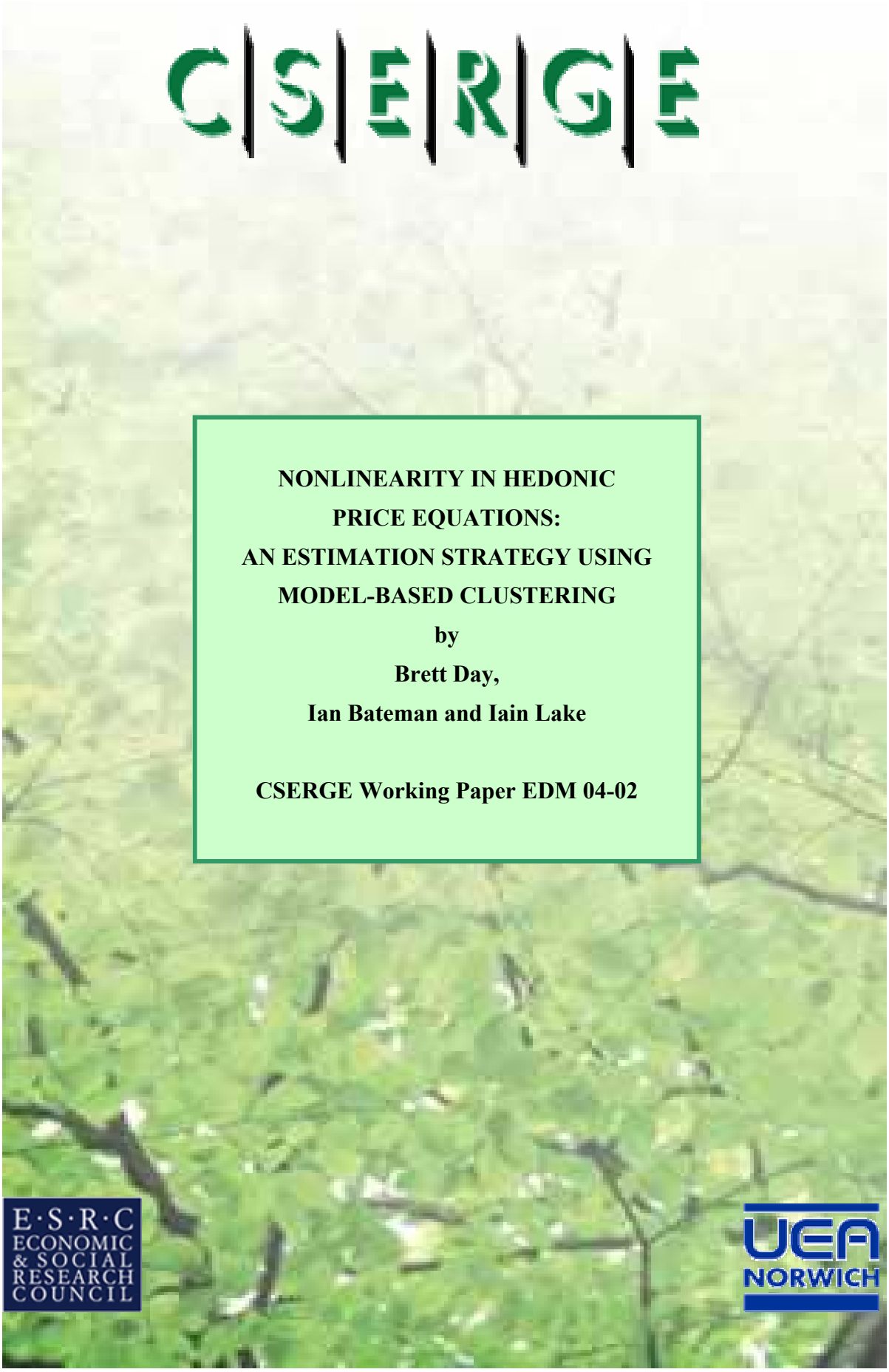
Nonlinearity in hedonic price equation: An estimation strategy using model-based clustering

CSERGE Working Paper EDM, No. 04-02

Provided in Cooperation with:
The Centre for Social and Economic Research on the Global Environment (CSERGE), University of East Anglia

Suggested Citation: Day, Brett; Bateman, Ian; Lake, Iain (2004) : Nonlinearity in hedonic price equation: An estimation strategy using model-based clustering, CSERGE Working Paper EDM, No. 04-02, University of East Anglia, The Centre for Social and Economic Research on the Global Environment (CSERGE), Norwich

This Version is available at:
https://hdl.handle.net/10419/80273


CSERGE
NONLINEARITY IN HEDONIC
PRICE EQUATIONS:
AN ESTIMATION STRATEGY USING
MODEL-BASED CLUSTERING
by
Brett Day,
Ian Bateman and Iain Lake
CSERGE Working Paper EDM 04-02
E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL
UEA
NORWICH

NONLINEARITY IN HEDONIC PRICE EQUATIONS: AN ESTIMATION STRATEGY USING MODEL-BASED CLUSTERING

by

Brett Day[1], Ian Bateman[1] and Iain Lake[2]

[1]Centre for Social & Economic Research on the Global Environment (CSERGE) School of Environmental Sciences University of East Anglia, UK

[2]Centre for Environmental Risk, School of Environmental Sciences University of East Anglia, UK

Acknowledgements

The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. This work was part of the interdisciplinary research programme of the ESRC Centre for Social and Economic Research on the Global Environment (CSERGE).

This work was funded in part by the UK Department for Transport as part of the project entitled *Valuation of Transport Related Noise in Birmingham and Benefit Transfer to UK*.

ISSN 0967-8875

Abstract

Recent advances in the theoretical understanding of equilibria in property markets predict that the equilibrium hedonic price function will typically be highly nonlinear. Rather than adopting progressively more flexible econometric specifications to deal with this nonlinearity we adopt an alternative estimation strategy based on a further insight provided by the theoretical literature. That insight is that in equilibrium the market may not be characterised by a continuum of properties over attribute space. Rather the market may well be lumpy, being well-provided with properties exhibiting certain combinations of characteristics and sparsely provided elsewhere. We test the predictions of two different models; one that suggests that the market will be characterised by clusters of properties with similar physical attributes, one that the market will be characterised by clusters of neighbourhoods exhibiting similar socioeconomic compositions. We identify clusters by applying techniques of model-based clustering which allow the data to inform on the nature and the number of clusters. Our estimation strategy for handling nonlinearity, therefore, is to avoid estimating the hedonic price function over the entire attribute space. Rather, we fit separate price functions for the properties in each cluster thereby forming local approximations to the hedonic price surface over the attribute area spanned by the properties in each cluster. Finally we test to see which partitioning of the data, either according to the attributes of properties or the socioeconomics of neighbourhoods, is capable of explaining more of the variability in the data.

Key words: Hedonics, functional form, model-based clustering

1. Introduction

In his seminal paper Sherwin Rosen (1974) endowed economists with a substantive theoretical framework within which to study market equilibria for differentiated commodities. Despite the generality of this theoretical framework attention has focused, to a large extent, on a particularly simple specification of the Rosen model that allows for a closed-form solution (e.g. Epple, 1987; Tauchen and Witte, 2001).

Ekeland et al. (2002) designate this the Normal-Linear-Quadratic (NLQ) model; the *Normal* and *Linear* referring to the distributional and functional assumptions underpinning the model, the *Quadratic* to the shape of the equilibrium hedonic price function that forms the solution to the model. The constant curvature of this quadratic hedonic determines that the implicit price functions for each property attribute are themselves linear. Furthermore, as Heckman et al. (2003) demonstrate, the market equilibrium described by the NLQ model is characterised by a set of products exhibiting a range of differing attributes. Given the regularity of this formulation of the model, it is no surprise to discover that the density of properties boasting different combinations of property characteristics is found to follow a normal distribution.

Hedonic analyses of data from property markets, the focus of this paper, have found little evidence to support the NLQ model as an adequate description of real world markets. Indeed, the received wisdom purports that empirical investigators can learn little from theory in their attempts to estimate hedonic price functions. [1]

However, recent years have witnessed renewed interest in the theoretical modelling of markets for differentiated goods. One set of models, typified by the work of Ekeland et al. (2002, 2003) and Heckman et al. (2002, 2003), have investigated the nature of the market equilibrium when some of the restrictive assumptions of the NLQ model are relaxed. In the context of the property market these models assume that households' choices are determined by the attributes of the properties themselves. In a parallel theoretical literature, Nesheim (2002) investigates the nature of property market equilibria when households' choices depend not on the characteristics of the properties themselves but on the characteristics of the equilibrium sets of people that choose to inhabit the neighbourhood in which those properties are located.

---

[1] Witness the papers by Halvorsen and Pollakowski (1981), Cassel and Mendelhsohn (1984), Blackley et al. (1984), Cropper et al. (1988) and Rasmussen and Zuehlke (1990) who all chose to introduce their papers by stressing that economic theory does not suggest an appropriate functional form for the empirical estimation of hedonic price functions.

In this paper, we examine these theoretical developments and observe that this literature throws up a number of important insights that may be of use in empirical applications. In particular, both sets of models predict that under all but the most contrived of assumptions, the equilibrium hedonic price will be far from quadratic. Rather the implicit price functions for property (or neighbourhood) attributes are shown to be generically nonlinear (Ekeland et al., 2003). Furthermore, the models suggest that property markets in equilibrium may be characterised by lumpy provision in attribute space. The market may be well-provided with properties (or neighbourhoods) with certain combinations of attributes and sparsely provided elsewhere.

In this paper we describe an application that exploits these insights. Using data from the City of Birmingham in the UK, we examine the attributes of properties and their neighbourhoods for evidence of clustering. The method by which we propose allocating properties to clusters is known as model-based clustering. In contrast to other clustering techniques model-based clustering provides a purely data driven methodology that simultaneously identifies the most appropriate clustering method and the most appropriate number of clusters into which the data should be partitioned (Fraley and Raftery, 1998). Following the two strands of the theoretical literature, we define two initial partitions of the data. In the first, we use attributes of properties to define clusters. In the second, we use the characteristics of the inhabitants of neighbourhoods to define clusters.

If the data confirms the existence of clusters then by definition the properties within them must lie in close proximity in attribute space. By extension, these properties must also lie close to each other on the hedonic price surface. Rather than employing increasingly more general econometric specifications to capture the nonlinearity of the equilibrium hedonic price function, our estimation strategy is to avoid estimating the hedonic price function over the entire attribute space. Rather, we fit separate price functions for the properties in each cluster thereby forming local approximations to the hedonic price surface over the attribute area spanned by the properties in each cluster.

Of course, if the parameters of the hedonic price function do not differ substantially over attribute space then such an estimation strategy will be inefficient. We test this hypothesis by establishing whether there are significant differences in the parameters of the hedonic price functions estimated for each separate cluster of properties.

Furthermore, we are interested to ascertain whether the property characteristics model (Ekeland et al., 2002, 2003; Heckman et al., 2002, 2003) or the neighbourhood characteristics model Nesheim (2002) forms the better approximation to the processes generating the data. To this end, we take our two initial partitions of the data; one based on the characteristics of the properties, the second on the characteristics of the inhabitants of neighbourhoods. Within

each partitioning we follow our estimation strategy of fitting separate price functions for the properties in each cluster. Following Goodman and Dubin (1990) we then employ non-nested tests to compare the two empirical models of the property market.

If the hedonic price function estimated using the property characteristics partitioning is found to statistically dominate those based on other partitions of the data, then the evidence indicates that the property characteristics model best reflects the processes at force in the market. Alternatively the hedonic price function estimated using the neighbourhood characteristics partitioning may be found to statistically dominate those based on other partitions of the data. In that case we might conclude that the neighbourhood characteristics model best reflects the processes at force in the market.

The rest of this paper is organised as follows. In Section 2 we briefly review the theoretical literature. In section 3 we discuss model-based clustering, our approach to defining clusters in the data. Section 4 describes the data collected from the City of Birmingham in the UK that is used in this study. Section 5, reports the results of the model-based clustering. Section 6 describes the results of the econometric exercise of fitting hedonic price functions to the different partitions of the data. Finally, Section 7 reports on the application of non-nested tests designed to answer the question of whether real world data best resembles a model in which households' property decisions are primarily driven by the attributes of properties themselves or by the characteristics of the inhabitants of neighbourhoods in which properties are located.

2. Models of Hedonic Markets

Rosen (1974) envisaged a market in which heterogeneous suppliers and heterogeneous consumers interact so as to establish an equilibrium maintained by an equilibrium hedonic price function. For example, in the property market, the focus of attention of this paper, households with differing preferences for the characteristics of properties interact with landlords differing in their costs of transforming the characteristics of their properties. An equilibrium is attained when the market settles on a hedonic price schedule that ensures households (within their limited budgets) cannot increase their utility by choosing a different property and landlords cannot increase their profits by increasing the property's rent or changing its characteristics.

Whilst the theoretical framework is quite general, attention tends to have focused on a particularly simple specification of the Rosen model designated the NLQ model (examined by Tinbergen, 1956, Epple, 1987 and Tauchen and Witte, 2001, amongst others). The NLQ model assumes that the heterogeneity of households and landlords is *normally* distributed in the population, imposes *linear* demand and supply functions and thereby provides a closed-form solution in the shape of a simple *quadratic* equilibrium hedonic price function with linear implicit prices. Heckman et al. (2003) show that in equilibrium the NLQ model predicts that a range of properties exhibiting different combinations of attributes will be provided to the market in equilibrium. Moreover, the density of these properties in attribute space is found to follow a normal distribution.

In a series of recent papers, Ivar Ekeland, James Heckman, Rosa Matzkin and Lars Nesheim (Ekeland et al., 2002, 2003; Heckman et al., 2002, 2003) have investigated the nature of the market equilibrium when some of the restrictive assumptions of the NLQ model are relaxed. Since, these problems no longer provide closed-form solutions, numerical methods are used to approximate the hedonic price function and characterise the market in equilibrium. Their analysis reveals that even minor perturbations from the assumptions of the NLQ model disrupt the neat simplicity of the equilibrium solution.

For example, Ekeland et al. (2003) abandon the assumption of normally distributed heterogeneity within the populations of households and landlords. Instead they model heterogeneity as the mixture of two different normal distributions. Naturally, the greater the degree of mixing, the further the distribution of heterogeneity strays from normal. Ekeland et al. (2003) discover that when the distribution of heterogeneity is non-normal, the market equilibrium is no longer characterised by a quadratic hedonic price function with linear implicit prices. Rather, the greater the degree of mixing, the greater the degree of non-linearity observed in the implicit price functions. Imposing

economically reasonable restrictions (i.e. positive implicit prices, only positive quantities of attributes demanded and supplied) only serves to exaggerate the nonlinearity of implicit prices. Indeed as Ekeland et al. (2003) prove in the context of the NLQ model, the implicit price schedules of the equilibrium hedonic price function are generically nonlinear.

Heckman et al. (2003) also examine the equilibrium density of properties exhibiting different levels of attributes. In the NLQ model, this density follows a normal distribution. They observe that as the distribution of heterogeneity is made increasingly non-normal, the density of properties in attribute space follows suit.

Heckman et al. (2003) extend their investigation by examining models in which the quadratic specifications of the household utility and landlord cost functions are replaced with higher order polynomials. Again, the increased flexibility of these specifications precipitates increasing nonlinearity in the implicit price schedules of the equilibrium hedonic price function. What is more, in these more flexible models, the equilibrium density of properties in attribute space is far from normally distributed. Indeed, in the cases illustrated in Heckman et al. (2003) the density of supply exhibits many modes; that is, in equilibrium there exists clusters of properties exhibiting similar combinations of attributes, whilst properties with other combinations of attributes are sparsely represented in the equilibrium market. As Heckman et al (2003) point out, "the model is capable of generating equilibria in which there are nearly gaps in the range of products marketed".

In a parallel theoretical literature. Nesheim (2002) investigates the nature of property market equilibria in which households choose where to live based on their willingness to pay for locational quality. In particular, he concerns himself with neighbourhood effects. That is, a model where households' valuations of properties depend not on the characteristics of the properties themselves but on the characteristics of the equilibrium sets of people that choose to inhabit the neighbourhood in which that property is located.

Paralleling the work of Ekeland et al. (2003), Nesheim finds that in all but the simplest cases, the curvature of the equilibrium hedonic price schedule is highly nonlinear. Indeed, Nesheim reports that for certain parameter values, a kinked price function is required in order to attain an equilibrium.

Similarly, Nesheim (2002) finds that in equilibrium the property market may be characterised by lumpy provision. Neighbourhoods boasting high and low levels of quality are relatively more common than those at intermediate levels. Moreover, Nesheim (2002) shows that households will sort themselves across the urban area such that the traits of households within a neighbourhood are likely to be less varied than those of the population as a whole. Indeed, the more correlated a trait is with WTP for locational quality, the more homogenous

neighbourhoods are likely to be in this trait and the greater will be the differences in the average level of this trait across neighbourhoods.

Clearly, once one moves outside the contrived realm of the NLQ model, theoretical investigations of hedonic property markets provide descriptions of a rich and varied urban landscape. The property characteristics models are capable of generating equilibria in which there exist clusters of properties exhibiting similar combinations of attributes, whilst properties with other combinations of attributes are sparsely represented. Likewise in the neighbourhood characteristics models, households are shown to sort themselves across the urban space such that neighbourhoods with residents showing particular combinations of characteristics may be well-represented in the equilibrium market, whilst neighbourhoods with residents exhibiting other combinations may be relatively rare. In both the property characteristics and the neighbourhood characteristics models, the market equilibrium is maintained by a hedonic price function that may be highly nonlinear and quite possibly kinked.

3. Identifying Clusters in Property Market Data

3.1 Submarkets versus clusters of properties with similar attributes

There is a long-established literature on the existence and identification of housing submarkets within an urban area that bears some resemblance to the work presented here (e.g. Straszheim, 1973, 1974; Ball and Kirwan, 1977; Schnare and Struyk, 1976; Sonstelie and Portney, 1980; Goodman, 1978; Michaels and Smith, 1990; Allen et al., 1995; Wolverton et al., 1999; Goodman and Thibodeau, 1998, 2003). However, contrary to the argument advanced in this paper, these papers motivate the existence of clusters of properties exhibiting different pricing structures through imperfections in the market mechanism. For example, Goodman and Thibodeau (2003) state that "due to either supply-or demand-related factors, the normal arbitrage that would be expected to equalize prices both within and across metropolitan areas may work either slowly, or not at all". Likewise Can (1992) states that "… varying attribute prices … indicate the presence of independent price schedules, thus the existence of a segmented market. The presence of geographic submarkets violates the assumption of a long-run equilibrium in urban housing markets since there will be independent hedonic price schedules within a single metropolitan area reflecting the demand and supply structures of submarkets."

Of course, the theoretical literature described in the introduction paints a quite different picture of the mechanisms at work in property markets. In particular, it shows that differences in prices across urban areas are not the result of market imperfection or disequilibrium, but rather are an integral part of the price mechanism establishing equilibrium in the property market. The Nesheim (2002) model for example predicts that identical properties in different neighbourhoods can command radically different prices depending on the characteristics of the inhabitants of that neighbourhood. Likewise the models described by Ekeland et al. (2003) allow for the fact that properties in the same neighbourhood may command radically different implicit prices for attributes depending on the characteristics of the particular property.

Within the housing submarket literature, therefore, the definition of submarkets has tended to be dominated by the identification of property or neighbourhood characteristics that define market barriers. For instance Goodman and Thibodeau (2003) suggest that racial discrimination may produce separate submarkets for those of different ethnic origin, or that distinct sub-populations of households with strong preferences for either newly constructed properties or for historic properties may segment property markets according to the ages of properties.

In this application, however, partitioning of the data is not motivated by the supposed existence of different market segments but by the prediction of the

theoretical models that property markets in equilibrium may be characterised by lumpy provision in attribute space. That is, that the market may be well-provided for certain combinations of property or neighbourhood characteristics and sparsely-provided elsewhere.

The existence of such clusters of properties is distinct from the notion of market segmentation. As such, our approach to identifying clusters is not shackled by the need to provide a formal definition of the process driving market segmentation or to formally define the property or neighbourhood characteristics by which such segments should be delineated. Rather in this paper, the data itself is used to inform on the pattern of clustering in the property market. The method by which we propose allocating properties to clusters is known as *model-based clustering*.

Clustering techniques have seen some application to the classification of properties into submarkets, notably Abraham et al. (1994), Goetzmann and Wachter (1995), Hoesli and MacGregor (1995), Bourassa et al. (1999), Day et al. (2003) and Day (2003). Though, since this literature is predicated by the existence of barriers to the attainment of market equilibrium, these papers do not provide a coherent justification for the use of data driven clustering techniques. Furthermore, these studies all use relatively simple clustering algorithms that provide no independent statistical indication of the nature or number of clusters to be found in the data.

3.2 Model-based cluster analysis

The basic aim of cluster analysis is to sort observations into a classification based on a set of P variables defining the characteristics of each observation. A common starting point is to define each observation as a point in P-space whose location is determined by how highly that observation scores for each variable. Clearly, observations holding similar values for the different variables will be located close to each other in this P-space. Clusters can be identified as concentrations of observations falling into the same region of this P-space. Individual observations can be classified according to their proximity to different clusters.

In recent years a number of new approaches to identifying clusters in data have been proposed.[2] One approach that has shown particular promise is that of model-based clustering (McLachlan and Basford, 1988; Banfield and Raftery, 1993; Fraley and Raftery, 1998; Fraley and Raftery, 2002a). This clustering approach has been successfully applied to a variety of data problems across a broad range of disciplines. For example, in the biological sciences to analyse

[2] For a review of recent advances see Fasulo (1999) and Jain et al. (1999).

gene expression data (e.g. Ghosh and Chinnaiyan, 2002; McLachlan et al., 2002; Yeung et al., 2001), in ecology to study community composition (e.g. Ter Braak et al., 2003), in atmospheric sciences to study circulation patterns (e.g. Smyth, 2000; Smyth et al., 1999), in astronomy to classify gamma ray bursts (e.g. Mukherjee et al, 1998) and in various fields for image analysis (e.g. Gopal and Hebet, 1998; Campbell et al, 1999; Wehrens et al., 2003). In contrast, model-based clustering techniques are relatively unknown in economic analysis. As far as the authors are aware this is the first application of these techniques in this field

The fundamental assumption of model-based clustering is that each data point is drawn from a population of such points constituting all the members of the cluster. Moreover, the location, size and shape of this underlying population can be approximated by a probability distribution. Assuming a Gaussian distribution, for example, would imply that clusters are ellipsoidal. It would also assume that the likelihood of observing data points belonging to a particular cluster is greater near the mean location of that cluster than at its periphery. The data observed by the researcher is the composite of data points drawn from a finite number of such clusters.

To formalise, each P-dimensional data point $\boldsymbol{x}$ arises from a super population comprising a mixture of M populations, $C_1, C_2, \ldots, C_M$, in some proportions $\pi_1, \pi_2, \ldots, \pi_M$ respectively, where;

$$\sum_{j=1}^{M} \pi_j = 1 \quad \text{and} \quad \pi_j \geq 0 \quad (j = 1, 2, \ldots, M) \tag{1}$$

If we assume that each population, C_j, can be modelled as a P-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, then the probability density function (pdf) of an observation $\boldsymbol{x}$ is of the form;

$$\begin{aligned} f(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) &= \sum_{j=1}^{M} \pi_j f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j) = \sum_{j=1}^{M} \pi_j \phi_j^p(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\ &= \sum_{j=1}^{M} \pi_j \left(2\pi^p \left|\boldsymbol{\Sigma}_j\right|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right) \end{aligned} \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$ is the vector of parameters associated with the assumed distributions of the M clusters, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_M)$ is the vector of mixing proportions, $f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)$ is the pdf of cluster C_j which is given the specific P-dimensional Gaussian form denoted $\phi_j^p(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Thus, in the Gaussian case $\boldsymbol{\theta}_j$ comprises the elements of the vector $\boldsymbol{\mu}_j$, which determine the mean location of each cluster in P-space, and the distinct elements of the covariance matrices $\boldsymbol{\Sigma}_j$, which determine the geometric proportions of each cluster. Given (2) The mixture model describing the pattern of clustering can be formalised into the likelihood function;

$$L_M(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j f_j(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j) \right] \tag{3}$$

where $\boldsymbol{X}$ is the $N \times P$ matrix of data by which the N observations are to be clustered.

To allow for comparison of different assumptions concerning the geometric characteristics of the different clusters, Banfield and Raferty (1993) reparameterise each covariance matrix $\boldsymbol{\Sigma}_j$ using the eigenvalue decomposition;

$$\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{D}_j' \quad (j = 1, 2, \ldots, M) \tag{4}$$

where $\boldsymbol{D}_j$ is the matrix of eigenvectors, λ_j is the first eigenvalue of $\boldsymbol{\Sigma}_j$, and $\boldsymbol{A}_j$ is a diagonal matrix with diagonal elements $1 = \alpha_{1j} \geq \alpha_{2j} \geq \ldots \geq \alpha_{pj} > 0$.

The advantage of Banfield and Raferty's decomposition is to isolate different geometric properties of each cluster into different components. Hence λ_j determines cluster volume, $\boldsymbol{D}_j$ cluster orientation and $\boldsymbol{A}_j$ other properties of the cluster shape. Thus imposing the restriction $\lambda_j = \lambda \ (j = 1, 2, \ldots, M)$ enforces equality of volume across all clusters. Similarly, imposing the restriction $\boldsymbol{A}_j = I \ (j = 1, 2, \ldots, M)$, where I is the P-dimensional identity matrix, generates strictly spherical clusters. Clearly differing combinations of restrictions imply different imposed similarities between clusters. As we shall see shortly, the great advantage of model-based clustering is that it provides a formal framework in which such restrictions can be compared.

For now, imagine that the number of clusters, M, in the data is known. Also imagine we know to which of these clusters each data point belongs. In that

case Equation 3 can be reformulated as the complete-data log likelihood as follows;

$$\ln L_M(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \sum_{j=1}^{M} d_{ij} \ln\left[\pi_j \phi_j^p(\boldsymbol{x} \mid \boldsymbol{\theta}_j)\right] \tag{5}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \lambda_j, \boldsymbol{A}_j, \boldsymbol{D}_j)$ (j = 1, 2, …, M) and d_{ij} (i =1, 2, …, N; j = 1, 2, …, M) are indicator variables whose value is 1 if observation i belongs to cluster C_j and 0 otherwise.

Of course, we do not know the provenance of each data point; from the researcher's point of view the d_{ij} are missing data. As Celeux and Govaert (1995) describe, this motivates a simple application of the EM algorithm (Dempster et al., 1977).

The E-step of the algorithm calculates;

$$\bar{d}_{im} = E\left[d_{im} \mid \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}}\right] = \frac{\hat{\pi}_m \phi(\boldsymbol{x}_i \mid \hat{\boldsymbol{\theta}}_m)}{\sum_{j=1}^{M} \hat{\pi}_j \phi(\boldsymbol{x}_i \mid \hat{\boldsymbol{\theta}}_j)} \quad (m = 1, 2, \ldots, M) \tag{6}$$

where $\bar{d}_{im}$ is the expected value of the indicator variable for membership of cluster m conditional on the data and current parameter estimates denoted $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\pi}}$ with cluster specific components $\hat{\boldsymbol{\theta}}_j$ and $\hat{\pi}_j$ respectively.

In the M-step, $\bar{d}_{im}$ replaces d_{im} in the complete-data log-likelihood (Equation 5), which is then maximised with respect to the parameters. Solution of this maximisation problem provides simple closed forms for the mean cluster locations and mixing probabilities;

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{i=1}^{N} \bar{d}_{im} \boldsymbol{x}_i}{N_m}; \quad \hat{\pi}_m = \frac{N_m}{N}; \quad N_m \equiv \sum_{i=1}^{N} \bar{d}_{im} \quad (m = 1, 2, \ldots, M) \tag{7}$$

Estimating the elements of the covariance matrix $\hat{\boldsymbol{\Sigma}}_m$, in the M-Step, depends on the particular parameterisation. Further details of these computations using the eigenvalue decomposition in (Equation 4) can be found in Celeux and Govaert (1995). The E-step and M-step are iterated until convergence of the parameters.

The value $\bar{d}^*_{im}$ of $\bar{d}_{im}$ that maximises (Equation 5) gives the conditional probability that observation i belongs to cluster C_m. A maximum likelihood classification of the data can be derived by associating each observation with the cluster to which it is most likely to belong. That is, observation i is classified as belonging to cluster C_m if $d^*_{im} = \max_j d^*_{ij}$. Furthermore, $1 - \max_j d^*_{ij}$ gives a measure of the uncertainty associated with each observation's classification (Bensmail et al., 1997)

As is clear from Equations 6 and 7, the EM algorithm decomposes the problem of maximising the mixture model log-likelihood (Equation 3) into a series of relatively simple calculations. As described by Fraley and Raftery (2002a), this simplicity comes at a cost. In particular, the conditions under which the algorithm can be proven to converge to a local maximum do not always hold for mixture models. Nonetheless, Fraley and Raftery (2002a) indicate that EM estimation has been applied with considerable success in this context. Furthermore, the rate of convergence of the algorithm may be very slow and may encounter difficulties if there are a large number of clusters or the data is ill-conditioned. As with all maximisation problems, the chances of reaching a satisfactory solution are greatly enhanced by initialising the algorithm with reasonable starting values, a subject we shall return to discuss shortly.

One limitation of the model as presented so far is that it assumes that all data points belong to a cluster. A more general model would allow for the presence of noise or outliers. Banfield and Raftery (1993) suggest that data points belonging to the noise could be modelled as being draws from a homogeneous Poisson process. That is, having removed observations belonging to clusters, the distribution of the remaining data points is one in which the expected number of "noise" observations in any location in the P-space defined by the clustering variables is identical.

The existence of "noise" observations adds an extra component to the mixture distribution of Equation 2. That is there is a constant $1/V$ density of observations over the entire P-space where V is the volume of that space. In this case the likelihood function can be rewritten;

$$L_M\left(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi}\right) = \prod_{i=1}^{N} \left[\frac{\pi_0}{V} + \sum_{j=1}^{M} \pi_j f_j\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j\right) \right] \tag{8}$$

If an observation belongs to the noise it contributes $1/V$ to the mixture likelihood, if it belongs to one of the clusters it contributes a Gaussian term. The estimation of the mixing proportion for the Poisson process, π_0, is easily

achieved within the EM procedure discussed previously (Fraley and Raftery, 1998, 2002a, 2002b).

3.3 Model selection

One problem that remains is how to choose between clustering solutions allowing different numbers of components and differing parameterisations of cluster shapes. In contrast to other clustering algorithms, the probabilistic basis of model-based clustering provides a framework within in which these comparisons can be made.

A Bayesian approach to model selection is to choose the model that is most likely *a posteriori*. Given that *a priori* all models are considered equally likely, this amounts to comparing the integrated likelihood of the different models. Unfortunately, even for relatively simple Gaussian mixture models this integral has no closed form. An alternative then, is to use penalized likelihood methods that approximate the integrated likelihood. One such method is the Bayesian Information Criterion (BIC) (Schwartz, 1978);

$$BIC_g \;=\; 2\,\ln L\left(\boldsymbol{\pi}_g^*, \boldsymbol{\theta}_g^*\right) - \upsilon_g \log\left(N\right) \tag{9}$$

where g indexes the particular model being evaluated, $\boldsymbol{\pi}_g^*$ and $\boldsymbol{\theta}_g^*$ are the maximum likelihood estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ respectively and υ_g are the total number of independent parameters in $\boldsymbol{\pi}_g^*$ and $\boldsymbol{\theta}_g^*$. If a BIC statistic is calculated for two different models, the difference between their BICs is what will indicate the superiority of one model over the other. If the difference is large enough, one can be reasonably certain that one model gives a better fit than the other. A standard convention for calibrating BIC differences is that differences of less than 2 correspond to weak evidence, differences between 2 and 6 to positive evidence, differences between 6 and 10 to strong evidence, and differences greater than 10 to very strong evidence.

Whilst, the regularity conditions necessary for the BIC to approximate the integrated likelihood do not hold for finite mixture models (Titterington et al., 1985), a growing weight of theoretical and empirical evidence supports the use of the BIC in this context (Leroux, 1992; Keribin, 1998; Roeder and Wasserman, 1997; Campbell et al., 1999; Dasgupta and Raftery, 1998; Fraley and Raftery, 1998, 2002; Stanford and Raftery, 2000).[3]

---

[3] Other approaches to model comparison include the approximate weight of evidence (AWE) criterion employed by Banfield and Raftery (1993), the NEC, an entropy criterion proposed by Biernacki et al. (1999), a

The approach followed here then is that outlined in Fraley and Raftery (1998). In the first instance select a range for M, the number of clusters. Then select a series of parameterizations of the covariance matrix by applying one or more equality restrictions to Equation 4. For each value of M and each parameterization, use the EM algorithm to calculate the maximum likelihood estimates of the model parameters. Compute the BIC for each model. The model providing the highest value of the BIC is then selected.

In large datasets, the BIC tends to favour models with many clusters (Posse, 2001). Thus we follow the example of Posse (2001) who suggests "picking a good candidate in the region where the rate of change of the BIC drops significantly".

3.4 Initialisation of the EM Algorithm

Since the likelihood surface is typically characterised by many local maxima, finding appropriate starting values for the EM algorithm is a very important issue (Biernacki et al., 2003, compare various initialisation strategies). One approach is to derive starting values through a two-step methodology (Fraley and Raftery, 1998). First the data must be partitioned in to those observations that are thought to fall into the clusters and those that are thought to be part of the noise. Second, using just the denoised data, observations are given an initial allocation to clusters using hierarchical clustering techniques.

In particular, Fraley and Raftery (1998) propose initialisation of the denoised data using model-based hierarchical clustering (Banfield and Raftery, 1993) with an unconstrained covariance matrix. Here each observation begins in a cluster of its own and at each stage a pair of clusters are merged so as to maximise a log-likelihood function (see Banfield and Raftery, 1993, for details). Each step in the hierarchical clustering defines a unique number of clusters until in the final step all observations are tied together in one cluster. The output from this hierarchical clustering can be illustrated as a dendrogram revealing the association between observations.

To categorise the observations into M partitions, a section can be taken through the dendrogram at the level isolating M clusters. Fraley and Raftery (1998) propose using this categorisation to provide starting values for the cluster membership indicators d_{ij} $\left(i = 1, 2, \ldots, N; j = 1, 2, \ldots, M\right)$. These, in turn, can replace the conditional probabilities (Equation 5) to feed into the initial M-Step of the EM iteration.

---

bootstrap approach followed by McClachlan (1987) and the cross-validated likelihood approach of Symthe (2000). Comparisons of some of these different approaches can be found in Pan et al. (2002) and Biernacki and Govaert (1999).

One shortcoming of this approach is the onerous computing requirements of hierarchical clustering methods. In particular, the initial step of agglomerative hierarchical methods requires a measure of distance to be calculated between each observation in the data. As a result computing time and storage requirements are at least quadratic in the number of observations. Consequently, hierarchical clustering of large datasets may prove unfeasible.

Two basic approaches have been forwarded to overcome these constraints. Banfield and Raftery (1993) propose clustering a subsample of the data then using discriminant analysis to classify the remaining observations (see also Maitra, 2001; Wehrens et al., 2003). Alternatively, Posse (2001) suggests an approach that takes into account all of the observations in the dataset. Rather than beginning the hierarchical agglomeration from the set of singleton clusters, Posse (2001) proposes initially categorising observation into a smaller number of well-defined clusters. Provided the initial clustering is efficient, such that it only groups observations that would naturally fall into the same cluster at a relatively early stage in the hierarchical agglomeration, this approach should result in a similar classification as that achieved through a hierarchical clustering of the entire dataset.

Here we adopt the Posse (2001) approach that draws on graph theoretic approaches to clustering. In particular, Posse (2001) suggests generating clusters from the minimum spanning tree (MST) of the data. A spanning tree is a graph that connects all the data points in *P*-space such that there is only one path connecting each pair of data points. The MST is the spanning tree in which the total length of the connections or edges joining each point is at a minimum.

Posse's (2001) approach involves two steps in which the MST is first "peeled" and then "pruned". The peeling step involves trimming out the longest edges of the MST. In effect this divides the well-separated groups in the dataset into discrete clusters whilst also isolating observations on the periphery of clusters that would not be assigned to a cluster until late on in the hierarchical agglomeration. The pruning step involves dividing the surviving connected observations in the MST into small groups each of roughly the same size. These groups should consist of close neighbours that would have been merged early on in the hierarchical clustering. Observations that are connected after peeling and pruning are given the same classification and this acts as the partition from which the hierarchical clustering is initiated.

To determine which edges in the MST are considered sufficiently long to warrant "peeling", Posse (2001) proposes the use of plots comparing the observed distribution of the longest edge lengths in the data with those that would be expected if the data had come from a single Gaussian population. To this end, Posse (2001) extends a theorem of Penrose (1998) that describes the expected distribution of edge lengths in the MST if data points were drawn from

a single standard *P*-dimensional Gaussian distribution. In particular, the theorem states that for the standard Gaussian distribution, the probability of observing the l^{th} longest edge in the MST to have length e_l, is given by the l^{th} order Gumbel distribution once e_l has been suitably centred and scaled. More formally;

$$\lim_{N\to\infty} P\left[a_N e_l - b_N \le x\right] = G_l(x)$$

where

$$\begin{aligned} G_l(x) &= \exp\left(-e^x\right)\left(1+\sum_{j=1}^{l-1}\exp(-jx)\right)/\Gamma(l)\right) \\ a_N &= 2^{1/2}\left(\ln N + \left((p/2)-1\right)\ln_2 N - \ln\Gamma(p/2)\right)^{1/2} \\ b_N &= (p-1)\ln_2 N - \left((p-1)/2\right)\ln_3 N - \ln\kappa_p \\ \kappa_p &= 2^{-p/2}(2\pi)^{-1/2}\Gamma(p/2)(p-1)^{(p-1)/2} \end{aligned}$$

and

$$\ln_2 N = \ln\ln N, \quad \ln_3 N = \ln\ln\ln N \tag{10}$$

Thus, the quantities $u_1 = G_1(a_N e_1 - b_N)$, $u_2 = G_2(a_N e_2 - b_N)$, …, $u_N = G_N(a_N e_N - b_N)$ are identically (though not independently) distributed according to the uniform distribution in the interval [0,1].[4]

However, if the data are not in fact drawn from a single standardised Gaussian population the u_l-sequence will not show this pattern. As Posse (2001) describes, if separate clusters are present in the data, the ordered sequence of edge lengths given by e_l will tend to be longer than expected for early elements in the sequence. Similarly, if clusters are not homogenous but are elongated in some dimension then early values of the edge length sequence e_l will also be longer than expected. As a consequence, in either or both of these cases, the observed u_l-sequence will be characterised by initial values close to 1 before decreasing rapidly towards 0.

Posse (2001) suggests that the number of edges to be peeled should be determined by plotting both the u_l-sequence and the e_l-sequence. These plots should reveal the point at which the u_l-sequence stabilises around 0 and the point at which the rate of decay in the e_l-sequence drops significantly. Posse

---

[4] Note that the u_l-equence can be easily calculated since for l > 7 the Gl distribution is accurately approximated by the Gaussian distribution with mean $\mu_l = -\ln l + 1/2l$ and standard deviation $\sigma_l - (l-1/2)^{-1/2}$. Further, Posse (2001) notes the slow rate of convergence of the limit in Equation (10) and provides second order corrective terms for a_N and b_N obtained from Monte-Carlo simulations. As acknowledged by the author, there is a small error in Posse's (2001) Equation (6) where the Monte Carlo correction term should in fact be subtracted from b_N rather than added.

(2001) indicates that a suitable choice for the number of edges to peel is the largest of these two quantities.

The Posse procedure, partitions the data into a large number of well-defined small clusters. Observations falling into these clusters will be located close to each other in *P*-space. As such, a hierarchical clustering of these initial clusters should be little different from a hierarchical clustering based on the individual data points.

Thus far, our discussion of the initialisation of the EM algorithm for model-based clustering has failed to deal with the issue of denoising the data; that is, which data points are ascribed to the noise and which are included in the hierarchical clustering used to identify an efficient initial partitioning of the denoised data into clusters. Fraley and Raftery (1998) use a nearest neighbour denoising procedure proposed by Byers and Raftery (1998). In effect this procedure assumes that the data can be viewed as a mixture of two homogenous poisson processes with different intensities. The data in the clusters is drawn from the process with the greater intensity and hence will tend to be closer to its neighbours. In contrast data in the noise will be drawn from the less intense process and will be more distant from its neighbours. The procedure works by allocating an observation to the noise or the clusters according to its proximity to its neighbours.

Here we propose an alternative procedure for allocating observations to the noise based on Posse's (2001) procedure for peeling and pruning the MST. During peeling and pruning the longest edges of the MST are trimmed thereby isolating well-observations in the dataset. Consequently, we allocate all observations that have been isolated into single observations cluster following the application of the Posse procedure to the noise. The remaining observations are allocated to the denoised data and classified using hierarchical clustering. Subsequently, the data is recombined and the partitioning of the data into noise and separate clusters is used to initialise the EM algorithm for model based clustering.

3.5 Geographic smoothing

One last issue remains to be resolved. We might expect that for properties, geographical location will play an important role in determining submarket membership. Unfortunately directly including locational variables in a cluster analysis when observations are spread reasonably homogeneously across space, tends to result in a large number of clusters that are nearly circular when spatially mapped (Fovell, 1997). As a result we follow Posse (2001) and post-process the clustering classification to take account of the spatial information.

Here we adopt a very simple rule. The six closest observations in geographical space to each observation are identified. These seven observations are examined and their classification noted. If the majority of these observations favour one classification and this differs from the classification of the target observation then the probabilities of belonging to these two different clusters (as given by the respective values of d_{im}^{*}) are compared. Only if the target observation is less that twice as likely to belong to its current classification is the classification switched. This spatial smoothing rule is applied to all observations and the process iterated until no observations change classification.

3.6 Overall clustering strategy

The clustering strategy followed in this paper, therefore, follows a number of steps;

1. Construct the MST. In our application the MST is constructed using Prim's (1957) algorithm. To account for different scaling in the P clustering variables, inter-point distance is measured using a Mahalanobis metric.
2. Peel and prune the MST. Data points that are isolated in single-observation clusters following peeling and pruning of the MST are allocated to the noise. Data ponits that are connected in multiple-observation clusters are given the same classification and this classification acts as the partition from which the hierarchical clustering is initiated.
3. Perform an agglomerative model-based hierarchical clustering on the denoised data, starting from the initial partition determined in step 2
4. Determine the number of clusters, M, and a parameterisation for the cluster covariance matrices.
5. Perform a model-based clustering of the data using the EM algorithm. The EM algorithm is initiated with observations allocated to the noise or clusters according to the rule in step 2. For observations that are not part of the noise, initial cluster membership is determined from the hierarchical clustering in step 3.
6. Calculate the *BIC* for this model
7. Repeat steps 4 to 6 for various numbers of clusters and parameterisations of the cluster covariance matrices.
8. Plot the values of the *BIC* for the different models and select the model with the largest value of the *BIC* or choose a good candidate model as that giving the highest value for the BIC in the region where the rate of change of the BIC drops significantly.

3.7 Software

Software implementing the MST initial partition has been written by the authors in the GAUSS programming language. This code has subsequently been verified through comparison and with Christian Posse's original code designed to interface with the S-PLUS software package.

The model-based clustering (both hierarchical and EM) and BIC calculation have been implemented using Fraley and Raftery's (2002) MCLUST package designed to interface with either S-PLUS or R. The MCLUST software is available over the internet at http://www.stat.washington.edu/mclust.

4. The City of Birmingham Dataset

Hedonic valuation is a data intensive technique. The success or failure of a study hinges upon the quality of the data upon which it is based. In general, researchers require information on the selling price of properties, the structural characteristics of those properties, indicators of each property's proximity to (dis)amenities, descriptors of the socioeconomic characteristics of property neighbourhoods and data on the environmental quality of each property location.

The case study described in this paper is from the City of Birmingham in the UK. Records of all property sales in Birmingham during 1997 were obtained from the databases of the UK Land Registry[5]. These records indicated selling prices, dates of sales and full property address for each residential property transaction.

The Valuation Office Agency (VOA) provided property characteristics data. The VOA is an executive agency of the Inland Revenue, one of whose main functions is to value property for taxation purposes. In order to perform this function, the VOA maintains a database describing the structural characteristics of every residential property in England.[6] Amongst other details, the VOA provided data on the number of bedrooms and bathrooms in each property, total floor area, the property's age, whether the property was a bungalow or house (flats are not included in the analysis), whether the property was detached, semi-detached, in a terrace or at the end of a terrace, whether the property had central heating and access to off-road parking. Furthermore, the VOA classifies properties according to age and style of construction into one of around 30 property types called Beacon Groups. This information was also recorded as it provides a useful additional indication of property quality that cannot be determined from size and age alone.

Addresses were geolocated using a GIS. Subsequently GIS datasets were used to provide details of the garden area and aspect of each property and to calculate straight line distances, car travel times and walking distances from each property to (dis)amenities including schools, shops[7], railway stations and industrial sites.

---

[5] The Land Registry database is not publicly accessible information for England and Wales. However, the UK Department for Transport (DfT), who funded this study, arranged access for the purposes of this research.

[6] Unfortunately, the VOA data sources are currently held as paper records. Consequently, the process of matching addresses to the structural characteristics of each property required laborious trawling through ranks of filing cabinets.

[7] Specifically businesses registered as "Delicatessens", "Grocers", "Newsagents" or "Supermarkets".

When considering the accessibility of properties to shops, any measure based on proximity to only one facility has disadvantages. For example, a property 200m from ten shops is likely to be perceived as having better accessibility than another property 200m from one shop. As a result, measures for access to shops were constructed using a weighted sum of distances to all shops. This is a common procedure in accessibility studies and formalises to:

$$A_i = \sum_{j=1}^{J} \alpha_j e^{-\delta d_{ij}} \tag{11}$$

Where, A_i is accessibility at property i, α_j is the attractiveness of shop j, d_{ij} is the walking distance in kilometres between property i and shop j, δ exponent for distance decay and J is the number of shops in the region. Here we set $\delta = 2$ (such that a shop 100m from the property receives a weight over 6 times that of a shop at 1km distance and shops at over 2km distance receive almost no weight at all) and $\alpha_j = \alpha = 1$ (such that all shops are considered equally attractive). This shop accessibility variable is illustrated in Figure 1.

A similar procedure was used when considering accessibility to primary schools. Recent research suggests that selection procedures for primary school intake that favour local residents can considerably inflate house prices around high performing schools (Gibbons and Machin, 2003).[8] For each primary school in the Birmingham area an estimate of school quality was calculated as the percentage of pupils achieving Level 4 or above in Science, Mathematics and English (the level expected of 11 year olds).[9] A primary school accessibility index was constructed using (9) with the weight α_j set to this measure of school quality and $\delta = 1$. Figure 2 presents the primary school quality/accessibility variable is depicted for a region of the study area.

[8] As Gibbons and Machin (2001) argue, the issue is thought less important for secondary schools that typically draw from much wider catchments. Also, high educational achievement at primary school level may be a pre-requisite for admission to selective secondary schools. For example, the five selective Grammar Schools of King Edward the Sixth in Birmingham make offers " … solely on the basis of performance in the entrance test. Special allowances are not made for brothers or sisters or distance from the school." (quote taken from the Grammar Schools of King Edward VI in Birmingham web site http://www.kingedwardthesixth.org/eligibility.htm)

[9] This information was obtained for 1997 from the Department for Education and Employment website (http://www.dfee.gov.uk/performance/primary_97.htm).

Figure 1: Shop accessibility scores for a selection of properties

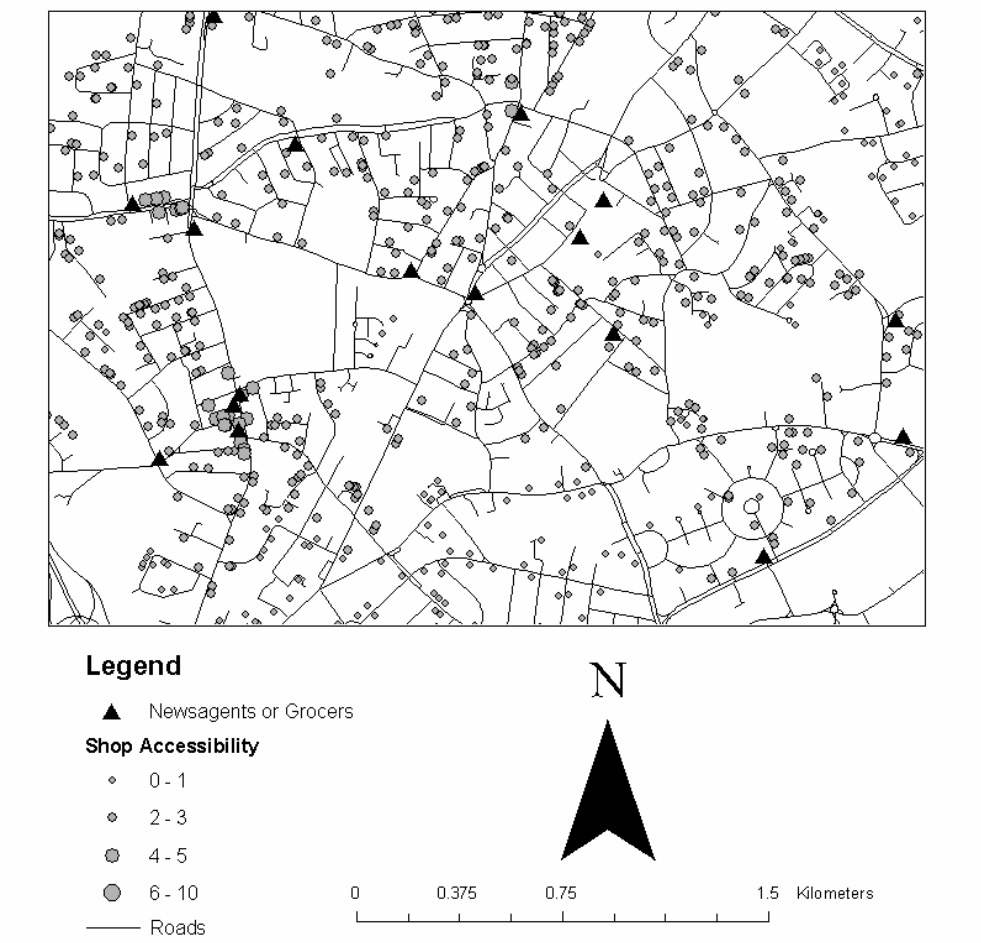
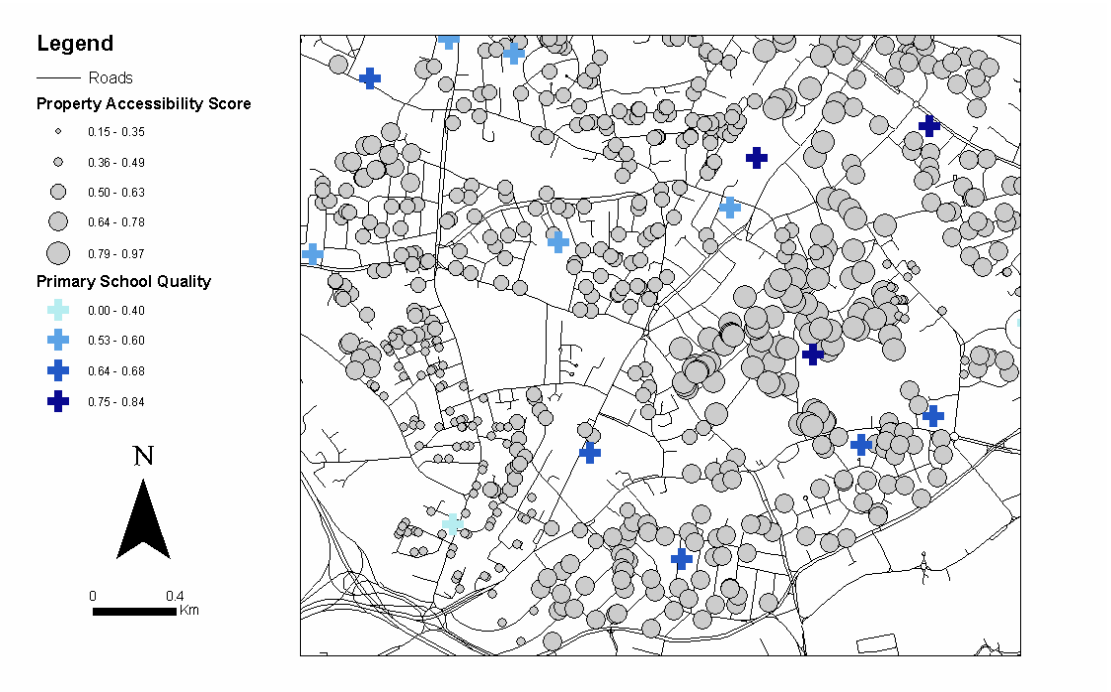


Figure 2: Primary School accessibility scores for a selection of properties



Using a procedure outlined in (Lake et al., 2000) data on land uses and the location and orientation of each property was combined with information on the landscape topology and building heights to calculate indices of the views available from the front and back of each property. For example indices were constructed for visible road surface, recreational park land and water surfaces.

Finally, road traffic and rail traffic noise data was provided by the Birmingham 1 project (DETR, 2000). The aircraft noise level at each property was identified by digitising a 1999 aircraft noise contour map of Birmingham International Airport. This map displayed aircraft noise levels in 3dB steps. Each property was assigned a noise level by interpolating linearly between the contours. All noise measurements are in decibels L_{EQ}.

Data on the socioeconomic composition of property neighbourhoods were drawn from the 1991 UK census provided by the Office for National Statistics (ONS). For the purposes of this research we recognise two levels of neighbourhoods. The smallest area over which census data is provided by the ONS is an enumeration district (ED). Birmingham is divided into 1,940 EDs, with each ED containing an average of 191 households. EDs are gathered into larger scale political units known as wards. Birmingham contains 39 wards such that each ward comprises an average of 50 EDs and 9,500 households. The organisation of these spatial units are shown in Figure 3.

The census provides a myriad of information on the socioeconomic characteristics of the population living in each ED. As we shall discuss in the Section 3, census data are ideal for constructing indicators of the attributes of the neighbourhood in which a property is located.

Descriptions of the variables used in the hedonic analysis are listed in Table 1. Complete data records were successfully compiled for some 10,848 residential property transactions in Birmingham in 1997. Further examination of the data lead to the exclusion of another 57 observations for various reasons. For example, 16 adjoining properties along one road were sold within a few months of each other at prices well below the apparent market rate. Examination of recent aerial photographs of this area provided an explanation; the houses had since been demolished to make way for a road widening scheme. The final data set used in this analysis consists of 10,791 observations.

Figure 3: Hierarchy of administrative areas in Birmingham

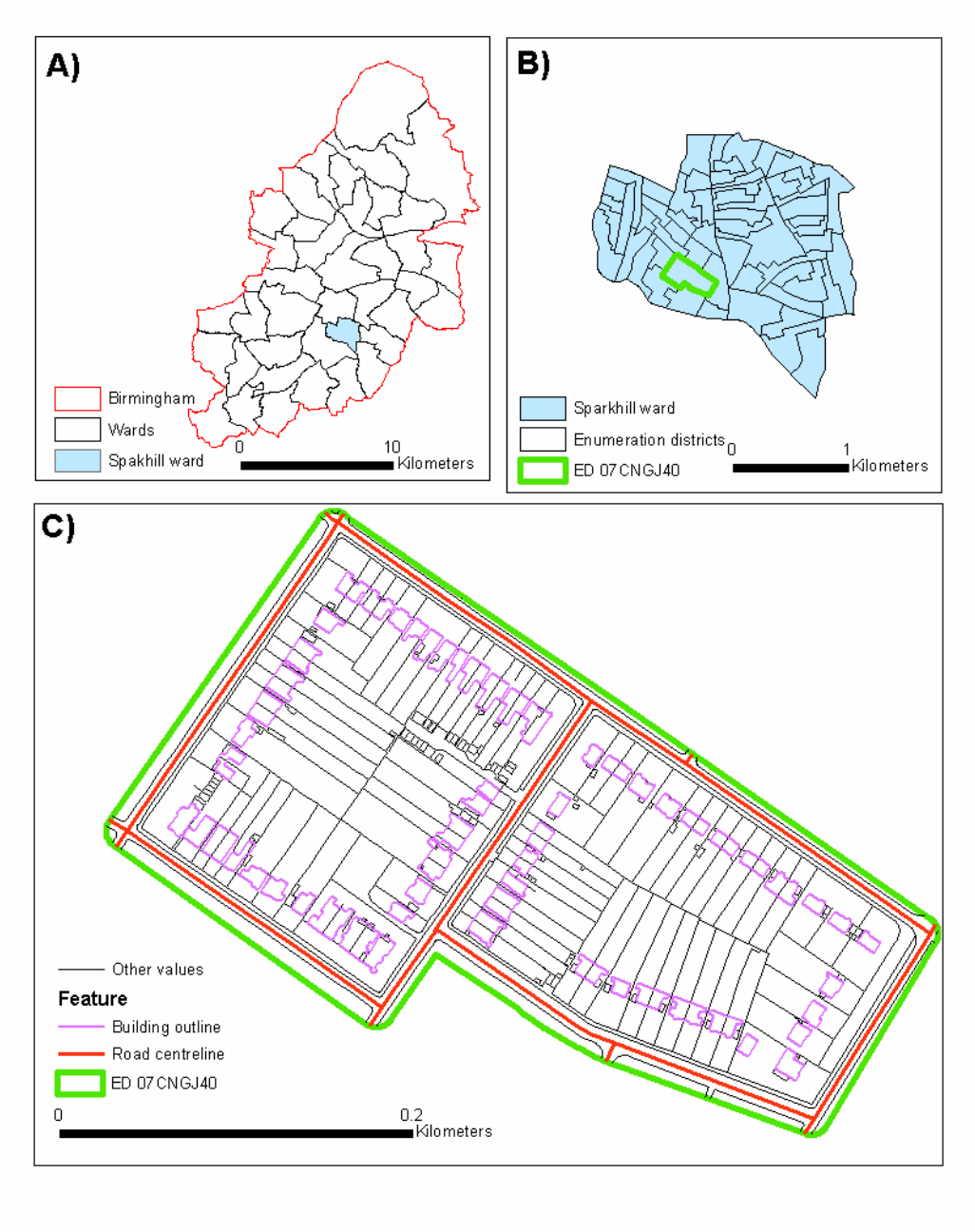


Table 1: Data Descriptions

Variable	Mean	Std. Dev.	Min	Max
Sale Price (£)	58,986	36,099	11,000	645,003
Structural Characteristics				
Floor Area (m^2)	102.6	32.7	42	645
Garden Area (m^2)	226.1	208	0	5,164
Garage (proportion)	0.436	0.496	0	1
Central Heating (proportion)	0.728	0.268	0	1
Age (decades)	6.1	2.76	0	11
WCs (proportion)				
One	0.794	0.404	0	1
Two	0.196	0.397	0	1
Three	0.009	0.094	0	1
> Three	0.001	0.029	0	1
Bedrooms (proportion)				
One	0.005	0.069	0	1
Two	0.172	0.377	0	1
Three	0.716	0.451	0	1
Four	0.083	0.276	0	1
Five	0.016	0.127	0	1
> Five	0.007	0.084	0	1
Storeys (proportion)				
One	0.021	0.145	0	1
Two	0.954	0.209	0	1
Three	0.021	0.143	0	1
> Three	0.003	0.058	0	1
Construction Type (proportion)				
Detached Bungalow	0.013	0.111	0	1
Semi-Detached Bungalow	0.008	0.090	0	1
End Terrace Bungalow	0.000	0.022	0	1
Terrace Bungalow	0.000	0.017	0	1
Detached House	0.116	0.320	0	1

Variable	Mean	Std. Dev.	Min	Max
Semi-Detached House	0.396	0.489	0	1
End Terrace House	0.115	0.319	0	1
Terrace House	0.352	0.478	0	1
Beacon Group (proportion)				
1. Unrenovated cottage pre 1919	0.000	0.019	0	1
2. Renovated cottage pre 1919	0.001	0.027	0	1
3. Small “industrial” pre 1919	0.040	0.195	0	1
4. Medium “industrial” pre 1919	0.226	0.418	0	1
5. Large terrace pre 1919	0.006	0.078	0	1
8. Small “villa” pre 1919	0.020	0.138	0	1
9. Large “villas” pre 1919	0.009	0.093	0	1
10. Large detached pre 1919	0.003	0.058	0	1
19. Houses 1908 to 1930	0.011	0.103	0	1
20. Subsidy houses 1920s & 30s	0.140	0.347	0	1
21. Standard houses 1919-45	0.257	0.437	0	1
24. Large houses 1919-45	0.016	0.124	0	1
25. Individual houses 1919-45	0.000	0.022	0	1
30. Standard houses 1945-53	0.045	0.207	0	1
31. Standard houses post 1953	0.190	0.392	0	1
32. Large houses post 1953	0.032	0.177	0	1
35. Individual houses post 1945	0.001	0.038	0	1
36. “Town Houses” post 1950	0.004	0.062	0	1
Sale Date (proportion)				
1st Quarter (Jan. to Mar.)	0.214	0.410	0	1
2nd Quarter (Apr. to June)	0.247	0.431	0	1
3rd Quarter (July to Sept.)	0.287	0.452	0	1
4th Quarter (Oct. to Dec.)	0.252	0.434	0	1
Neighbourhood Characteristics				
Poverty Factor	-0.375	0.855	-1.934	2.363
Sills Factor	0.180	1.000	-1.398	4.198
Age Factor	0.055	0.807	-3.216	3.143
Family Factor	-0.029	0.842	-3.198	3.791

Variable	Mean	Std. Dev.	Min	Max
Asian Factor	-0.045	0.942	-1.131	5.152
Black Factor	-0.240	0.750	-2.016	8.214
Locational Characteristics				
Proximity to City Centre (secs)	1,313	478	208	3,187
Proximity and Quantity of Shops	2.276	1.273	0.07	9.56
Proximity and Quality of Primary Schools	0.602	0.177	0.15	0.97
Walking time to Rail Station (secs)	1,846	1,013	21	5,525
Walking time to a Park (secs)	900	558	3	3,425
Driving time to Airport (secs)	2,388	655	602	4,386
Proximity to A-Type Industrial Processes (m)	2,463	1,821	21	10,204
Proximity to B-Type Industrial Processes (m)	814	528	10	3,333
Proximity to Land Fill sites (m)	947	608	10	3,472
Wards (proportion)				
Acock's Green	0.039	0.194	0	1
Aston	0.015	0.122	0	1
Bartley Green	0.018	0.131	0	1
Billesley	0.027	0.162	0	1
Bournville	0.038	0.191	0	1
Brandwood	0.022	0.147	0	1
Edgbaston	0.020	0.139	0	1
Erdington	0.029	0.168	0	1
Fox Hollies	0.028	0.165	0	1
Hall Green	0.041	0.198	0	1
Handsworth	0.016	0.125	0	1
Harborne	0.036	0.186	0	1
Hodge Hill	0.024	0.154	0	1
King's Norton	0.016	0.125	0	1
Kingsbury	0.010	0.101	0	1
Kingstanding	0.022	0.146	0	1
Ladywood	0.014	0.118	0	1

Variable	Mean	Std. Dev.	Min	Max
Longbridge	0.023	0.150	0	1
Moseley	0.024	0.152	0	1
Nechells	0.019	0.136	0	1
Northfield	0.028	0.164	0	1
Oscott	0.026	0.158	0	1
Perry Barr	0.033	0.180	0	1
Quinton	0.024	0.152	0	1
Sandwell	0.027	0.163	0	1
Selly Oak	0.044	0.205	0	1
Shard End	0.020	0.138	0	1
Sheldon	0.021	0.144	0	1
Small Heath	0.028	0.164	0	1
Soho	0.018	0.134	0	1
Sparkbrook	0.013	0.111	0	1
Sparkhill	0.020	0.142	0	1
Stockland Green	0.028	0.166	0	1
Sutton Four Oaks	0.038	0.191	0	1
Sutton New Hall	0.044	0.206	0	1
Sutton Vesey	0.039	0.194	0	1
Washwood Heath	0.028	0.164	0	1
Weoley	0.017	0.130	0	1
Yardley	0.024	0.154	0	1
Environmental Characteristics				
Views of Water (weighted m^2)	0.480	7.543	0	348
Views of Parkland (weighted m^2)	6.290	36.83	0	664
Road Traffic Noise (dB)	49.8	9.4	31.6	75.8
Rail Traffic Noise (dB)	36.8	12.6	0	74.7
Aircraft Noise (dB)	4.8	16.0	0	69

5. Application of cluster analysis strategy to the City of Birmingham dataset

The theoretical discussion in the introduction describes the predictions of two models of property markets. In the first model, typified by the work of Ekeland et al. (2003), households choose a property based on the attributes of properties themselves. The model predicts that for certain parameter values, the market will be characterised by distinct clusters of properties exhibiting similar levels of attributes.

In contrast, Nesheim (2002) presents a model in which households choose properties based on the characteristics of the other households in the neighbourhood. Nesheim's model predicts that households will sort themselves spatially according to socioeconomic characteristics. Again, the model suggests that for certain parameter values the market will be characterised by the existence of distinct clusters of (not necessarily spatially contiguous) similar neighbourhoods; that is, neighbourhoods that are composed of a set of households exhibiting similar socioeconomic profiles.

In this section we attempt to identify clustering of these two different forms by applying model-based clustering techniques to the City of Birmingham dataset.

5.1 Choice of clustering variables

The first step in the cluster analysis is to choose the set of P variables defining the characteristics of each observation.

One exceedingly practical consideration in making this choice is that model-based clustering, as applied here, requires the clustering variables to be continuous. Examination of the data descriptions in Table 1 reveals that the majority of variables detailing the structural characteristics of properties are discrete. That is, they are binary variables indicating the presence or absence of a particular feature. Indeed, once the discrete variables have been eliminated, we are left with only three candidates; floor area, garden area, and property age. Fortunately, between them these three variables capture a substantial proportion of the variability in the structural characteristics of properties. Witness the fact that a simple linear regression of log sales price on log floor area, log garden area and property age, returns an R^2 statistic of .52. That is, on their own, these three variables explain some 52% of the variation in property prices. Adding in the remaining 46 discrete structural covariates only elevates the R^2 statistic to 0.63. In terms of property prices, therefore, the three variables of floor area, garden area, and property age capture the majority of variability that can be accounted for by the structural attributes of properties. We contend that this is because the structural features of properties are highly correlated (e.g. number

of beds, WCs and storeys with a property's floor space; beacon group with a property's age and floor space; property type with a property's floor space and garden size) such that combinations of these three variables provide a reasonably precise description of the different structural types available in the property market.

In contrast, in defining the socioeconomic characteristics of neighbourhoods we are faced with a surfeit of candidate variables. The census data provides literally hundreds of variables describing the socioeconomic characteristics of the households inhabiting each enumerator district. As a result, we adopt a simple two-step procedure that condenses the excess of neighbourhood attributes into a more manageable set of indices or factors. In the first step, variables from the census data are grouped into five categories. These categories are as follows; variables describing the age composition of inhabitants of an ED, variables describing the family composition of households in an ED, variables describing the wealth of households in an ED, variables describing the ethnicity of inhabitants of an ED, and variables describing the education and employment of inhabitants of an ED. In the second step, the variables in each category are subjected to a factor analysis. A summary of the factor analysis is provided in Table 2. Following standard practice, for each group of variables, only factors with eigenvalues greater than one are retained. In all but one case, this results in the retention of only one factor for each category. As can be surmised from the third column of Table 2, on the whole, the retained factors capture the greater portion of the variability in the variables included in each category. The factors are rotated to aid interpretation and those variables with loadings greater than |0.50| are listed in the final column of Table 2. The loadings suggest meaningful interpretations for the dimensions captured by each factor. These interpretations are summarised in the first column of Table 2. The final step is to define a score for each ED for each factor. In effect, EDs that exhibit high values for attributes that load positively on a factor receive high scores for that factor whilst neighbourhoods that exhibit high values for attributes that load negatively on that factor receive low scores.

The six factor scores are used as summary variables describing the major features of the socioeconomic characteristics of property neighbourhoods for use in the model-based clustering. Again the six factors describing the socioeconomic are found to be major determinants of property prices. A simple linear regression of log property price against the six factors returns and R^2 statistic of 0.57.

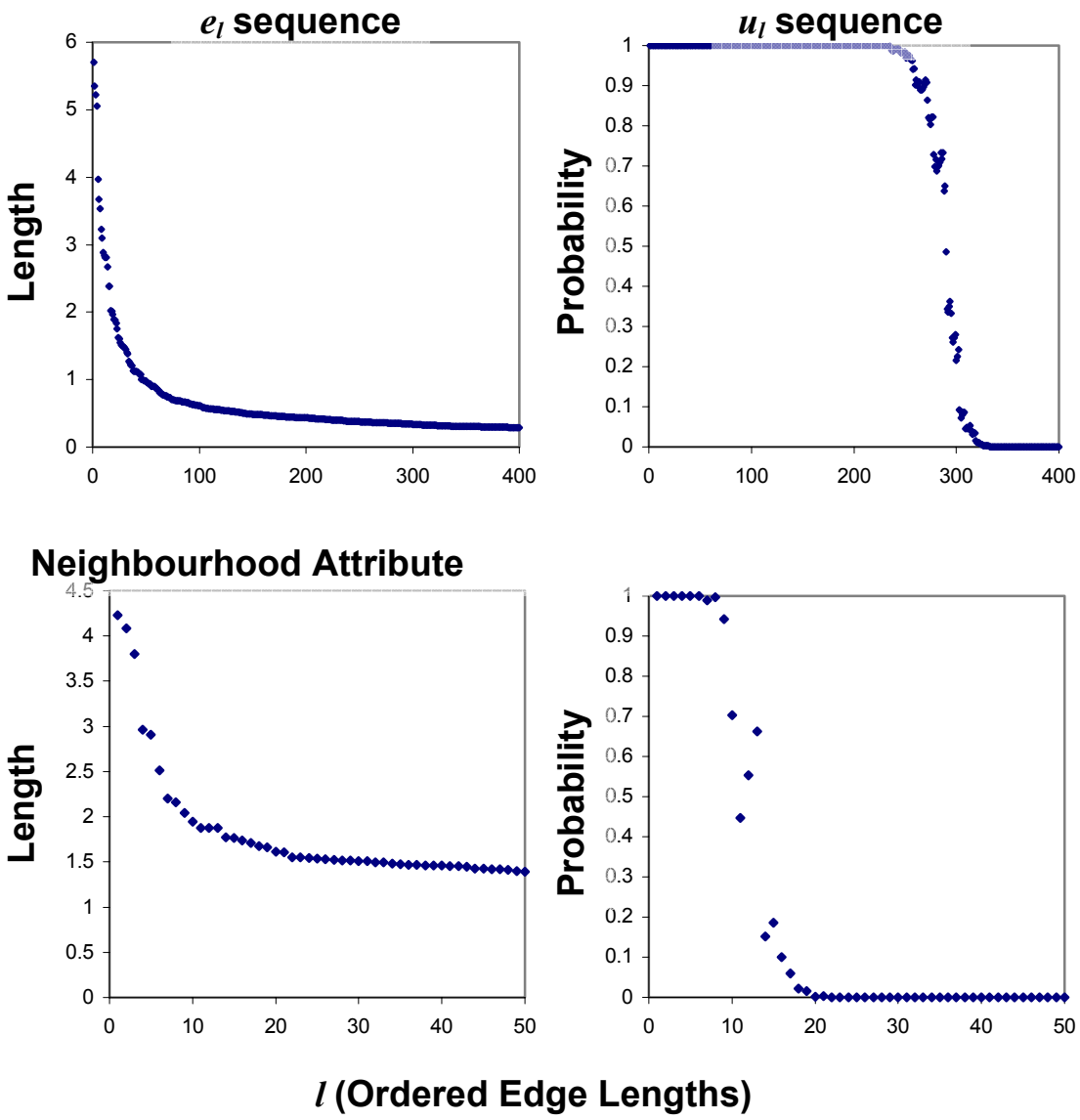
Table 2: Factor analysis of census data describing the socioeconomic characteristics of enumerator districts

Factor Name & Description	Eigenvalues (>1)	Percent Variance Explained	Variable Loadings (>|0.50|)	
I. Household Age Composition (Using 5 variables):				
a. AGE FACTOR: Increasing Age of Inhabitants	1.76	61	% Age 18-24	-0.72
			% Age 25-34	-0.64
			% Age 50-64	0.59
			% Age > 65	0.63
II. Family Composition (Using 4 variables):				
b. FAMILY FACTOR: Increasing Proportion of Households with Children	2.66	81	% Young Family	0.86
			% Old Family	0.82
			% Age 0-10	0.78
			% Age 10-17	0.80
III. Wealth of Households (Using 4 variables):				
c. POVERTY FACTOR: Increasing Poverty of Households	3.18	97	% No car	1.00
			% Two cars	-0.85
			% Unemployed	0.85
			% Local Authority Housing	0.85
IV. Ethnicity (Using 6 variables):				
d. ASIAN FACTOR: Increasing Proportion of Asians Households	3.00	56	% Pakistani	0.96
			% Bangladeshi	0.67
			% White	-0.75
e. BLACK FACTOR: Increasing Proportion of Black Households	1.13	21	% Caribbean	0.89
			% African	0.77
V. Education and Employment (Using 15 variables):				
f. SKILLS FACTOR: Increasingly Skilled Households	2.97	34	% professional	0.61
			% diploma	0.73
			% degree	0.83

5.2 MST initial partition

The data contains 10,791 property observations. Using the three clustering variables of floor area, garden area and property age, the MST for the property attributes data was constructed using Mahalanobis distances. The top two graphs in Figure 4 plot the first 400 elements of the e_l-sequence and u_l-sequence calculated from the edge lengths of this MST. Likewise, there are 1,940 EDs in the City of Birmingham. Again the MST for the six factors describing the socioeconomic composition of inhabitants of these EDs has been constructed using Mahalanobis distances. The e_l-sequence and u_l-sequence for the edge lengths of this MST are reproduced in the lower two plots in Figure 4.

Figure 4: The e_l-sequence and u_l-sequence of the Minimal Spanning Tree



The e_l-sequence plots for both MSTs show the expected pattern. A few observations are well-separated from the others and have comparatively large edge lengths in the MST but the rate of change in the sequence of ordered edge lengths declines rapidly.

In both cases the u_l-sequence plots show clear evidence of clustering. If the data had been drawn from a single standardised Gaussian population then we would expect to see no discernible pattern in the u_l-sequence. Rather, in both cases we see that the longer edge lengths of the MST exceed the lengths that might be expected through chance whilst the shorter edge lengths of the MST are somewhat shorter than might be expected.

To determine which edges in the MST are considered sufficiently long to warrant "peeling", Posse (2001) proposes identifying the edge length at which the u_l-sequence stabilises around 0 and at which the rate of decay in the e_l-sequence has dropped significantly. Since edges longer than this length separate observations that are more distant from each other than might be expected, Posse suggests an initial partitioning of the MST that breaks these edges.

From Figure 4 it is clear that the rate of decline of the e_l-sequence reduces significantly after the first 75 to 125 longest edge lengths in the case of the property attribute MST and after the first 20 to 25 longest edge lengths in the case of the neighbourhood attributes MST. Likewise, the u_l-sequence stabilises around zero shortly after the 350^{th} longest edge length for the property attribute MST and after the 20^{th} longest edge length for the neighbourhood attribute MST. Following Posse's (2001) proposition, therefore, we choose to peel the first 350 longest edges of the property attribute MST and the first 30 longest edges of the neighbourhood attribute MST.

As detailed in Table 3, we subsequently "prune" the MSTs so as to form a large number of roughly equal sized clusters. In the case of the property attributes data, the average number of observations in a cluster following pruning is 3.48. Similarly the average cluster size for the neighbourhood attribute is 3.41.

Furthermore, the Posse procedure isolates 765 property observations and 161 neighbourhood observations into clusters of their own. Since these singleton clusters are likely to be well-separated from other observations they are taken as an initial indication of observations that do not belong to any cluster but are part of the noise.

5.3 Model based clustering of properties with geographical smoothing

Clusters derived from the MSTs are used to initialise the model-based clustering algorithms. For both data sets, a variety of models corresponding to different numbers of clusters and different cross-cluster restrictions on the cluster covariance matrices have been estimated. BIC values for a selection of these

models are presented in Figure 5. The three covariance models described in the figure performed significantly better than other possible parameterisations. Indeed, no other model estimated returned BIC scores that would register on these graphs.[10]

Table 3: Initial Partition of the data set using the MST procedure suggested by Posse

	Property Attribute Clustering	Neighbourhood Attribute Clustering
Num. Obs	10,791	1,940
Num. Peel	350	30
Num. Prune	2,500	400
Num. Clusters	3,100	569
Num. Singletons	765	161
Avg. Obs. per Cluster	3.48	3.41
Max. Obs. per Cluster	6	6

The BIC scores for the neighbourhood attribute clustering reveal the unconstrained model, in which different clusters may differ in size, shape and orientation, outperforms the other models. The BIC reaches a maximum at a model containing seven clusters and following Fraley and Raftery (1998) this model is selected as the one best describing the patterns of clustering in the data.

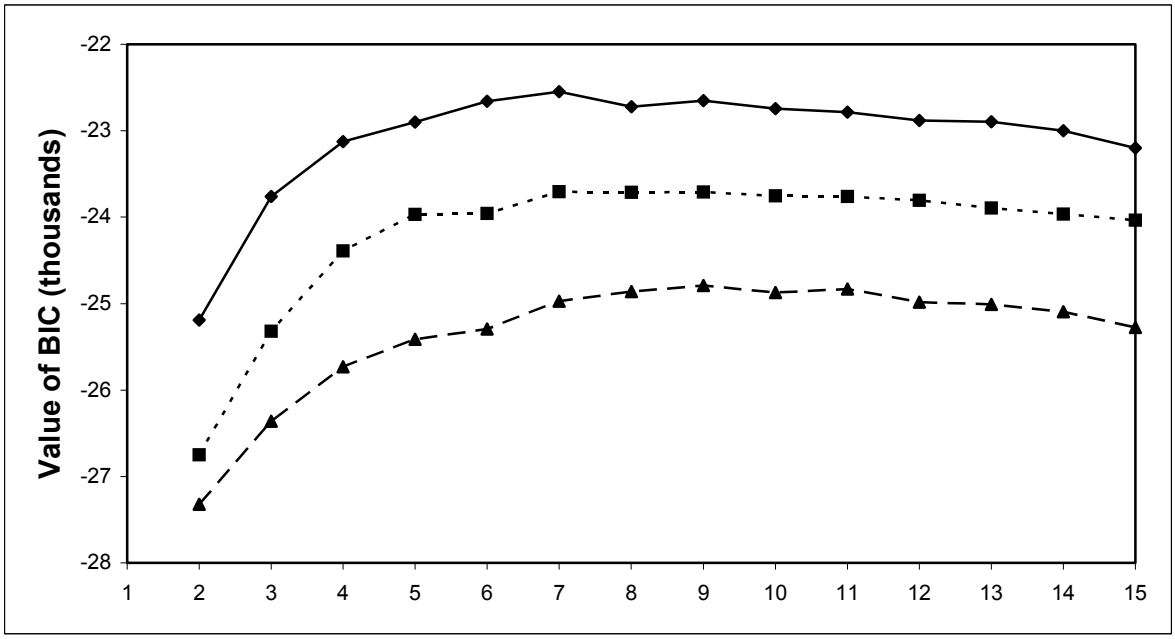
For the property attribute data the picture is less clear. A model in which the shape of each cluster is constrained to be equal performs only marginally less well than the unconstrained model. Also, there is no single maximum for the BIC scores. Rather, the BIC scores for models with progressively larger numbers of clusters tend to increase but a progressively slower rate. This pattern is not uncommon in large data sets where the BIC tends to prefer partitions with many clusters (Posse, 2001). Here we follow the suggestion of Banfield and Raftery (1993) taken up by Posse (2001) and choose the 6 cluster unconstrained model as this gives a particularly high value for the BIC in the region where the rate of change of the BIC drops significantly.

[10] This observation indicates that traditional approaches to clustering such as Ward's (1963) method may be inappropriate since this method is equivalent to restricting the covariance matrices to be spherical but with differing volumes (Fraley and Raftery, 1998).

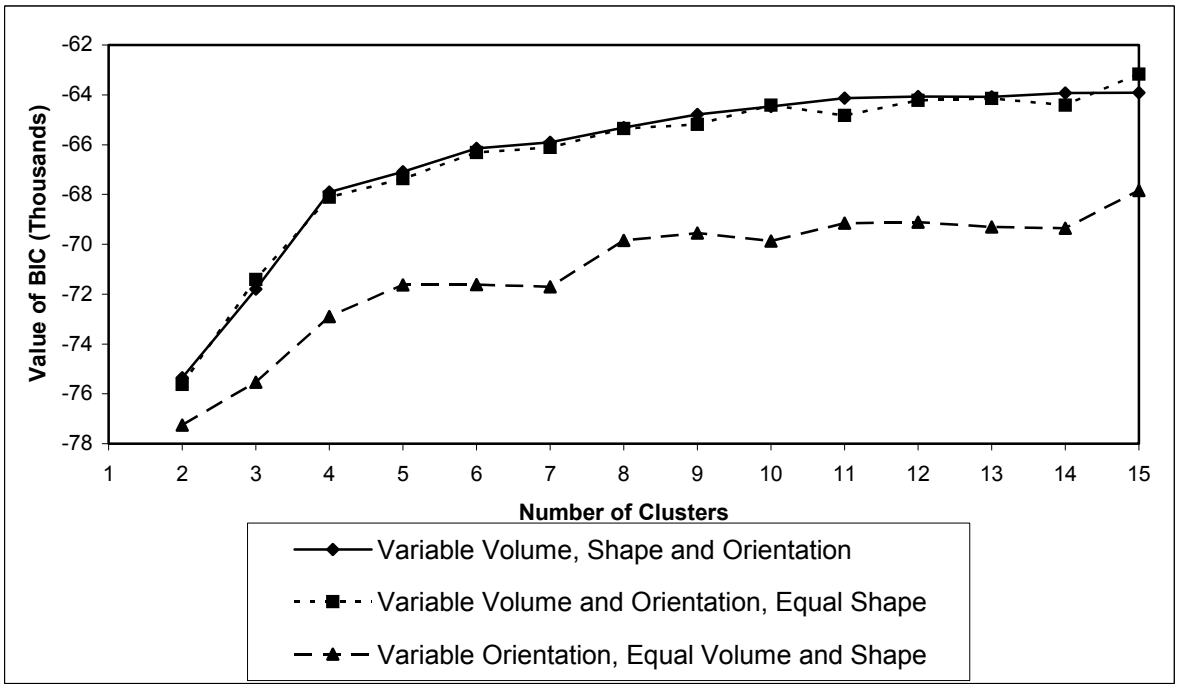
Finally, the spatial smoothing algorithm was applied to the two clustering solutions. In the case of the neighbourhood attribute partitioning the classification stabilised after 4 iterations, with some 154 EDs having changed classification. In the case the property attribute partitioning the classification stabilised after 3 iterations once 642 properties had changed classification.

Figure 5: BIC scores for clustering models assuming different numbers of clusters and different parameterisations of the covariance matrices

Neighbourhood Attribute Clustering:



Property Attribute Clustering:



Tables 4 and 5 present summary statistics that report the number of observations and the means of selected variables for each cluster. Figures 6 and 7 plot the locations of the properties in the different clusters for the two partitionings of the data.

In general the neighbourhood clusters are readily interpretable. Clusters 1, 3, 4 and 5 pick out neighbourhoods that are populated, in the main, by ethnically white inhabitants. Of these Cluster 1 identifies relatively poor neighbourhoods, with low-skilled inhabitants. These neighbourhoods tend to be located to the south and west of the city but not in the city centre nor in the relatively affluent north-eastern suburbs. Cluster 4 comprises middle income neighbourhoods that are averagely skilled and relatively old. Clusters 3 and 5 pick out wealthy neighbourhoods with highly skilled inhabitants. These neighbourhoods tend to be in suburban locations with especially high concentrations in the desirable north-eastern region of the city.

In contrast, Clusters 6 and 7 define neighbourhoods whose inhabitants come mainly from the ethnic minorities. Whilst these neighbourhoods share the same inner city locations and are characterised by relative poverty and low-skilled inhabitants, they remain ethnically distinct. Cluster 6 defines neighbourhoods that are majority black, Cluster 7 neighbourhoods that are majority Asian. Perhaps unsurprisingly, average adult ages in these neighbourhoods are relatively low whilst, especially in the Asian neighbourhoods, there are a relatively large number of households with children.

Cluster 2 is somewhat more difficult to interpret. The population is ethnically diverse and comprised almost exclusively of young adults without children. Whilst the inhabitants of these neighbourhoods are relatively skilled they are only moderately wealthy. We surmise that these neighbourhoods are those inhabited by young professionals. The geographic distribution of properties in this Cluster accords with this interpretation. Neighbourhoods in Cluster 2 are located outside the inner city, but within easy commuting distance of the city centre. Further, a particularly large concentration of neighbourhoods in this cluster can be found to the south and west of the city centre, located around the University and Hospital complex.

Finally only a very few neighbourhoods cannot be assigned to one of the clusters and fall into the noise category.

The clusters identified by partitioning according to the age, floor space and garden size of properties are also readily interpretable.

Cluster 1 picks out modern developments. Indeed, 87% of properties in this cluster fall into Beacon Group (BG) 31 defined as standard houses built post 1953. These properties tend to be provided with gardens and cover a range of sizes and construction designs; some detached, some semi-detached and some terraced. Notice from Figure 6 that properties in this cluster are widely

dispersed over the cityscape reflecting recent planning trends that have encouraged infilling rather than expansion of the urban area.

Table 4: Summary of neighbourhood attribute clusters reporting the number of EDs in each cluster and the mean values for the clustering variables

Cluster	Num	Poor	Skill	Age	Family	Black	Muslim
Cluster 1	424	0.500	-0.623	0.252	0.014	-0.290	-0.409
Cluster 2	255	0.172	0.586	-0.804	-0.907	0.442	-0.297
Cluster 3	328	-1.081	0.437	0.470	-0.370	-0.602	-0.399
Cluster 4	148	0.226	-0.227	0.631	-0.259	-0.405	-0.349
Cluster 5	309	-0.990	0.875	0.392	-0.274	-0.366	-0.374
Cluster 6	256	0.915	-0.631	-0.458	0.550	1.470	0.023
Cluster 7	214	0.646	-0.538	-0.668	1.538	-0.047	2.533
Noise	6	-0.197	1.923	-1.706	-0.109	2.348	-0.142

At the other extreme Cluster 2 is comprised almost exclusively of small turn-of the-century terraces with relatively small associated plots of land. 93% of these properties are classfied as BG 3 or 4, that is small or medium "industrial" properties built before 1919. In accordance with the historical development of the city, these properties encircle the city centre.

Similarly, Cluster 3 identifies turn-of the-century properties located in a similar geographic region to those in Cluster 2. However, unlike Cluster 2 these are not exclusively small terraces. In fact, the properties in Cluster 3 are larger with more bedrooms and much larger gardens. Cluster 3 comprises properties constructed for the more affluent members of turn-of-the-century Birmingham society; properties that estate agents like to call "town houses" or "villas".

Clusters 4 and 5 identify standard mostly terraced or semi-detached properties with gardens. Notice in Figure 7 that the properties in Cluster 5 fall in a broad swathe that encircles the inner city. Indeed, these properties are part of the rapid expansion of Birmingham that took place between the wars. 97.5% of properties in this cluster are classified as BG 20 or 21, standard (frequently state-subsidised) properties constructed in the 1920s and 1930s. Geographically, properties in Cluster 4 appear to comprise a final ring of development surrounding the properties built between the wars. Indeed, these properties comprise standard, post-war properties. Some 70% of properties in Cluster 5 are classified as standard houses constructed between 1945-53 (BG 30) or post 1953 (BG 31).

Cluster 6 isolates the large properties in Birmingham. 86% of the properties in this cluster are detached or semidetached. They tend to have large gardens and are located in mainly suburban area with a large concentration in the desirable north-eastern region of the city.

From the descriptive statistics it would appear that many of the 67 properties allocated to the noise are the extremely large properties. The properties in the noise are mostly detached and, on average, have the most bedrooms, floor space and garden area of any of the clusters. It appears that the clustering procedure has isolated this small number of seeming outliers from the larger groupings of more moderately proportioned properties in the data.

Table 5: Summary of property attribute clusters reporting the number of EDs in each cluster and the mean values for the clustering variables and other variates

Cluster	Num	Area	Garden	Age	Beds	%Terrace	%Detached	Price
Cluster 1	1,540	91.7	162.4	19.5	2.85	0.42	0.24	61,749
Cluster 2	2,324	95.2	84.1	93.7	2.7	0.95	0.00	38,916
Cluster 3	878	142	195.8	95	3.36	0.65	0.03	63,365
Cluster 4	1,176	97.5	266.8	49.5	2.95	0.31	0.06	57,064
Cluster 5	3,453	87.5	205.8	66	2.85	0.34	0.03	48,530
Cluster 6	1,353	136.8	506.7	57.3	3.46	0.06	0.46	107,734
Noise	67	276.3	1694.1	66.4	5.1	0.04	0.82	243,415

Figure 6: Geographical distribution of properties in clusters defined by neighbourhood soicioeconomics partitioning

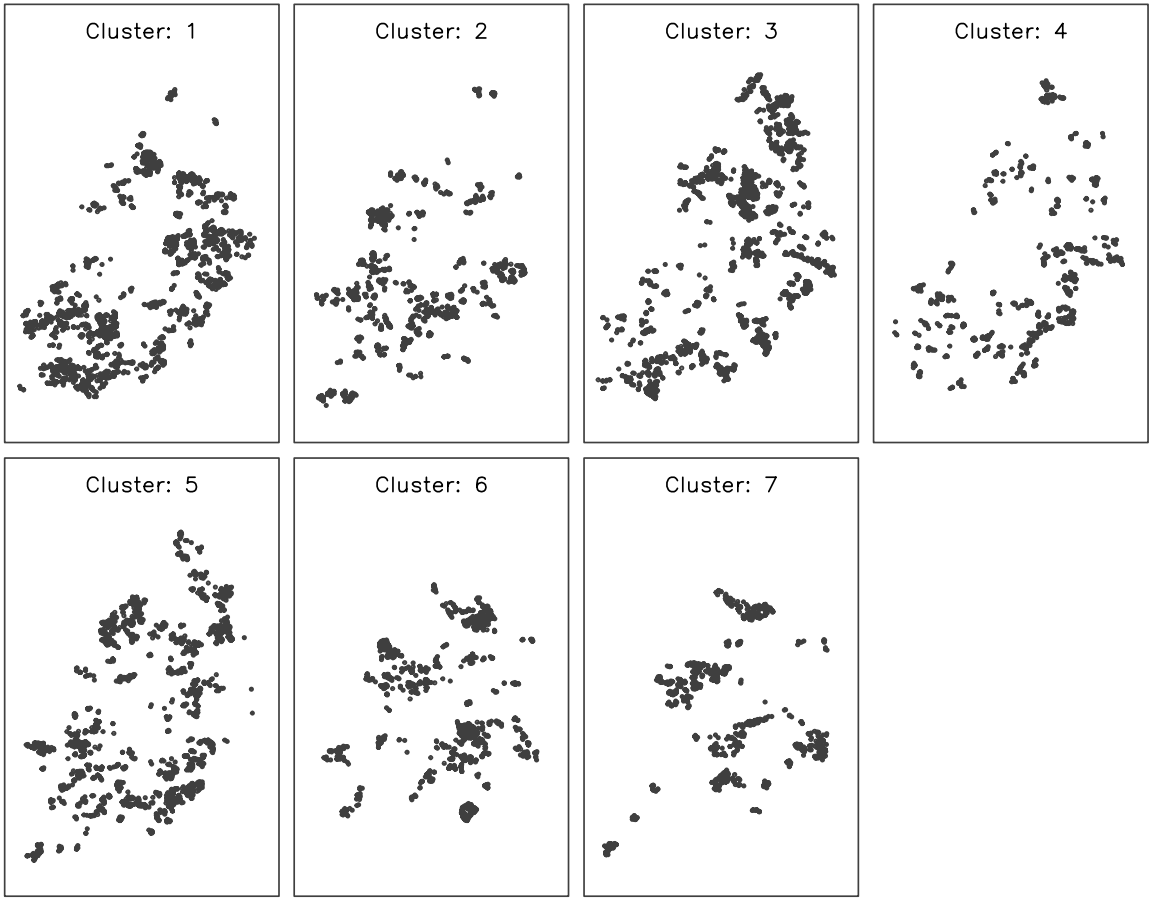
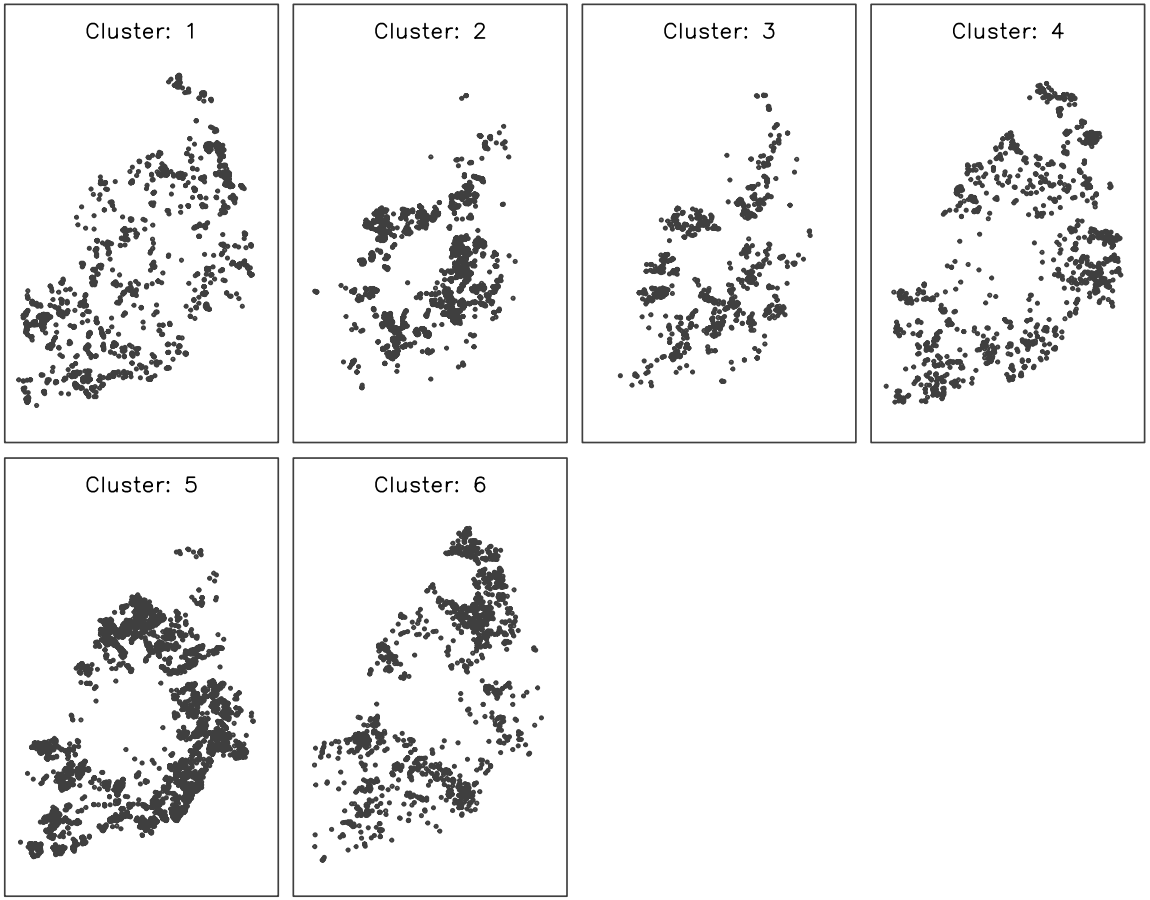


Figure 7: Geographical distribution of properties in clusters defined by property attributes partitioning



6. Estimation of Hedonic Price Functions by Cluster

The theoretical models described in the introduction both predict that the hedonic price surface may be highly non-linear. As such, following the standard procedure and fitting a simple linear regression usually with log-transformed price as the dependent variable is unlikely to provide anything but a poor approximation to the true hedonic price function. A number of alternative estimation strategies suggest themselves.

Foremost amongst these strategies is to adopt more flexible functional forms. There is a long established literature pursuing this line of reasoning. A number of researchers have investigated the use of parametric specifications such as the Box-Cox flexible functional form (e.g. Milon, Gressel and Mulkey, 1984; Blackley, Follain and Ondrich, 1984; Cassel and Mendelsohn, 1985; Cropper, Deck and McConnell, 1988; Gençay and Yang, 1996; Huh and Kwak, 1997; Cheshire and Sheppard, 1998) though as discussed by Ramussen and Zuehlke (1990) there are some theoretical difficulties with this approach.

Even more flexible semiparametric approaches have been employed by a number of authors. Anglin and Gençay (1996) and Gençay and Yang (1996), for example, use a partially linear model to allow a subset of the variables to be included into the model specification in nonparametric form.

In the extreme, some researchers have opted to estimate the whole hedonic price function by nonparametric regression (e.g. Pace 1993, 1995). The use of nonparametric regression allows the data to dictate the nature of the relationship between property characteristics and price. Unfortunately, it is evident that a large number of factors affect property prices and, as such, the approach will likely fall foul of the well-known curse of dimensionality.

Rather than employing increasingly more general econometric specifications to capture the nonlinearity of the equilibrium hedonic price function, our estimation strategy is to avoid estimating the hedonic price function over the entire attribute space. Rather, we fit separate price functions for the properties in each cluster thereby forming local approximations to the hedonic price surface over the attribute area spanned by the properties in each cluster.

For each of the two partitionings of the data we adopt the following set of simple linear regression functions;

$$\ln(\boldsymbol{P}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}_j \qquad j = 1, 2, ..., M \tag{12}$$

where j indexes clusters, $\boldsymbol{P}_j$ is the $N_j \times 1$ vector of property prices for data allocated to cluster j, $\boldsymbol{X}_j$ is the associated $N_j \times K_j$ regressor matrix, $\boldsymbol{\beta}_j$ is the $K_j \times 1$ vector of parameters and $\boldsymbol{e}_j$ is the $N_j \times 1$ vector of residuals that we

assume to have $E\left[\boldsymbol{e}_j\right]=\mathbf{0}$ and $E\left[\boldsymbol{e}_j\boldsymbol{e}_j'\right]=\sigma_j^2\boldsymbol{I}_{N_j}$. We estimate the models using ordinary least squares (OLS).

6.1 A discussion of the parameter estimates

A selection of parameter results from these two sets of linear regressions are provided in Tables 5 and 6. Full details can be found in appendices A and B at the end of this paper. For want of space, we do not discuss all the results but highlight some of the more interesting findings.

First, let us examine the partitioning based on the socioeconomic characteristics of neighbourhoods (Table 5). All in all, the parameters estimated for the structural characteristics of properties exhibit similar patterns for all seven clusters. Not surprisingly the two structural attributes describing the overall dimensions of properties, floor area and garden area, are highly significant in all clusters; the bigger the property the more it sells for, all else equal. Furthermore, in clusters where the presence of a garage and/or central-heating makes a statistically meaningful difference, it is always to make those properties more valuable.

For all clusters the parameter estimated on the age variable is negative, though it is only statistically significant at over a 90% level of confidence in two of the seven clusters. Accordingly, all else equal, properties lose market value with age. Of course, a fuller appreciation of differences in property values brought about by construction date would have to consider the parameters estimated on the eighteen Beacon Group dummy variables (to be found in the appendices) since these also isolate important aspects of a properties age and design.

The set of dummy variables indicating the number of bedrooms possessed by a property shows a similar pattern across all clusters. Compared to the baseline case of a three-bedroom house, properties boasting more or fewer bedrooms tend to command higher prices in the market. The most statistically significant premium is for five-bedroomed properties. Of course, this is under the important caveat that all else, including floor area, is held equal.

With regards to the number of storeys over which a property is divided, there appears little to distinguish properties with one storey from the baseline case of a two-storied property. Again, to see the full picture one would need to consider the full set of dummy variables for construction type (to be found in the appendices) which include three variables indicating types of bungalow. In nearly all cases, and in all cases that make a statistically meaningful difference, properties with three or four stories command lower market prices than a two-storied property. It appears that the market values short, fat properties more highly than tall, skinny ones. In a similar vein, the dummy variables on construction type shown in Table 5 reveal that in all clusters, detached

properties are valued more highly than semi-detached properties which are in turn valued more highly than terraced properties.

In all seven clusters, there is clear evidence of prices changing over the course of the study year (1997). Prices appear to have risen between 3% and 8% (depending on cluster) between the first and third quarter of the year, remaining stable in the final quarter.

In contrast to the structural characteristics, the influence of neighbourhood characteristics on property prices displays a number of interesting contrasts across clusters. As might be expected, property prices are depressed if the area in which the property is located is relatively poor, are inflated if the neighbourhood is inhabited by relatively highly skilled households. Perhaps not so predictably but also showing a consistent pattern across clusters we find that property prices tend to be higher in neighbourhoods with relatively older inhabitants but lower when the neighbourhood has a proportionately larger population of households with children.

In contrast, observe the parameter estimates on the Asian and Black factors. In the first five clusters, clusters whose populations appear to be majority white, neighbourhoods with larger Black or Asian contingents are characterised by lower priced properties. However, a different pattern emerges for Cluster 6, the cluster isolating neighbourhoods with mainly Black communities. Here the parameter estimates on the Asian and Black factors take the opposite sign; properties in neighbourhoods containing proportionately more Black or Asian households command significantly greater market prices. A similar pattern can be seen in Cluster 7, the cluster isolating majority Asian neighbourhoods. In this cluster, properties in neighbourhoods with proportionately more Asian households are significantly more expensive. Without wishing to over-elaborate the significance of this result, the implication is that within clusters properties in racially homogeneous neighbourhoods tend to be more highly valued than those in ethnically diverse neighbourhoods.

Consider now the locational characteristics of properties with respect to their proximity to amenities and disamenities. The parameters on the proximity to the city centre present a somewhat confused pattern, being negative and significant for some clusters, positive and significant for others. For example, proximity to the city centre deflates property prices in clusters 1 and 6 (the poor ethnically white and ethnically black clusters respectively) whilst inflating prices in clusters 3 and 7 (the wealthy and Asian clusters respectively). Since, proximity to the city centre does not induce a coherent influence on property prices across all clusters it seems likely that this variable is proxying for other features of the urban geography that are not captured by the model.

The patterns displayed by the parameters on the shops variable, which provides an indication of the size and proximity of local commercial centres, are again

somewhat complex. The model indicates that in clusters 1 and 6 the (the poor ethnically white and ethnically black clusters respectively) property prices increase with proximity to shops though in clusters 3 and 7 (the wealthy and Asian clusters respectively) prices are reduced by proximity to shops. A possible, though not entirely coherent, explanation of these results is that within the less affluent socioeconomic clusters, proximity to shops is considered an advantage whilst amongst more affluent suburban groups differing shopping habits reduce the attractiveness of such convenience.

The variable for primary schools combines distance and school quality into a single index. High scores indicate increasing quality and/or ease of access. The results here corroborate anecdotal evidence and that of recent studies (for example, Gibbons and Machin, 2003) suggesting that increasing primary school quality and proximity inflates property prices. Whilst the parameters for all clusters are positive, those in the ethnic minority socioeconomic clusters (clusters 6 and 7) are not significant.

Some fairly general patterns emerge with regards to the other locational variables. Proximity to railway stations and parks tend to have little influence on property prices. In the clusters where these locational characteristics make a difference, they act so as to decrease property prices with increasing proximity. Whilst, these locational features could nominally be considered as amenities, it appears that other issues, perhaps including security and noisy activity, may detract from the benefits of proximity to either a railway station or park. Where significant, proximity to Type-A industrial processes and proximity to landfill sites act so as to reduce property prices. In contrast, proximity to the airport and proximity to Type-B industrial processes act so as to increase prices.

In accordance with prior expectations all the parameter estimates on road and rail noise pollution are negative and in the majority of cases are statistically significant. A similar pattern emerges for estimates of the parameter on the aircraft noise pollution variable, though here only one parameter estimate is statistically significant and another is positive (though not significant at a 90% level of confidence). Unfortunately, air traffic noise is considerably less localised than that arising from either road or rail traffic. Indeed properties over a large area will experience very similar levels of air traffic noise. A short-coming of the modelling approach adopted in this research is that much of the influence of these wide-area spatial effects will be subsumed into the locational constants indicating ward membership (parameter estimates for these ward constants can be found in the appendices).

Parameter estimates for the model based on partitioning the data according to the attributes of the properties are displayed in Table 6. Conclusions concerning the impact of structural attributes on property prices for this partitioning of the

data are broadly similar to those for the partitioning based on the socioeconomic composition of neighbourhoods.

Table 6: A selection of parameters from hedonic price equations for clusters defined by partitioning according to the socioeconomics of neighbourhoods

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Constant	8.9067***	8.2327***	8.7047***	8.4688***	8.3764***	8.4239***	8.9403***
Structural Characteristics:							
Floor Area (log)	0.3827***	0.3991***	0.4383***	0.3612***	0.4879***	0.3864***	0.3670***
Garden Area (log)	0.0838***	0.1662***	0.0973***	0.1005***	0.0940***	0.1393***	0.1446***
Garage	0.0448***	0.0579***	0.0550***	0.0350*	0.0524***	0.0607***	0.0369
Central Heating	0.0464*	0.0653*	0.0577**	-0.0283	0.1032***	0.0828**	-0.0716
Age	-0.0148**	-0.0067	-0.0096	-0.0058	-0.0204***	-0.0091	-0.0106
WCs							
One	*b*	*b*	*b*	*b*	*b*	*b*	*b*
Two	0.0243*	-0.0399**	0.0315**	-0.0059	0.0297**	-0.022	-0.0244
Three	0.0198	0.2056**	-0.0112	-0.022	0.1304***	0.0222	0.0075
Four	.	0.8627***	-0.2295*	.	0.4666**	.	.
Bedrooms							
One	0.0727	0.0675	0.0414	0.2473**	0.0351	0.3394	0.1957
Two	0.007	-0.0013	0.0127	-0.0299	0.0152	-0.0062	0.0560**
Three	*b*	*b*	*b*	*b*	*b*	*b*	*b*
Four	0.0278	0.0029	0.0165	0.0279	0.0407*	0.0677**	0.047
Five	0.0452	-0.0609	0.1349***	0.1474**	0.1459***	0.1758***	0.044
Storeys							
One	-0.07	-0.4751	-0.037	0.0449	0.1903***	0.2255	0.1281
Two	*b*	*b*	*b*	*b*	*b*	*b*	*b*
Three	-0.0481	-0.2195***	-0.1069***	-0.0672	-0.1115***	-0.0166	0.0183
Four	-0.2106*	-0.8875***	-0.4576***	-0.1522	-0.1909*	-0.1956	-0.4995**
Construction Type							

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Detached	0.1396***	0.1477***	0.1220***	0.1386***	0.1087***	0.0721***	0.0884**
Semi-Detchd	*b*	*b*	*b*	*b*	*b*	*b*	*b*
End Terrace	-0.0887***	-0.0981***	-0.0440**	-0.0493*	-0.0780***	-0.0309	-0.1012***
Terrace	-0.0795***	-0.0418*	-0.0647***	-0.0407*	-0.0917***	-0.0763***	-0.0833***
Sale Date							
1st Quarter	-0.0508***	-0.0313*	-0.0564***	-0.0400**	-0.0407***	-0.0716***	-0.0675***
2nd Quarter	-0.0231*	-0.017	-0.0224*	-0.0495***	-0.009	0.0246	-0.0262
3rd Quarter	*b*	*b*	*b*	*b*	*b*	*b*	*b*
4th Quarter	-0.0039	-0.0094	0.0023	-0.0171	0.015	0.0269	-0.0101
Neighbourhood Characteristics							
Poverty	-0.0871***	-0.0736***	-0.0648***	-0.1284***	-0.0471***	-0.1260***	-0.0483**
Skills	0.0628***	0.0305***	0.0524***	0.0401**	0.0662***	0.0055	0.0404**
Age	0.0209***	0.0303**	0.014	0.0264*	0.0098	0.0587***	0.0404**
Family	-0.0075	-0.0489***	-0.0205*	-0.0255	-0.0106	-0.0248	-0.0506***
Asian	0.0137	-0.0482***	-0.0368***	-0.0697**	-0.0118	0.0314**	0.0438**
Black	-0.0254**	0.0046	-0.0517***	-0.0314	-0.0520***	0.0269**	-0.011
Locational Characteristics							
City Centre	0.0001**	-0.0001	-0.0001**	-0.0001	0	0.0001*	-0.0001*
Shops	0.0230***	-0.0134	-0.0347***	-0.0023	0.0126	0.0271**	-0.0363***
Primary Schools	0.0961**	0.1593***	0.1089***	0.1766***	0.0942**	0.0134	0.0404
Rail Station	0	0	0	0.0001***	0.0000**	0	0
Park	0	0	0.0000**	0	0	0	0
Airport	-0.0001***	-0.0001	0	0	-0.0001***	-0.0001**	-0.0001
A-Type Industry	0	0.0001***	0	0.0000**	0.0000**	0	0
B-Type Industry	0	-0.0001***	0	-0.0001*	0	-0.0001***	0
Land Fill	0	0.0000**	0.0000**	0.0001***	0.0000*	0.0000**	0
Environmental Characteristics							
Water View	0.0055**	-0.0001	0	0.0029	-0.0009	-0.0008	0.0002
Parkland View	0	-0.0002	-0.0002	0	0.0002	0.0003	0

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Road Noise	-0.0004	-0.0002	-0.0024**	-0.0037**	-0.0038***	-0.0035**	-0.0035*
Rail Noise	-0.0026	-0.0126*	-0.0086**	-0.0089**	-0.0023	-0.0046	-0.0119**
Air Noise	-0.0906*	-0.1413	0.0102	-0.0637	.	.	-0.0109
K	96	90	96	93	97	85	82
N	2261	1258	2173	895	2018	1207	970
R^2	0.721	0.830	0.800	0.807	0.790	0.847	0.829
s^2	0.0455	0.0471	0.0456	0.0382	0.0457	0.0514	0.0588

b Base case for a set of dummy variables
* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence

In Table 6, parameters for the socioeconomic variables describe an almost identical pattern to that found with the socioeconomic partitioning of the data. Property prices are depressed if the area in which the property is located is relatively poor, are inflated if the neighbourhood is inhabited by relatively highly skilled households, tend to be higher in neighbourhoods with relatively older inhabitants but lower when the neighbourhood has a proportionately larger population of households with children. In all but cluster 2, prices tend to be driven down in neighbourhoods with higher proportions of Asian and/or Black households. Cluster 2, separates the inner city properties where the majority of Asian and Black residents of Birmingham are located. Within this cluster properties in neighbourhoods with proportionately more Asian of Black households are significantly more expensive. Again the data suggests that the market rewards ethnic homogeneity.

With regards to locational characteristics, our conclusions concerning the impact on property prices from the proximity of (dis)amenities are little changed from those arrived at for the partitioning of the data according to the socioeconomic composition of neighbourhoods. One point of contrast concerns the variable describing the proximity and quality of primary schools. Notice that the parameter on primary schools is significant in cluster 2, the cluster which isolates the inner city properties. This contrasts with the results for the socioeconomic partitioning where properties in this cluster were divided between the Asian and Black socioeconomic clusters (clusters 6 and 7 of the socioeconomic partitioning) and were found not to be significant. In contrast, we find that the only cluster in which primary school proximity and quality does not exert a significant influence on property price is in cluster 6. Since this cluster identifies the large properties in the Birmingham property market this finding may simply reflect the relative lack of households with young children

and/or the availability of alternative educational opportunities that reduce the perceived importance of state funded educational institutions.

Once again the parameters on road and rail noise pollution variables are negative for all clusters. However, as a general observation, these tend to show less significance in than was exhibited in the socioeconomic partitioning.

Table 7: A selection of parameters from hedonic price equations for clusters defined by partitioning attributes of properties

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Constant	8.9319***	8.6194***	8.4596***	8.0269***	8.2942***	8.5659***
Structural Characteristics:						
Floor Area (log)	0.2795***	0.4408***	0.3860***	0.5532***	0.4595***	0.4090***
Garden Area (log)	0.0822***	0.0659***	0.1256***	0.0766***	0.0587***	0.1667***
Garage	0.0762***	0.0063	0.0349	0.0567***	0.0416***	0.0467**
Central Heating	0.0185	0.0204	-0.008	0.0701**	0.0742***	0.2206***
Age	-0.0556***	-0.0131**	0.0149	-0.0116	-0.01	-0.0281***
WCs						
One	*b*	*b*	*b*	*b*	*b*	*b*
Two	-0.0216	0.0148	0.0268	0.0211	0.0037	-0.0027
Three	0.0207	.	0.1565	.	-0.032	0.0373
Four	-0.0882	.	.	.	.	0.3829
Bedrooms						
One	-0.0434	0.1536	0.8127***	-0.0036	0.104	-0.4211***
Two	-0.0117	-0.0006	-0.0607*	0.0131	0.006	0.0404
Three	*b*	*b*	*b*	*b*	*b*	*b*
Four	0.0552**	0.0891**	0.0114	0.0323	-0.0143	0.0224
Five	0.2768***	0.2126	0.0113	.	0.0546	0.0523*
Storeys						
One	0.1283	.	-0.3993	0.0474	-0.0201	-0.0447
Two	*b*	*b*	*b*	*b*	*b*	*b*
Three	-0.1098**	0.0676	-0.1193***	-0.0598	-0.0684***	-0.1682***
Four	-0.3953***	.	-0.4555***	.	.	-0.0463

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Construction Type						
Detached	0.1697***	-0.2648***	0.1527**	0.0836***	0.0611***	0.1185***
Semi-Detached	*b*	*b*	*b*	*b*	*b*	*b*
End Terrace	-0.0607***	-0.0531*	-0.036	-0.0818***	-0.0474***	-0.1310***
Terrace	-0.0760***	-0.0740***	0.0019	-0.0764***	-0.0663***	-0.017
Sale Date:						
1st Quarter	-0.0435***	-0.0756***	-0.0626**	-0.0447***	-0.0363***	-0.0555***
2nd Quarter	-0.0147	-0.0276**	0.0142	-0.0222	-0.0097	-0.0404**
3rd Quarter	*b*	*b*	*b*	*b*	*b*	*b*
4th Quarter	-0.0016	-0.0038	0.0283	0.0241	0.0039	-0.0237
Neighbourhood Characteristics						
Poverty Factor	-0.1069***	-0.0496***	-0.0719***	-0.1023***	-0.0756***	-0.0163
Skills Factor	0.0078	0.0469***	0.0742***	0.0261**	0.0293***	0.0582***
Age Factor	0.0173**	0.0261**	0.0385**	0.0126	0.0370***	0.0522***
Family Factor	-0.011	-0.0386***	-0.0074	-0.0035	-0.0051	-0.0084
Asian Factor	-0.011	0.0326***	-0.0441**	0.0316	0.0072	-0.0681**
Black Factor	-0.0500***	0.0190**	-0.0346**	-0.0217	-0.0431***	-0.0632***
Locational Characteristics						
City Centre	0	0.0001	-0.0002	0	0	-0.0001**
Shops	0.0048	0.0115*	0.0117	-0.0168	0.0105*	-0.0350***
Primary Schools	0.1033**	0.0726*	0.2103**	0.1804***	0.0897***	0.0252
Rail Station	0.0000**	0.0000**	0	0.0000**	0.0000***	0.0000**
Park	0	0	0.0001*	0.0000*	0	0
Airport	-0.0001**	0	-0.0001	-0.0001***	-0.0001***	-0.0001**
A-Type Industry	0	0.0000***	0.0001**	0	0.0000**	0.0000***
B-Type Industry	0	0	0	0	0.0000***	0
Land Fill sites	0.0000**	0	-0.0001*	0.0001***	0	0.0000***
Environmental Characteristics						
Views of Water	-0.0023	0	0.0003	-0.0029	0.001	0.0004
Views of Parkland	-0.0002	0	0.0001	-0.0001	0	0
Road Noise	-0.0024	-0.0022**	-0.0052***	-0.0019	-0.0016*	-0.0017

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Rail Noise	-0.0063*	-0.0074**	-0.0055	-0.0128**	-0.0042	-0.0039
Aircraft Noise	0.0092	.	.	-0.0072	-0.0123	-0.0095
K	88	80	91	87	86	101
N	1540	2324	878	1176	3453	1353
R^2	0.760	0.646	0.655	0.686	0.574	0.763
s^2	0.0387	0.0525	0.0772	0.038	0.0355	0.0486

b Base case for a set of dummy variables
* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence

6.2 Comparison of the hedonic price functions across clusters

The estimation strategy followed in this paper is to capture the nonlinearity of the equilibrium hedonic price function by fitting separate price functions for the properties in each cluster. We do not estimate the hedonic price function over the entire attribute space, rather we form local approximations to the hedonic price surface over the attribute area spanned by the properties in each cluster.

Clearly, a question we would like to answer is whether this estimation strategy makes a difference. In particular, we need to test whether the hedonic price functions estimated for the different clusters differ from each other in statistically meaningful ways.

We do this by carrying out a series of pairwise comparisons. For example, we may wish to test the hypothesis that the parameters of the hedonic price function estimated from the first cluster do not differ significantly from those estimated from the second cluster. That is, we wish to test the hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ in the two linear regressions;

$$\ln(\boldsymbol{P}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}_j \qquad j = 1, 2 \qquad (13)$$

where $\boldsymbol{P}^j$, $\boldsymbol{X}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{e}_j$ are defined as before but we also assume that $\boldsymbol{e}_j$ follows a multivariate normal distribution with mean zero and covariance matrix $\sigma_j^2\boldsymbol{I}$.

In the special case in which we can assume that $\sigma^2{}_1 = \sigma^2{}_2$ the stability of the parameters can be tested using a small sample test such as the Chow Test. This approach has been adopted by a number of previous authors in this field (e.g. Michaels and Smith, 1990; Allen et al., 1995). A quick glance across the values

for s^2 (the OLS estimates of σ^2) in Tables 5 and 6 indicates that the equality of error variances is unlikely to hold true in this case. Unfortunately, when $\sigma^2{}_1 \neq \sigma^2{}_2$ the Chow test is invalid (Toyoda, 1974).

An alternative test is offered by the Wald statistic given by;

$$W = (\boldsymbol{b}_1 - \boldsymbol{b}_2)' \left(s^2{}_1 \boldsymbol{\Sigma}_1 + s^2{}_2 \boldsymbol{\Sigma}_2\right)^{-1} (\boldsymbol{b}_1 - \boldsymbol{b}_2) \tag{14}$$

where $\boldsymbol{b}_j$ and s_j^2 are the least squares estimates of $\boldsymbol{\beta}_j$ and σ_j^2 respectively and $\boldsymbol{\Sigma}_j$ is $\left(\boldsymbol{X}_j'\boldsymbol{X}_j'\right)^{-1}$. Nominally, this statistic has a chi-squared distribution with k degrees of freedom, where k is the number of parameters in common between the two models.[11] In matter of fact, the actual significance level of the Wald statistic is larger than that given by the chi-squared distribution (Kobayahsi, 1986). Unfortunately, the exact distribution of the test statistic is a complex function of the regressor variables and the error variances such that the exact significance level of a test score is almost impossible to obtain. However, Kobayashi (1986) shows that the distribution of W/k (that is, the Wald statistic divided by the number of regressors) is asymptotically bounded by the distribution of two F variates; $F(k, N_1 + N_2 - 2k)$ and $F(k, \min(N_1 - k, N_2 - k))$. The actual probability of observing a particular Wald statistic will lie between the bounds defined by these two variates.

Tables 7 and 8 present a series of pairwise comparisons of parameters for the clusters defined by neighbourhood socioeconomics and property attributes respectively. To err on the side of conservatism, the Wald statistics are based upon contrasts in only the continuous parameters of the models (including the constant, garage and central heating dummy variables). The p-values presented in these tables are the upper bound of the range identified by Kobayashi. Again, these will tend to favour acceptance of the hypothesis of equality in parameters.

Nevertheless, for all comparisons in both partitions, the test statistics are significant at a greater than 95% level of confidence[12]. In accordance with theory, there are significant between the prices that characterise the localities on the hedonic price surface isolated in the different clusters.

[11] Parameters unique to one of the models being tested were dropped from the calculation of the statistic.

[12] Wald tests based on contrasts in all the parameters of the model are significant at a greater than 99% level of confidence for all comparisons.

Table 8: Wald test chi-squared statistics for differences between hedonic price functions for neighbourhood socioeconomic characteristics partitioning

Wald Test Statistics
(p-values – Kobayashi's upper bound)

Submarket	1	2	3	4	5	6
2	118.917 (0.000)					
3	78.426 (0.000)	65.259 (0.000)				
4	74.288 (0.000)	54.36 (0.001)	55.508 (0.001)			
5	41.065 (0.024)	72.541 (0.000)	43.228 (0.014)	74.035 (0.000)		
6	50.252 (0.002)	65.859 (0.000)	86.774 (0.000)	65.997 (0.000)	48.79 (0.003)	
7	70.708 (0.000)	64.644 (0.000)	50.467 (0.003)	73.952 (0.000)	55.005 (0.001)	49.07 (0.003)

Table 9: Wald test chi-squared statistics for differences between hedonic price functions for property attribute partitioning

Wald Test Statistics
(p-values – Kobayashi's upper bound)

Submarket	1	2	3	4	5
2	152.083 (0.000)				
3	96.017 (0.000)	70.865 (0.000)			
4	54.575 (0.001)	86.429 (0.000)	52.873 (0.001)		
5	92.327 (0.000)	95.03 (0.000)	55.162 (0.001)	60.795 (0.000)	
6	132.302 (0.0000	125.608 (0.0000	44.997 (0.010)	68.651 (0.000)	101.968 (0.000)

7. Comparison of Data Partitions

The Wald tests carried out in the previous section confirm that the hedonic price function cannot be adequately approximated by a single linear regression. Rather partitioning the data and estimating a set of linear regressions one for each partition, reveals significant differences between the marginal prices of property attributes in different clusters. One further comparison needs to be made, that between the different partitions of the data. We wish to test which of the two partitionings of the data is better at isolating those regions of the hedonic price surface between which marginal prices differ significantly.

In effect we have two competing economic theories that imply different linear regression models. For example, the set of M^a linear regressions estimated for the clusters defined by partitioning according to the attributes of properties is equivalent to the single linear regression;

$$\begin{bmatrix} \ln(\boldsymbol{P}_1) \\ \ln(\boldsymbol{P}_2) \\ \vdots \\ \ln(\boldsymbol{P}_{M^a}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_{m^a} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{M^a} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_{M^a} \end{bmatrix} \tag{15}$$

or more succinctly;

$$\boldsymbol{y}^a = \boldsymbol{X}^a \boldsymbol{\beta}^a + \boldsymbol{e}^a \tag{16}$$

Likewise the set of M^b linear regressions estimated for the clusters defined by partitioning according to the socioeconomics of neighbourhoods could be represented by the single linear regression;

$$\boldsymbol{y}^b = \boldsymbol{X}^b \boldsymbol{\beta}^b + \boldsymbol{e}^b \tag{17}$$

Since the data for the model in (16) is partitioned differently to that in (17) it must be the case that neither model is a special case of the other.

Goodman and Dubin (1990) were the first to propose the use of the *J*-test (Davidson and Mackinnon, 1981) in order to compare the two hypotheses defined by the specifications in (16) and (17). The *J*-test requires artificially nesting the two models by including the fitted values from one specification as an explanatory variable in the other.

Consider first model *a* in which the data is partitioned according to the attributes of the properties themselves. Let us suppose that partitioning the data

in this way generates clusters that isolate those regions of the hedonic price surface between which the marginal prices of property characteristics differ markedly. In this case, we would expect the model in Equation (16) to fit the data very well. Imagine also, that the opposite is true of model *b*. That is, the partitioning defined by the socioeconomics of neighbourhoods does not isolate regions of the hedonic price surface characterised by markedly different marginal prices. We could test this hypothesis by artificially nesting the two models according to;

$$\boldsymbol{y}^a = \boldsymbol{X}^a \boldsymbol{\beta}^a + \alpha^b \, \hat{\boldsymbol{y}}^{b(a)} + \boldsymbol{e}^a \tag{18}$$

where $\hat{\boldsymbol{y}}^{b(a)}$ is the $N \times 1$ vector of fitted values from the linear regression in (17) with the observations reordered to conform with the arrangement of the observations in (16).

Now if our hypothesis were correct then we would not expect $\hat{\boldsymbol{y}}^{b(a)}$, the fitted values from the socioeconomic partitioning of the data, to add significantly to the explanatory power of the model in (16). Indeed, a simple *t*-test of the single parameter α^b, can be used as test of the hypothesis. If α^b is not statistically different from zero then we can conclude that partitioning the data by the socioeconomics of neighbourhoods adds nothing to the model that is not already captured by partitioning the data according to the attributes of properties themselves.

Of course we could also test the alternative hypothesis; that partitioning the data according to the attributes of the properties themselves adds nothing to our model of the hedonic price function that is not captured by partitioning the data according to the socioeconomic composition of neighbourhoods. To test this hypothesis we can artificially nest the two models according to;

$$\boldsymbol{y}^b = \boldsymbol{X}^b \boldsymbol{\beta}^b + \alpha^a \, \hat{\boldsymbol{y}}^{a(b)} + \boldsymbol{e}^b \tag{19}$$

where $\hat{\boldsymbol{y}}^{a(b)}$ is the $N \times 1$ vector of fitted values from the linear regression in (16) with the observations reordered to conform with the arrangement of the observations in (17). Again an insignificant *t*-test would allow us to accept the hypothesis that little is gained through partitioning the data according to property attributes that is not already accounted for through partitioning the data according to neighbourhood socioeconomic composition.

The results of the pair of *J*-tests defined by the models in Equations (18) and (19) are recorded in Table 9.

Table 10: J-tests of alternative partitionings of data

	Coefficient (s.e.)	p-value
H₀: *Neighbourhood Partition does not provide information beyond that already captured by Property Partition*	0.0566 (0.0086)	0.000
H₀: *Property Partition does not provide information beyond that already captured by Neighbourhood Partition*	-0.0064 (0.0038)	0.088

In this application, the *J*-test provides a clear conclusion. Including fitted values from the socioeconomics of neighbourhood partition in the model based on partitioning the data according to property characteristics significantly improves the fit of the model; α^b is significantly different from zero at over the 99.9% level of confidence. In contrast including the fitted values from the property characteristics partition into the model based on partitioning according to the socioeconomics of neighbourhoods does not significantly improve the model; α^a is not significantly different from zero at the 95% level of confidence.

Modelling the hedonic price function by partitioning the data according to the socioeconomics of neighbourhoods statistically dominates models defined by partitioning the data according to the attributes of properties themselves.

8. Conclusion

This paper has addressed the issue of estimating hedonic price equations when the hedonic price function is assumed to be highly nonlinear. Recent theoretical research has shown that such nonlinearity is likely to be a generic feature of the equilibrium hedonic price function. Furthermore, this theoretical work suggests that in equilibrium, the market may not provide a continuum of products. Rather, the equilibrium market may be characterised by clusters of properties exhibiting similar combinations of attributes whilst properties with other combinations of attributes may be sparsely represented.

In this paper we describe an application that exploits these insights. Using model-based clustering we examine the property market for evidence of clustering. Two different sets of clustering variables are used. The first set defines attributes of the properties themselves, whilst the second set defines the socioeconomic composition of the neighbourhoods in which the properties are located. In both cases there is clear evidence of clustering. We describe the characteristics of these clusters and map their distributions. In both cases we find that the clusters define readily interpretable partitions of the market.

In previous applications, researchers have addressed the issue of nonlinearity in the hedonic price function by allowing for more and more flexibility in the specification of their econometric model. Here we follow a different estimation strategy. First we note that properties categorised into the same cluster lie in close proximity to each other in certain dimensions of the attribute space. By extension, these properties must also lie close to each other in these dimensions on the hedonic price surface. We hypothesise that partitioning the data generates clusters that isolate regions of the hedonic price surface between which the marginal prices of property characteristics differ markedly.

Thus, rather than employing increasingly more general econometric specifications to capture the nonlinearity of the equilibrium hedonic price function, our estimation strategy is to avoid estimating the hedonic price function over the entire attribute space. Rather, we fit separate price functions for the properties in each cluster thereby forming local approximations to the hedonic price surface over the attribute area spanned by the properties in each cluster.

In the application described here, we find that the hedonic price function cannot be adequately approximated by a single linear regression. Rather partitioning the data and estimating a set of linear regressions, one for each partition, reveals significant differences between the marginal prices of property attributes in different clusters. Indeed, one of the advantages of this approach when compared to estimation strategies based on nonparametric regression, is that the parameters estimated on the various covariates can be examined for interesting contrasts across clusters. In the application described here, for example, we find

that the market tends to reward ethnic homogeneity within neighbourhoods. That is, in clusters whose properties are located in majority white neighbourhoods, property prices tend to be depressed the greater the proportion of ethnic minority households present in those neighbourhoods. Likewise, in clusters whose properties are located in majority ethnic minority neighbourhoods, property prices tend to be inflated the greater the proportion of ethnic minority households present in those neighbourhoods.

Finally we test to see whether one of the two proposed partitionings can be said to provide a better description of the data in the model than the other. Using a *J*-test we discover that partitioning the data according to the socioeconomic characteristics of neighbourhoods, statistically dominates a model in which the data has been partitioned according to the attributes of properties. It appears that differences in property prices can better be captured by looking at the differences that exist between socioeconomically differing neighbourhoods than by examining the differences that exist between different structural types of property. This lends a certain amount of credence to one of the estate agents fundamental laws; “Always buy the worst house in the best area, never the best house in the worst area”.

References

Abraham, J. M., W. N. Goetzmann and S. M. Wachter (1994). "Homogenous groupings of metropolitan housing markets", Journal of Housing Economics, 3 (3), pp. 186-206.

Allen, M. T., T. M. Springer and N. G. Waller (1995). "Implicit pricing across residential rental markets", *Journal of Real Estate Finance and Economics*, 11, pp 137-151.

Anglin, P. M., and R. Gencay (1996). "Semiparametric estimation of a hedonic price function", *Journal of Applied Econometrics*, 11, pp 633-648.

Ball, M. J. and R. M. Kirwan (1977). "Accessibility and supply constraints in the urban housing market", *Urban Studies*, 14, pp 11-32.

Banfield, J. D., and A. E. Raftery (1993). "Model-based Gaussian and non-Gaussian clustering", *Biometrics*, 49, pp 803-821.

Bensmail, H., G. Celeux, A. E. Raftery, and C. P. Robert (1997). "Inference in model-based cluster analysis", *Statistics and Computing*, 7, pp 1-10.

Biernacki, C., G. Celeux and G. Govaert (1999). "An improvement of the NEC criterion for assessing the number of clusters in a mixture model", *Pattern Recognition Letters*, 20(3), pp. 267-272.

Biernacki, C., G. Celeux and G. Govaert (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models", *Computational Statistics and Data Analysis*, 41 (3-4), pp. 561-575.

Biernacki, C., and G. Govaert (1999). "Choosing models in model-based clustering and discriminant analysis", *Journal of Statistical Computation and Simulation*, 64(1), pp. 49-71.

Blackley, P., J. R. Follain, Jr., and J. Ondrich (1984). "Box-Cox estimation of hedonic models: How serious is the iterative OLS variance bias?", *Review of Economics and Statistics*, 66, pp. 348-353.

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

Byers, S. D., and A. E. Raftery (1998). "Nearest neighbour clutter removal for estimating features in a point process", *Journal of the American Statistical Association*, 93, pp 577-584.

Campbell, J. G., C. Fraley, D. Stanford, F. Murtagh and A. E. Raftery (1999). "Model-based methods for textile fault detection", *International Journal of Imaging Systems And Technology*, 10 (4), pp. 339-346.

Can, A., (1992). "Specification and estimation of hedonic housing price models", *Regional Science and Urban Economics*, 22, pp 453-474.

Cassel, E., and R. Mendelsohn (1985). "The choice of functional forms for hedonic price equations: Comment", *Journal of Urban Economics*, 18, pp 135-142.

Celeux, G., G. and Govaert (1995). "Gaussian parsimonious clustering models", *Pattern Recognitio*n, 28, pp. 781–793.

Cheshire, P. and S. Sheppard (1998). "Estimating the demand for housing, land and neighbourhood characteristics", *Oxford Bulletin of Economics and Statistics*, 60(3), pp357-382.

Cropper, M. L., L. B. Deck and K. E. McConnell (1988). "On the choice of functional form for hedonic price functions", *Review of Economics and Statistics*, 70, pp 668-75.

Dasgupta, A., and A. E. Raftery (1998). " Detecting features in spatial point processes with clutter via model-based clustering", *Journal of the American Statistical Association*, 93, pp. 294-302.

Davidson, R., and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses", *Econometrica*, 49, pp. 781-793.

Day, B. H., (2003). "Submarket Identification in Property Markets: A Hedonic Housing Price Model for Glasgow", *CSERGE Working Paper*, EDM 03-09, University of East Anglia, UK.

Day, B. H., and I. J. Bateman and I. Lake (2003). "What Price Peace? A Comprehensive Approach to the Specification and Estimation of Hedonic Housing Price Models", *CSERGE Working Paper*, EDM 03-08, University of East Anglia, UK.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood for incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, 39, pp. 1–38.

Department of the Environment Transport and the Regions (2000). *A report on the production of noise maps of the City of Birmingham*. London: HMSO.

Ekeland, I., J. J. Heckman and L. Nesheim (2002). "Identifying hedonic models", *American Economic Review*, 92(2), pp. 304-309.

Ekeland, I., J. J. Heckman and L. Nesheim (2003). "Identification and estimation of hedonic models", *Journal of Political Economy*, forthcoming.

Epple, D., (1987). "Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products", *Journal of Political Economy,* 95, pp 59-80.

Fasulo, D., (1999). "An analysis of recent work on clustering algorithms", Technical Report No. 01-03-02, Department of Computer Science and Engineering, University of Washington, Seattle, WA.

Fovell, R. G., (1997). "Consensus Clustering of U.S. Temperature and Precipitation Data Authors", *Journal of Climate*, 10 (6), pp. 1405-1427.

Fraley, C., and A. E. Raftery (1998). "How many clusters? Which clustering method? Answers via model-based cluster analysis", *The Computer Journal*, 41(8), pp. 578-588.

Fraley, C., and A. E. Raftery (2002a). "Model-based clustering, discriminant analysis and density estimation", *Journal of the American Statistical Association*, 97(458), pp. 611-631.

Fraley, C., and A. E. Raftery (2002b). "MCLUST: Software for model-based clustering, density estimation and discriminant analysis", *Technical Report No. 415*, Department of Statistics, University of Washington.

Ghosh, D., and A. M. Chinnaiyan (2002). "Mixture modelling of gene expression data from microarray experiments", *Bioinformatics*, 18, pp. 275-286.

Gibbons, S., and S. Machin (2001). "Valuing primary schools", *Centre for the Economics of Learning Discussion Paper*, 15.

Gibbons, S., and S. Machin (2003). "Valuing English primary schools", *Journal of Urban Economics*, 53(2), 15.

Goetzmann, W. N., and S. M. Wachter (1995). "Clustering methods for real estate portfolios", *Real Estate Economics*, 23 (3), pp. 271-310.

Goodman, A. C., (1978). "Hedonic prices, price indexes and housing markets", *Journal of Urban Economics*, 5, pp 471-484.

Goodman, A. C., and R. A. Dubin (1990). "Sample Stratification with Non-Nested Alternatives: Theory and a Hedonic Example," *Review of Economics and Statistics*, 72, pp 168-173.

Goodman, A. C., and T. G. Thibodeau (1998). "Housing market segmentation", *Journal of Housing Economics*, 7, pp. 175-185.

Goodman, A. C., and T. G. Thibodeau (2003). "Housing market segmentation and hedonic prediction accuracy", *Journal of Housing Economics*, 12(3), pp.181-201.

Gopal, S. S., and J. Hebet (1998). "Bayesian Pixel Classification using Spatially Variant Finite Mixures and Generalised EM Algorithm", *IEEE Trans. on Image Processing*, 7, pp. 1014-1028.

Halvorsen, R. and H. O. Pollakowski (1981). "Choice of functional form for hedonic price equations", *Journal of Urban Economics*, 10, pp 37-49.

Heckman, J., R. Matzkin and L. Nesheim (2002). "Non-parametric estimation of nonadditive hedonic models", unpublished working paper.

Heckman, J., R. Matzkin and L. Nesheim (2003). "Simulation and estimation of hedonic models", *CENMAP Working Paper*, CWP10/03, Institute of Fiscal Studies, UCL, London.

Hoesli, M., and B. D. MacGregor (1995). "The classification of local property markets in the UK using cluster analysis", in *The Cutting Edge: Proceedings of the RICS Property Research Conference, 1995, Vol. 1*, London: Royal Institution of Chartered Surveyors.

Huh, S., and S.-J. Kwak (1997). "The choice of functional forma and variables in the hedonic price model in Seoul", *Urban Studies*, 34(7), pp 989-998.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). "Data Clustering: A Review", *ACM Computing Surveys*, 31(3), pp. 264-323.

Keribin, C., (1998). "Consistent estimate of the order of mixture models", *Comptes Rendues de l'Academie des Sciences, série I – Mathématiques*, 326, pp. 243-248. *Annals of Statistics*, 20, pp. 1350-1360.

Kobayashi, M., (1986). "A bounds test for equality between sets of coefficients in two linear regressions when disturbance variances are unequal", *Journal of the American Statistical Association*, 81(394), pp. 510-513.

Lake, I. R., A. A. Lovett, I. J. Bateman and B. H. Day (2000). "Using GIS and large-scale digital data to implement hedonic pricing studies", *International Journal of Geographical Information Science*, 14(6), pp. 521-541

Leroux, M., (1992). "Consistent estimation of a mixing distribution", *The Annals of Statistics*, 20, pp. 1350-1360.

Maitra, R., (2001). "Clustering massive datasets with applications in software metrics and tomography", Technometrics, 20 (3), pp. 336-346.

McLachlan, G., (1987). "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture", *Applied Statistics*, 36, pp. 318-324.

McLachlan, G., and K. Basford (1988). *Mixture Models*, New York: Marcel Dekker.

McLachlan, G. J., R. W. Bean and D. Peel (2002). "A mixture model-based approach to the clustering of microarray expression data", *Bioinformatics*, 18(3), pp.413-422.

Michaels, R. G., and V. K. Smith (1990). "Market-segmentation and valuing amenities with hedonic models - the case of hazardous-waste sites", *Journal of Urban Economics*, 28 (2), pp 223-242.

Milon, J. W., J. Gressel and D. Mulkey (1984). "Hedonic amenity valuation and functional form specification", *Land Economics*, 60 (4), pp 378-387.

Mukherjee, S., E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley and A. E. Raftery (1998). "Three types of gamma-ray bursts", *Astrophysical Journal*, 508 (1), pp. 314-327.

Nesheim, L., (2002). "Equilibrium sorting of heterogeneous consumers across locations: theory and empirical implications", C*EMMAP Working Paper*, CWP08/02, Insititute of Fiscal Studies, Department of Economics, University College London.

Pace, R. K., (1993). "Nonparametric methods with applications to hedonic models", *Journal of Real Estate Finance and Economics*, 7(3), pp.185-204.

Pace, R. K., (1995). "Parametric, semiparametric and nonparametric estimation of characteristic values with mass assessment and hedonic pricing models", *Journal of Real Estate Finance and Economics*, 11, pp.195-217.

Pan, W., J. Lin and C. T. Le (2002). "Model-based cluster analysis of microarray gene-expression data", *Genome Biology*, 3(2): research 0009.1–0009.8.

Penrose, M. D., (1998). "Extremes for the minimal spanning tree on normally distributed points", *Advances in Applied Probability*, 30, pp 628-639.

Posse, C., (2001). "Hierarchical model-based clustering for large datasets", *Journal Of Computational And Graphical Statistics,* 10 (3), pp. 464-486.

Prim, R. C., (1957). "Shortest connection networks and some generalizations", *Bell Systems Technology Journal,* 36, pp. 1389-1401.

Ramussen, D. W., and T. W. Zuehlke (1990). "On the choice of functional form for hedonic price functions", *Applied Economics*, 22, pp 431-438.

Roeder, K., and L. Wasserman (1997). "Practical Bayesian density estimation using mixtures of normals", *Journal of the American Statistical Association*, 87, pp. 108-119.

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

Schnare, A., and R. Struyk (1976). "Segmentation in urban housing markets", *Journal of Urban Economics*, 3, pp 146-166.

Smyth, P., (2000). "Model selection for probabilistic clustering using cross-validated likelihood", *Statistics and Computing*, 10(1), pp. 63-72.

Smyth, P., K. Ide and M. Ghil (1999). "Multiple regimes in Northern Hemisphere height fields via mixture model clustering", *Journal of Atmospheric Sciences*, 56(21) pp. 3704-3723.

Sonstelie, J. C., and P. R. Portney (1980). "Gross rents and market values: Testing the implications of Trebout's hypothesis", *Journal of Urban Economics*, 7, pp 102-118.

Stanford, D., and A. E. Raftery (2000). "Principal curve clustering with noise", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, pp. 601-609.

Straszheim, M. R., (1973). "Estimation of the demand for urban housing services from household interview data", *Review of Economics and Statistics*, 55, pp 1-8.

Tauchen, H., and A. D. Witte (2001). "Estimating hedonic models: Implications of the theory", National Bureau of Economic Research, Technical Working Paper, 271.

Ter Braak, C. J. F., H. Hoijtink, W. Akkermans, P. F. M. Verdonschot (2003). "Bayesian model-based cluster analysis for predicting macrofaunal communities", *Ecological Modelling*, 160 (3), pp. 235-248.

Tinbergen, J., (1956). "On the theory of income distribution", Weltwirtschaftliches Archiv, 77, pp 155-73.

Titterington, D. M., A. F. M. Smith and U. E. Markov (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, Chichester, UK.

Toyoda, T., (1974). "Use of the Chow test under heteroskedasticity", *Econometrica*, 42, pp. 601–608.

Ward, J. H., (1963). "Hierarchical groupings to optimize an objective function", *Journal of the American Statistical Association*, 58, pp. 234–244.

Wehrens, R., L. M. C. Buydens, C. Fraley and A. E. Raftery (2003). "Model-based clustering for image segmentation and large datasets via sampling", *Technical Report No. 424*, Department of Statistics, University of Washington.

Wolverton, M. L., W. G. Hardin and P. Cheng (1999). "Disaggregation of local apartment markets by unit type", *Journal of Real Estate Finance and Economics*, 19 (3), pp. 243-257.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, W. L. Ruzzo (2001). "Model-based clustering and data transformations for gene expression data", *Bioinformatics*, 17, pp. 977-987.

Appendix A: Parameters of Hedonic Price Equations for Neighbourhood Socioeconomic Characteristics Partitioning

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Constant	8.9067***	8.2327***	8.7047***	8.4688***	8.3764***	8.4239***	8.9403***
Structural Characteristics:							
Floor Area (log)	0.3827***	0.3991***	0.4383***	0.3612***	0.4879***	0.3864***	0.3670***
Garden Area (log)	0.0838***	0.1662***	0.0973***	0.1005***	0.0940***	0.1393***	0.1446***
Garage	0.0448***	0.0579***	0.0550***	0.0350*	0.0524***	0.0607***	0.0369
Central Heating	0.0464*	0.0653*	0.0577**	-0.0283	0.1032***	0.0828**	-0.0716*
Age	-0.0148**	-0.0067	-0.0096	-0.0058	-0.0204***	-0.0091	-0.0106
WCs							
One	b	b	b	b	b	b	b
Two	0.0243*	-0.0399**	0.0315**	-0.0059	0.0297**	-0.022	-0.0244
Three	0.0198	0.2056**	-0.0112	-0.022	0.1304***	0.0222	0.0075
Four	.	0.8627***	-0.2295*	.	0.4666**	.	.
Five	.	0.372	.	.	.	.	.
Bedrooms							
One	0.0727	0.0675	0.0414	0.2473**	0.0351	0.3394	0.1957
Two	0.007	-0.0013	0.0127	-0.0299	0.0152	-0.0062	0.0560**

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Three	b	b	b	b	b	b	b
Four	0.0278	0.0029	0.0165	0.0279	0.0407*	0.0677**	0.047
Five	0.0452	-0.0609	0.1349***	0.1474**	0.1459***	0.1758***	0.044
Six	-0.1925*	-0.0346	0.0692	0.4663*	0.3435***	0.1821**	-0.0534
Seven	0.1105	-0.3969***	0.6348***	-0.2241	-0.5027***	0.2726*	0.4141
Eight	.	-0.0483	.	0.4797**	-0.0843	.	0.2517
Nine	.	.	0.0221	.	.	.	.
Storeys							
One	-0.07	-0.4751	-0.037	0.0449	0.1903***	0.2255	0.1281
Two	b	b	b	b	b	b	b
Three	-0.0481	-0.2195***	-0.1069***	-0.0672	-0.1115***	-0.0166	0.0183
Four	-0.2106*	-0.8875***	-0.4576***	-0.1522	-0.1909*	-0.1956	-0.4995**
Five	.	.	.	-0.3947*	-0.3526	.	-0.5758*
Construction Type							
Detached Bungalow	0.2031	0.8569***	0.1771	0.1213	.	-0.1787	.
Semi-Detached Bungalow	0.1699	.	0.0075	.	-0.0383	.	-0.0694
End Terrace Bungalow	-0.0122	.	.	.	.	.	.
Terrace Bungalow	.	.	.	.	.	.	0.0827

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Detached House	0.1396***	0.1477***	0.1220***	0.1386***	0.1087***	0.0721***	0.0884**
Semi-Detached House	b	b	b	b	b	b	b
End Terrace House	-0.0887***	-0.0981***	-0.0440**	-0.0493*	-0.0780***	-0.0309	-0.1012***
Terrace House	-0.0795***	-0.0418*	-0.0647***	-0.0407*	-0.0917***	-0.0763***	-0.0833***
Beacon Group							
1. Unrenovated cottage pre 1919	-0.1371	.	.	0.0228	-0.3349	-0.0215	.
2. Renovated cottage pre 1919	.	0.1031	0.3737*	.	0.6604***	0.2239	0.2426
3. Small “industrial” pre 1919	-0.1158***	0.0321	0.1031**	0.0414	-0.0325	-0.0665	-0.1759***
4. Medium “industrial” pre 1919	-0.0225	0.0257	-0.0485*	-0.0053	-0.0259	0.0092	-0.0545
5. Large terrace pre 1919	0.0134	0.0614	-0.0662	-0.1176	0.0702	0.029	-0.1293
8. Small “villa” pre 1919	0.0092	0.0927	-0.0518	0.0416	0.0324	-0.0365	0.2134***
9. Large “villas” pre 1919	0.036	0.0955	0.0222	0.1769*	-0.0274	0.1872**	0.1927**
10. Large detached pre 1919	0.4524**	-0.1677	0.2266***	0.0403	0.4721***	-0.1772	-0.4598*
19. Houses 1908 to 1930	0.1012**	0.072	-0.0585	0.0704	0.0748	0.1327**	0.1309*
20. Subsidy houses 1920s & 30s	-0.0805***	-0.0919***	-0.0880***	-0.1278***	-0.0292	0.0157	-0.0545

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
21. Standard houses 1919-45	b	b	b	b	b	b	b
24. Large houses 1919-45	0.2597***	0.1120**	0.2021***	0.1768**	0.1352***	0.2615***	0.1892**
25. Individual houses 1919-45	0.1885	.	0.0152	.	-0.3234	0.1769	.
30. Standard houses 1945-53	-0.0851***	-0.1605***	-0.1436***	-0.0127	-0.0845***	-0.0851*	-0.0498
31. Standard houses post 1953	-0.0491	-0.0169	-0.0204	-0.0261	-0.0543	0.0304	-0.011
32. Large houses post 1953	0.1536***	0.1201*	0.1202***	0.1581**	0.0157	0.1359**	0.1013
35. Individual houses post 1945	0.5353***	0.4387*	-0.214	0.0809	0.0312	0.0838	.
36. “Town Houses” post 1950	-0.2377***	-0.1489	-0.1329	.	-0.2121	-0.2559***	-0.1947
Sale Date							
1st Quarter (Jan. to Mar.)	-0.0508***	-0.0313*	-0.0564***	-0.0400**	-0.0407***	-0.0716***	-0.0675***
2nd Quarter (Apr. to June)	-0.0231*	-0.017	-0.0224*	-0.0495***	-0.009	0.0246	-0.0262
3rd Quarter (July to Sept.)	b	b	b	b	b	b	b
4th Quarter (Oct. to Dec.)	-0.0039	-0.0094	0.0023	-0.0171	0.015	0.0269	-0.0101
Neighbourhood Characteristics							
Poverty Factor	-0.0871***	-0.0736***	-0.0648***	-0.1284***	-0.0471***	-0.1260***	-0.0483**
Skills Factor	0.0628***	0.0305***	0.0524***	0.0401**	0.0662***	0.0055	0.0404**

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Age Factor	0.0209***	0.0303**	0.014	0.0264*	0.0098	0.0587***	0.0404**
Family Factor	-0.0075	-0.0489***	-0.0205*	-0.0255	-0.0106	-0.0248	-0.0506***
Asian Factor	0.0137	-0.0482***	-0.0368***	-0.0697**	-0.0118	0.0314**	0.0438**
Black Factor	-0.0254**	0.0046	-0.0517***	-0.0314	-0.0520***	0.0269**	-0.011
Locational Characteristics							
Proximity to City Centre	0.0001**	-0.0001	-0.0001**	-0.0001	0	0.0001*	-0.0001*
Proximity and Quantity of Shops	0.0230***	-0.0134	-0.0347***	-0.0023	0.0126	0.0271**	-0.0363***
Proximity and Quality of Primary Schools	0.0961**	0.1593***	0.1089***	0.1766***	0.0942**	0.0134	0.0404
Walking time to Rail Station	0	0	0	0.0001***	0.0000**	0	0
Walking time to a Park	0	0	0.0000**	0	0	0	0
Driving time to Airport	-0.0001***	-0.0001	0	0	-0.0001***	-0.0001**	-0.0001
Proximity to A-Type Industrial Processes	0	0.0001***	0	0.0000**	0.0000**	0	0
Proximity to B-Type Industrial Processes	0	-0.0001***	0	-0.0001*	0	-0.0001***	0
Proximity to Land Fill sites	0	0.0000**	0.0000**	0.0001***	0.0000*	0.0000**	0
Wards							
Acock's Green	-0.2896**	-0.0878	-0.2595***	0.0285	-0.1522***	-0.128	0.0541

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Aston	.	-0.5673	-0.3934**	.	.	-0.2820**	-0.1593
Bartley Green	-0.1273	.	-0.2058***	-0.1212	-0.1503*	.	.
Billesley	-0.0747	-0.0923	-0.1043**	0.0655	-0.0829*	.	.
Bournville	0.0683	0.2489*	-0.0475	0.078	0.0654	.	.
Brandwood	-0.1095	0.1651	-0.1362**	0.0854	0.0599	0.056	.
Edgbaston	0.0399	0.2047	-0.2538***	0.118	0.1255*	0.5427***	0.3242
Erdington	-0.2274**	0.0281	-0.1146***	0.0413	-0.1196***	-0.0675	.
Fox Hollies	-0.2115*	.	-0.2427***	0.1233*	-0.1544***	-0.1841*	0.1992
Hall Green	-0.1784	0.0166	-0.1187***	0.0472	-0.0839**	0.122	0.1605
Handsworth	.	-0.2885**	-0.2472*	.	-0.2132***	-0.2763**	-0.0937
Harborne	0.0748	0.3169**	0.0817	0.1066	0.3960***	.	.
Hodge Hill	-0.2692**	.	-0.2717***	0.1024	-0.0918**	-0.0024	.
King's Norton	-0.0621	.	-0.0571	-0.0436	-0.0245	.	.
Kingsbury	-0.2660**	.	-0.2635***	0.0237	-0.0245	.	.
Kingstanding	-0.1627	-0.1531	-0.1734***	-0.0853	-0.0694	0.0205	-0.2233
Ladywood	-0.084	0.1058	-0.3156***	0.1636	-0.0301	-0.221	.
Longbridge	-0.1124	0.169	-0.1032	-0.0639	0.0824	.	0.4550***
Moseley	.	0.155	-0.3543***	-0.3153**	-0.0606	0.1903	0.3409**

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Nechells	-0.3682***	-0.2002	-0.1422**	0.0916	-0.4287***	-0.7439***	-0.0714
Northfield	-0.0855	.	-0.0881	0.0913	0.1691	.	.
Oscott	-0.2088*	-0.3334**	-0.2240***	-0.2329**	-0.1163**	.	.
Perry Barr	-0.1629	-0.2357*	-0.2433***	-0.1577	-0.1565***	.	.
Quinton	0.0412	0.2236	-0.1776*	-0.0967	0.0335	0.1068	.
Sandwell	-0.1557	-0.3896***	-0.2749***	-0.0042	-0.2002***	-0.0467	-0.1068
Selly Oak	0.0982	0.2363	.	0.2023*	0.1423**	.	.
Shard End	-0.4113***	.	-0.2220***	0.118	-0.1345	-0.4445***	.
Sheldon	-0.2656**	.	-0.2177***	.	0.0045	-0.067	-0.0842
Small Heath	-0.2866**	0.0406	.	.	-0.2047	-0.1810*	-0.1147
Soho	.	-0.3117**	.	.	.	-0.4080***	-0.2027
Sparkbrook	.	0.0503	.	.	.	-0.1595	-0.0112
Sparkhill	-0.1936	0.0833	-0.058	0.2220*	-0.0819	-0.0868	0.2269
Stockland Green	.	.	-0.2485***	-0.0557	-0.1976***	.	.
Sutton Four Oaks	0.0501	-0.1293	0.0659*	0.1483	0.0454	.	.
Sutton New Hall	b	B	b	b	b	b	b
Sutton Vesey	-0.0638	.	.	0.0936	-0.0472	0.1214	0.2678*
Washwood Heath	-0.3167***	.	-0.3352***	0.0473	-0.1505***	-0.2755**	-0.4276

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Weoley	-0.11	0.0734	-0.2655***	0.0939	-0.0187	0.1514	0.1722
Yardley	-0.2692**	-0.0777	-0.2097***	-0.0155	.	-0.1086	0.0141
Environmental Characteristics							
Views of Water	0.0055**	-0.0001	0	0.0029	-0.0009	-0.0008	0.0002
Views of Parkland	0	-0.0002	-0.0002	0	0.0002	0.0003	0
Road Traffic Noise	-0.0004	-0.0002	-0.0024**	-0.0037**	-0.0038***	-0.0035**	-0.0035*
Rail Traffic Noise	-0.0026	-0.0126*	-0.0086**	-0.0089**	-0.0023	-0.0046	-0.0119**
Aircraft Noise	-0.0906*	-0.1413	0.0102	-0.0637	.	.	-0.0109
K	96	90	96	93	97	85	82
N	2261	1258	2173	895	2018	1207	970
R^2	0.721	0.830	0.800	0.807	0.790	0.847	0.829
s^2	0.0455	0.0471	0.0456	0.0382	0.0457	0.0514	0.0588

b Base case for a set of dummy variables
* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence

Appendix B: Parameters of Hedonic Price Equations for Property Attribute Partitioning

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Constant	8.9319***	8.6194***	8.4596***	8.0269***	8.2942***	8.5659***
Structural Characteristics:						
Floor Area (log)	0.2795***	0.4408***	0.3860***	0.5532***	0.4595***	0.4090***
Garden Area (log)	0.0822***	0.0659***	0.1256***	0.0766***	0.0587***	0.1667***
Garage	0.0762***	0.0063	0.0349	0.0567***	0.0416***	0.0467**
Central Heating	0.0185	0.0204	-0.008	0.0701**	0.0742***	0.2206***
Age	-0.0556***	-0.0131**	0.0149	-0.0116	-0.01	-0.0281***
WCs						
One	b	b	b	b	b	b
Two	-0.0216	0.0148	0.0268	0.0211	0.0037	-0.0027
Three	0.0207	.	0.1565	.	-0.032	0.0373
Four	-0.0882	.	.	.	.	0.3829
Five	.	.	.	.	.	0.2078
Bedrooms						
One	-0.0434	0.1536	0.8127***	-0.0036	0.104	-0.4211***
Two	-0.0117	-0.0006	-0.0607*	0.0131	0.006	0.0404
Three	b	b	b	b	b	b

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Four	0.0552**	0.0891**	0.0114	0.0323	-0.0143	0.0224
Five	0.2768***	0.2126	0.0113	.	0.0546	0.0523*
Six	0.0806	-0.1611	0.0624	.	.	0.0511
Seven	.	.	-0.1037	.	.	-0.0395
Eight	.	.	-0.4149	.	.	-0.0712
Nine	.	.	-0.5133*	.	.	.
Storeys						
One	0.1283	.	-0.3993	0.0474	-0.0201	-0.0447
Two	b	b	b	b	b	b
Three	-0.1098**	0.0676	-0.1193***	-0.0598	-0.0684***	-0.1682***
Four	-0.3953***	.	-0.4555***	.	.	-0.0463
Five	.	.	-0.7086***	.	.	-0.2824
Construction Type						
Detached Bungalow	0.0918	.	.	0.1858**	.	0.1756***
Semi-Detached Bungalow	-0.1722	.	.	.	0.1684**	.
End Terrace Bungalow	-0.2379	.	.	.	.	0.5389**
Terrace Bungalow	.	.	.	.	.	.
Detached House	0.1697***	-0.2648***	0.1527**	0.0836***	0.0611***	0.1185***

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Semi-Detached House	b	b	b	b	b	b
End Terrace House	-0.0607***	-0.0531*	-0.036	-0.0818***	-0.0474***	-0.1310***
Terrace House	-0.0760***	-0.0740***	0.0019	-0.0764***	-0.0663***	-0.017
Beacon Group						
1. Unrenovated cottage pre 1919	.	0.0157	0.0739	-0.2473	.	.
2. Renovated cottage pre 1919	.	.	0.0873	0.5262**	.	0.3516***
3. Small “industrial” pre 1919	.	-0.0458***	-0.0887	.	0.2271	0.0294
4. Medium “industrial” pre 1919	b	b	b	b	b	b
5. Large terrace pre 1919	.	-0.3192	0.0271	.	.	0.1018
8. Small “villa” pre 1919	.	0.1162***	0.0486	-0.0165	0.1986	-0.0999
9. Large “villas” pre 1919	.	.	0.0499	.	.	0.1009*
10. Large detached pre 1919	.	.	-0.2982*	.	.	-0.0102
19. Houses 1908 to 1930	.	0.0965	0.1787***	0.1524	0.1675**	-0.0186
20. Subsidy houses 1920s & 30s	.	0.0524	-0.0571	0.0032	0.1088**	-0.1699***
21. Standard houses 1919-45	.	0.1853***	-0.0717	0.0716	0.2073***	-0.1203***

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
24. Large houses 1919-45	.	.	-0.0338	0.4530***	0.2834***	0.0077
25. Individual houses 1919-45	.	.	.	.	.	-0.2203
30. Standard houses 1945-53	-0.287	.	.	-0.0253	0.1	-0.2545***
31. Standard houses post 1953	0.5236***	.	.	-0.0163	.	-0.2048***
32. Large houses post 1953	0.5639***	.	.	-0.2800**	0.4193***	-0.107
35. Individual houses post 1945	0.6099***	.	.	.	.	0.1258
36. “Town Houses” post 1950	0.3640***	.	.	-0.1128	.	.
Sale Date						
1st Quarter (Jan. to Mar.)	-0.0435***	-0.0756***	-0.0626**	-0.0447***	-0.0363***	-0.0555***
2nd Quarter (Apr. to June)	-0.0147	-0.0276**	0.0142	-0.0222	-0.0097	-0.0404**
3rd Quarter (July to Sept.)	b	b	b	b	b	b
4th Quarter (Oct. to Dec.)	-0.0016	-0.0038	0.0283	0.0241	0.0039	-0.0237
Neighbourhood Characteristics						
Poverty Factor	-0.1069***	-0.0496***	-0.0719***	-0.1023***	-0.0756***	-0.0163
Skills Factor	0.0078	0.0469***	0.0742***	0.0261**	0.0293***	0.0582***
Age Factor	0.0173**	0.0261**	0.0385**	0.0126	0.0370***	0.0522***

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Family Factor	-0.011	-0.0386***	-0.0074	-0.0035	-0.0051	-0.0084
Asian Factor	-0.011	0.0326***	-0.0441**	0.0316	0.0072	-0.0681**
Black Factor	-0.0500***	0.0190**	-0.0346**	-0.0217	-0.0431***	-0.0632***
Locational Characteristics						
Proximity to City Centre	0	0.0001	-0.0002	0	0	-0.0001**
Proximity and Quantity of Shops	0.0048	0.0115*	0.0117	-0.0168	0.0105*	-0.0350***
Proximity and Quality of Primary Schools	0.1033**	0.0726*	0.2103**	0.1804***	0.0897***	0.0252
Walking time to Rail Station	0.0000**	0.0000**	0	0.0000**	0.0000***	0.0000**
Walking time to a Park	0	0	0.0001*	0.0000*	0	0
Driving time to Airport	-0.0001**	0	-0.0001	-0.0001***	-0.0001***	-0.0001**
Proximity to A-Type Industrial Processes	0	0.0000***	0.0001**	0	0.0000**	0.0000***
Proximity to B-Type Industrial Processes	0	0	0	0	0.0000***	0
Proximity to Land Fill sites	0.0000**	0	-0.0001*	0.0001***	0	0.0000***
Wards						
Acock's Green	-0.0985**	-0.2980***	-0.1735	-0.2238***	-0.1338***	-0.3068***
Aston	-0.1998***	-0.4744***	-0.2	-0.2945*	-0.3373***	.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Bartley Green	-0.0526	-0.5340**	.	-0.1370*	-0.2129***	-0.1053
Billesley	-0.1131**	-0.6512***	0.2093	-0.0018	-0.0607	-0.0484
Bournville	0.0687	-0.0984	0.1722	0.0723	0.0207	0.1982**
Brandwood	-0.0394	-0.1043	-0.0247	0.0477	-0.056	0.0407
Edgbaston	0.1398**	0.0486	0.0294	0.1335	0.3624*	0.0787
Erdington	-0.1848***	-0.3205***	-0.1721	-0.1381***	-0.0773	-0.1442***
Fox Hollies	-0.1697*	-0.3561***	-0.2723**	-0.2320***	-0.0689	-0.0807
Hall Green	-0.0787	-0.2242***	-0.1633	-0.1694***	-0.0329	-0.0237
Handsworth	-0.2831***	-0.5141***	-0.2026	-0.0984	-0.1776***	-0.1319
Harborne	0.104	0.1831**	0.3256**	0.0684	0.0666	0.3190***
Hodge Hill	-0.2725***	-0.2894***	-0.5603*	-0.0917*	-0.1560***	-0.0930*
King's Norton	-0.0365	0.0785	0.195	-0.0082	-0.0606	0.1458
Kingsbury	-0.0874*	0.0114	0.1579	-0.2104***	-0.1471***	-0.0318
Kingstanding	-0.0694	-0.6128**	0.16	-0.3544***	-0.2166***	-0.2141**
Ladywood	-0.1048	-0.2405***	0.0595	0.0529	0.0634	-0.0252
Longbridge	0.0882	.	0.2941	0.0402	-0.084	0.1962*
Moseley	-0.1062	-0.0388	0.0472	0.0526	-0.0069	0.0964
Nechells	-0.0133	-0.4448***	0.1268	-0.0549	-0.0976	-0.3924**

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Northfield	-0.0606	-0.1584*	0.1384	0.0003	-0.0464	0.0858
Oscott	-0.1006	.	.	-0.0478	-0.2653***	-0.0749
Perry Barr	-0.1452**	-0.4767***	.	-0.1237*	-0.1902***	-0.2123**
Quinton	-0.0711	0.0591	0.3211*	0.0047	-0.0732	0.089
Sandwell	-0.0959	-0.5222***	-0.3727***	-0.1515**	-0.2528***	-0.2675***
Selly Oak	0.0689	0.0528	0.0966	0.0408	0.0666	0.1404*
Shard End	-0.2729***	.	-0.2846	-0.2451***	-0.2070***	-0.0312
Sheldon	-0.3771***	0.0495	.	-0.2283***	-0.1800***	-0.1985**
Small Heath	-0.6870***	-0.3508***	-0.0682	-0.1042	-0.1147*	-0.0814
Soho	-0.2091***	-0.5691***	-0.3294**	-0.2014	-0.2353***	0.1125
Sparkbrook	-0.0528	-0.3781***	-0.107	.	.	0.2386
Sparkhill	0.0673	-0.2401***	-0.0182	0.0984	0.1890*	0.0937
Stockland Green	-0.1466**	-0.4050***	-0.2686**	-0.1826***	-0.1295***	-0.2040***
Sutton Four Oaks	0.1189**	-0.0879	0.0258	0.0998*	-0.0204	0.0198
Sutton New Hall	b	b	b	b	b	b
Sutton Vesey	-0.054	-0.0491	-0.0267	0.0435	-0.0474	-0.0377
Washwood Heath	.	-0.4153***	-0.1691	-0.3485***	-0.1846***	0.0246
Weoley	-0.0086	-0.3608***	0.1447	0.0239	-0.1183**	0.1101

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6
Yardley	-0.1448**	-0.3017***	-0.1195	-0.2531***	-0.1809***	-0.1586***
Environmental Characteristics						
Views of Water	-0.0023	0	0.0003	-0.0029	0.001	0.0004
Views of Parkland	-0.0002	0	0.0001	-0.0001	0	0
Road Traffic Noise	-0.0024	-0.0022**	-0.0052***	-0.0019	-0.0016*	-0.0017
Rail Traffic Noise	-0.0063*	-0.0074**	-0.0055	-0.0128**	-0.0042	-0.0039
Aircraft Noise	0.0092	.	.	-0.0072	-0.0123	-0.0095
K	88	80	91	87	86	101
N	1540	2324	878	1176	3453	1353
R^2	0.760	0.646	0.655	0.686	0.574	0.763
s^2	0.0387	0.0525	0.0772	0.038	0.0355	0.0486

b Base case for a set of dummy variables
* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence