Day, Brett; Bateman, Ian; Lake, Iain

Working Paper

What price peace? A comprehensive approach to the specification and estimation of hedonic housing price models

CSERGE Working Paper EDM, No. 03-08

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 (2003) : What price peace? A comprehensive approach to the specification and estimation of hedonic housing price models, CSERGE Working Paper EDM, No. 03-08, 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/10419/80269

WHAT PRICE PEACE?
A COMPREHENSIVE APPROACH
TO THE SPECIFICATION AND ESTIMATION
OF HEDONIC HOUSING PRICE MODELS

by

Brett Day,
Ian Bateman and Iain Lake

CSERGE Working Paper EDM 03-08

WHAT PRICE PEACE?
A COMPREHENSIVE APPROACH
TO THE SPECIFICATION AND ESTIMATION
OF HEDONIC HOUSING PRICE MODELS

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, Norwich NR4 7TJ

[2]Centre for Environmental Risk,
School of Environmental Sciences
University of East Anglia, Norwich NR4 7TJ

Contact details:
email: brett.day@uea.ac.uk; tel: (44) (0)1603 592064
email: i.bateman@uea.ac.uk; tel: (44) (0)1603 593125

Acknowledgements

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

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).

ISSN 0967-8875

Abstract

Previous work on hedonic price functions tends to have focused on one of a number of specification and estimation issues; namely, market segmentation, choice of functional form, multicollinearity or spatial autocorrelation. The purpose of this paper is to bring together these various strands to provide a comprehensive modelling approach. In particular we use a combination of factor analysis and cluster analysis to define market segments and reduce collinearity in the data. We adopt Robinson's semiparametric specification of the hedonic price function and account for spatial autocorrelation using Kelejian and Prucha's generalized moments estimator. The modelling approach is applied to a large and extremely detailed dataset for the City of Birmingham constructed from multiple data sources and compiled with the use of GIS. The focus of this application is the identification of implicit prices for noise pollution from road, rail and air traffic sources.

Key words: Hedonics, market segmentation, factor analysis, clustering, partially linear model, spatial autocorrelation

1. Introduction

The hedonic price function describes the relationship that exists between a property's characteristics (denoted by the vector of values $\boldsymbol{z}$) and the price at which it sells in the market (which we denote P). The function can be written in the very general form;

$$P = P(\boldsymbol{z}) \tag{1}$$

Clearly, the researcher is faced by a number of important questions in specifying an econometric model that could be used to estimate (1). In brief, we identify four major issues that must be addressed;

Data Issues: Often the objective of a hedonic valuation study is to establish how property prices differ in response to differences in environmental quality. In the Birmingham case study presented here, for example, the objective is to establish the impact of transport related noise pollution. To tease out this relationship requires controlling for the myriad other determinants of property prices. Chief amongst these are size and structural characteristics, though proximity to amenities and the socioeconomic composition of neighbourhoods are also known to be significant determinants of a property's market price[1].

Fortunately, the development of geographically referenced datasets that can be manipulated with GIS (Geographical Information Systems) has greatly enhanced the ability of researchers to generate datasets containing variables that describe the locational characteristics of properties. In contrast to years gone past, the problem faced by researchers is not one of lack of data but one of data surfeit. Section 2 of this paper describes just such a dataset for the City of Birmingham.

Faced by an embarrassment of data riches, regression analysis is confounded by problems of interpretability and collinearity. Section 3 describes the application of factor analytical techniques to the plethora of variables describing the socioeconomic characteristics of property neighbourhoods. In essence, this procedure seeks to identify major dimensions of association between variables such that a smaller set of variables (factors) can be defined that approximate the variation shown in the original data. A number of goals are achieved by this procedure (1) the dimensionality of the regression problem is reduced, (2) the

[1] Whilst it would be inappropriate to list offenders, hedonic valuation studies that fail to control for this myriad potential determinants must answer to the criticism of omitted variable bias.

factors are orthogonal by design such that collinearity amongst the original variables is no longer a problem and (3) the factors describe the fundamental dimensions of difference and similarity underlying the original variables and hence are much easier to interpret in a regression analysis.

Market Segmentation: There are a number of theoretical and empirical reasons why accounting for market segmentation is both a desirable and necessary step in hedonic analysis. In particular, if a house price dataset contains data from more than one market segment then it is likely that the hedonic price functions for each segment are quite different. Failing to differentiate between these different submarkets may seriously bias estimates of the true hedonic price functions. Section 4 of this paper describes the application of a technique known as *Cluster Analysis* that is used to divide the data into various property submarkets.

Functional Form: Economic theory offers the researcher little guidance on the specific functional form of (1). On the whole, the less explicit the researcher is in specifying the relationship between the characteristics data and property prices, the less likely it is that the model will be misspecified. In Section 5 we describe a semiparametric model developed by Robinson (1988) which imposes relatively few assumptions on the relationship between prices and property characteristics.

Spatial Correlation: Despite the best efforts of the researcher, there still exists the possibility that important characteristics may be missing from the data set. For example, the proximity of an abattoir would inevitably deflate the price of neighbouring properties. Unfortunately, our data set does not contain information on the location of abattoirs and as a result the properties in the area will show prices that are lower than we might expect given their other characteristics. The regression error term for these properties will be unusually large and negative. In general, the existence of unmeasured similarities between properties in close proximity will result in correlation of error terms. In Section 6 we describe an estimation technique developed by Kelejian and Prucha (1999) that exploits this spatial correlation in order to improve our estimates of the hedonic price function.

Finally, Section 7 provides details of the regression results, providing estimates of the parameters of the hedonic price function for each of the property submarkets.

2. Data

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[2]. These records indicated selling prices, dates of sales and full property address for each residential property transaction. Initially, each property address was matched to an entry in the Ordnance Survey ADDRESS-POINT database providing a unique grid reference for each postal address in the UK (Ordnance Survey; 1996). Subsequently, each property grid reference was matched with a building outline on OS Land-Line.Plus (Ordnance Survey, 1996).

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.[3] 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.

[2] 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.

[3] 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.

The rest of the data set was constructed with the aid of GIS. OS Land-Line.Plus provided details of the garden area (plot area minus building area) and aspect of each property. Various data sources were used to identify (dis)amenities including schools, shops[4], railway stations and industrial sites. These were located using OS ADDRESS-POINT and straight line distances, car travel times and walking distances calculated from each property to each (dis)amenity using OS Land-Line.Plus. When considering the accessibility of properties to shops, any measure based on proximity to only one facility might have 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:

$$A_i = \sum_{j=1}^{J} \alpha_j e^{-\delta d_{ij}} \qquad (2)$$

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.

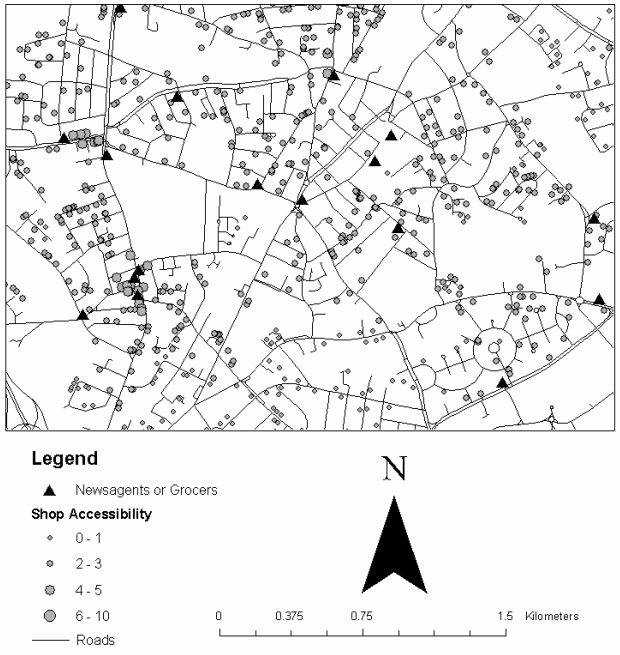
A similar procedure was used when considering accessibility to primary schools. Recent research suggests that selection procedures for primary school intake that favour local residents can considerably inflate house prices around high performing schools (Gibbons and Machin, 2001).[5] For each primary 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

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

[5] 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***)

and English (the level expected of 11 year olds).[6] A primary school accessibility index was constructed using (2) with the weight α_j set to this measure of school quality and $\delta = 1$.

Figure 1: Shop accessibility scores for a selection of properties in the data set



Data on the socio-economic composition of property neighbourhoods were drawn from the 1991 UK census provided by the Office for National Statistics (ONS). To preserve confidentiality census data is published as averages over small areas known as enumeration districts (EDs) (Openshaw, 1995). In Birmingham, on average, this consists of data for 191 households. 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

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

are ideal for constructing indicators of the attributes of the neighbourhood in which a property is located.

Using a procedure outlined in (Lake *et al*., 2000) data on land uses and the location and orientation of each property (taken from OS.Land-Line Plus) was combined with information on the landscape topology (extracted from OS.Land-Form PROFILE) 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). This project produced a traffic flow model predicting the number of vehicles travelling along the city's roads from data on the frequency and destination of vehicle journeys made in the Birmingham urban area. Naturally, the use of a traffic flow model introduces an element of uncertainty into the accuracy of the data for a property's road traffic noise exposure. We assume that mismeasurement, if it exists, is random but will account for this using instrumental variable estimation techniques. 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}.

A complete description and list of the variables used in the hedonic analysis are provided in Appendix A. Complete data records were successfully compiled for some 10,889 residential property transactions in Birmingham in 1997.

3. Defining Neighbourhood Characteristics

The properties in the study area came from 3,005 different EDs in Birmingham, and for each ED our dataset contained details of some 18 neighbourhood attributes. These are listed and described in the first two columns of table 2. Not surprisingly, many of these attributes are highly collinear. For example, the percentage of households not owning cars exhibits high positive correlation with the percentage of households that do not own their own property (correlation coefficient of .70) and high negative correlation with the percentage of households that own two cars (correlation coefficient of -.76).

Whilst each of these neighbourhood attributes might have a bearing on property prices, the presence of such collinearity creates a problem for researchers. As is well known, parameters estimated on highly collinear regressors are difficult to interpret. Parameter estimates may have implausible magnitude or, in the worst case, the wrong sign. Interpretation is further confounded by the fact that individual parameters may exhibit high standard errors and consequently low significance levels.

Moreover, it is not clear that each of these neighbourhood attributes will be independently capitalised into the property market. More likely, households in a market will consider more general indications of the characteristics neighbourhood of a property, the wealth of the area, its ethnic composition, the stage of life of its inhabitants etc.

As a result, we condensed the excess of neighbourhood attributes into a more manageable set of indices. Each index picked out a major dimension of difference or similarity between property neighbourhoods. For example one index indicated the wealth of a neighbourhood, effectively combining the myriad attributes that are indicators of wealth/poverty into one dimension. Subsequently, property neighbourhoods were scored along each dimension. In our example, poor neighbourhoods generated low scores on the wealth dimension, whilst affluent neighbourhoods generated high scores. The procedure by which dimensions are identified and property neighbourhoods are scored along these dimensions is known as *factor analysis*.

We do not intend presenting the intricacies of factor analysis here. For a highly accessible text on the subject see Lindeman *et al*., (1980). In essence, the procedure seeks to identify major dimensions of association between variables (in our case the attributes of neighbourhoods) such that a smaller set of variables can be defined that approximate the variation shown in the original data. These dimensions are called factors and one can define as many factors as there are variables in the original data.

Table 1 details the first eight factors for the neighbourhood attribute data. The second column in this table provides the eigenvalue of each factor. The third column indicates the percentage of the variation in the data explained exclusively by that factor. If the attributes were not correlated then each factor would explain $1/M^{th}$ of the variation (where M is the number of attributes in the factor analysis) and each eigenvalue would take a value of one. If the attributes were all perfectly correlated then the first factor would explain 100% of the variation with an eigenvalue of M. The fourth column provides the cumulative sum of this explained variation.

From Table 1 it can be seen that the first factor alone explains over 40% of the variation in the neighbourhood attribute data. This indicates that many of the attributes are highly correlated (positively or negatively) with a single underlying factor. Notice that successive factors explain progressively less of the remaining variation.

Table 1: Variation explained by the first ten factors of the Enumeration District neighbourhood attributes (estimated using Iterated Principal Factors)

Factor	Eigenvalue	Variation Explained by Factor	Cumulative Explained Variation
1	7.28	.43	.44
2	3.24	.20	.63
3	1.79	.11	.74
4	1.21	.25	.81
5	0.95	.30	.87
6	0.65	.11	.91
7	0.55	.15	.94
8	0.39	.17	.97

The question now is which of the factors should be taken as capturing the main dimensions of difference and similarity expressed in the neighbourhood attribute data. A good rule of thumb is to ignore factors with eigenvalues less than one as these dimensions explain less of the variation in the data than the dimensions defined by the original attributes themselves. This procedure leads

us to focus on the first 4 factors. As such, our constructed indices explain around 81% of the variation in the neighbourhood attribute data.

One of the arts of factor analysis is the interpretation of factors. Interpretation of factors is the process of describing the underlying dimension of similarity or difference between the neighbourhood attributes captured by a factor. Table 2 describes the (orthogonally rotated) loadings of the first four factors. A large positive loading indicates that high values of the original attribute are associated with high values of the factor. Similarly a large negative loading indicates that high values of the original attribute are associated with low values of the factor. The loadings in Table 2 suggest fairly obvious interpretations for the four factors. We suggest the following;

Factor 1: Wealth

This factor is very distinct and describes the general level of wealth of neighbourhoods. Not surprisingly, the factor is highly positively correlated with car ownership and being in the maximum earning bracket age range between 35 and 49. Conversely the factor is strongly negatively associated with lack of access to cars, unemployment, low home ownership and one parent families.

Factor 2: Ethnicity

The second factor also has a clear interpretation. This factor loads heavily on four attributes; the three attributes describing the ethnic composition of neighbourhoods and the attribute describing the degree of over-crowding in households. Since the loadings on all these four attributes are positively signed, high scores on this dimension reflect the increasing presence of members of the ethnic minorities in neighbourhoods.

Factor 3: Adult Age Composition

This fourth factor picks out a dimension defining the age composition of neighbourhoods. It loads negatively on young adults that is those in the age groups 18 to 24 and 25 to 34 but loads positively on adults in older generations, that is, age groups of 50 to 64 and older than 65. EDs scoring highly on this factor will be characterised by neighbourhoods with relatively older adult populations.

Factor 4: Family Composition

The final factor loads heavily on just three attributes those describing the percentage of households with children and the percentage of the ED populations in age groups 0 to 10 and 11 to 17. EDs that score highly on this factor are characterised by having a relatively large number of households with children. Notice the distinction here with the composition of adult ages

as described by the third factor. Clearly, it is possible to have EDs exhibiting the same distribution of adult age ranges but which differ according to the degree to which those adults have children.

Table 2: Neighbourhood attributes and rotated factor loadings

Attribute	Attribute Description	Factor 1	Factor 2	Factor 3	Factor 4
No car	% households with no access to a car	-0.96	+0.15	-0.07	+0.09
Two cars	% two-car households	+0.85	-0.21	+0.09	+0.02
Unemployment	% working age residents uemployed	-0.79	+0.35	-0.13	+0.19
Non-owners	% residents not owning their home	-0.89	-0.08	-0.03	+0.05
One-parent families	% lone parent households	-0.72	-0.10	-0.27	+0.38
Low Social Class	% residents in lower social classes	-0.44	+0.17	-0.02	+0.09
Families	% households with children	-0.10	+0.38	+0.12	+0.80
Age 0 to 10	% residents under 10 years old	-0.39	+0.30	-0.33	+0.74
Age 11 to 17	% residents aged 11 to 17	+0.18	+0.54	+0.16	+0.68
Age 18 to 24	% residents aged 18 to 24	-0.30	+0.33	-0.62	-0.16
Age 25 to 34	% residents aged 25 to 34	-0.15	-0.06	-0.85	-0.09
Age 35 to 49	% residents aged 35 to 49	+0.80	-0.30	+0.00	+0.11
Age 50 to 64	% residents aged 50 to 64	+0.27	-0.04	+0.57	-0.43
Age > 65	% residents over the age of 65	-0.22	-0.35	+0.64	-0.48
Over Crowding	% households with > 1 person per room	-0.30	+0.76	-0.04	+0.37
Non White	% ethnically non-white residents	-0.23	+0.92	-0.14	+0.19
Black	% ethnically black (African or Caribbean) residents	-0.48	+0.40	-0.29	+0.10
Asian	% ethnically Asian residents	-0.08	+0.94	-0.04	+0.20

The final step in a factor analysis is to define a score for each ED for each factor. Using the factor loadings a regression-like equation is calculated, the parameters of which indicate how greatly each attribute contributes to each factor. Given the attributes of each neighbourhood, the equation can be used to determine how highly a neighbourhood scores on each factor. In effect, neighbourhoods that exhibit high values for attributes that load positively on a factor receive high scores for that factor whilst neighbourhoods that that exhibit high values for attributes that load negatively on that factor receive low scores.

As shall be described subsequently, the factors can be used as proxies for the original attributes in regression analysis. Further, in the next section we employ the factors as indicators of the socioeconomic characteristics of property neighbourhoods in order to identify property submarkets.

As has been demonstrated the factors capture a good proportion of the variation shown in the original neighbourhood attributes. Moreover, the nature of their construction ensures that the factor scores are orthogonal overcoming the problem of collinearity in the original set of attributes.

4. Market Segmentation

A housing market will be typified by a unique hedonic price schedule determined by the particular characteristics of the households and housing stock that make up that market (for a more detailed discussion see Day, 2001). Property prices in two different markets may be determined by very different hedonic price functions. Hence a primary concern for hedonic researchers is to ensure that data are drawn from a single property market.

Whilst many hedonic researchers have taken data from a single urban conurbation as representing data from one market, there is much evidence to suggest that property markets are segmented within one urban area. Indeed, property valuation experts from the VOA in Birmingham suggested that a number of such market segments are identifiable in the City of Birmingham.

Various circumstances may precipitate segmentation of the property market. Straszheim (1975, p.28) states that "variation in housing characteristics and prices by location is a fundamental characteristic of the urban housing market". Indeed, geographic location may well define market segments; consider how important a property's postcode can be in determining its market price.

Furthermore, other characteristics of properties might drive or contribute to market segmentation. Schnare and Struyk (1976) argue that segmentation will result whenever household's demand for a particular locational, structural or neighbourhood characteristic is highly inelastic and when this preference is shared by a relatively large number of other households. Basu and Thibodeau (1998) identify a number of dimensions that might characterise market segmentation;

- s*tructure type*: households may wish to purchase a property of a certain type. For example, the market might segment between households looking to purchase houses with gardens and those looking to purchase flats or maisonettes.

- *structural characteristics*: households may have strong preferences for a particular property characteristic. For example, segmentation might result if certain households only consider buying period properties with "original features" whilst others only consider purchasing modern homes.

- *Neighourhood characteristics*: households may have strong preferences for localities providing certain amenities. For example, certain households may desire proximity to transport links or good quality schooling whilst others find no advantage in such proximity. Similarly, households may segment along income or racial lines particularly if households prefer to live in areas of relatively homogenous socio-economic characteristics.

Since, segmentation seems likely to pervade property markets, it is fortunate that statistical techniques are available to test for segmentation. Put simply, if it is suspected that the house prices used in a hedonic study may come from a segmented market then rather than estimating one hedonic price function, researchers can estimate a separate function for each suspected market segment. It is possible to test whether the separate functions are sufficiently similar to count as one market or whether they are significantly different and should be treated separately.

Evidence from the hedonic literature using this sort of test has returned ambivalent results. For example Butler (1980) tested to see whether a national housing market existed by comparing data from 36 cities in the USA. Though he concluded that the market in the sale and purchase of houses could not be considered a single market, he found it impossible to reject the possibility that the house rental market was a single national market. Smith and Huang (1995) surveyed hedonic pricing studies carried out between 1967 and 1988 and concluded that the estimated hedonic price functions differ across cities due to differences in local conditions. Other researchers have investigated the possibility that segmentation exists in the housing market within a single urban area. Straszheim (1974), for example, found that geographical segmentation was a feature of the housing market in San Francisco. On the other hand, Ball and Kirwan (1977) found that clusters of different housing types in the Bristol area did not result in separate submarkets with different hedonic prices.

Other studies include those by Straszheim (1973), Schnare and Struyk (1976), Sonstelie and Portney (1980), Goodman (1978) Michaels and Smith (1990) and Allen *et al*., (1995). Each of these studies have applied different rules by which properties in an urban area are allotted to a particular submarket. Criteria include, locational or political boundaries, characteristics of households (e.g income and race), property types and classifications based upon the judgement of estate agents. Here we suggest an approach that makes no *a priori* assumptions concerning the criteria defining submarkets, rather the data itself is used to suggest the pattern of market segmentation.

The procedure used to group properties into submarkets is known as *cluster analysis*. Cluster analysis divides a dataset into groups (clusters) of observations that are similar to each other. There are two basic approaches to cluster analysis, *partitioning methods* and *hierarchical methods*. With both methods, the researcher determines the *P* characteristics that are to be used to cluster the

observations. Here we follow the reasoning of previous researchers and choose characteristics reflecting three criteria that might generate segmentation of properties into submarkets;

- *Geographic Location* as defined by each properties grid reference (longitude and latitude)
- *Property Type* as defined by each properties floor and garden area
- *Socioeconomic Characteristics of Neighbourhoods* as defined by the four factors indicating neighbourhoods relative wealth, ethnicity, adult age composition and family composition

Each observation then can be plotted in *P*-space according to how highly it scores on each of these *P* characteristics. Clearly, observations holding similar values for the different characteristics will be located close to each other in this *P*-space.

With partitioning methods the researcher decides upon the number of clusters *a priori*. Let us denote this number of clusters *k*. The partitioning algorithm seeks to find *k* locations in *P*-space, known as medoids, such that the sum of the distances between each observation and its nearest medoid is minimised. Once the *k* medoids have been determined the observations are partitioned into clusters by assigning each observation to its nearest medoid.

Hierarchical methods work in a somewhat different manner. With a bottom-up approach, each observation is initially considered as a small cluster by itself. As a first step the two observations lying closest together are merged into a new cluster. At each subsequent step, the two nearest clusters are combined to form one larger cluster. Clusters are merged until one large cluster remains containing all the observations. The final result is a hierarchy of association appearing much like an inverted tree. The tree can be plotted in order to determine which branches of the hierarchy should be treated as separate clusters.

The advantage of hierarchical methods is that they do not impose any *a priori* assumptions on the pattern of association in the observations. The drawback with these methods, however, is that they are computationally burdensome with large data sets. Indeed, in this case, the sheer size of the data set precludes the use of hierarchical cluster analysis.[7]

[7] Though current research by the authors is investigating recent developments in model-based hierarchical clustering (Posse, 2001) that may allow their application to this dataset.

As a result, the researchers have been forced to determine the number of clusters *a priori*. The choice of the number of clusters has been decided through an iterative investigation of the data. Initially advice was sort from the valuation officers at the VOA who confirmed the researchers suspicions that the property market in the City of Birmingham contained a number of submarkets defined by location, property type and socioeconomic characteristics. Using cluster analysis various partitions of the properties in the sample have been attempted ranging from four submarkets (the number identified by the VOA as the minimum number of submarkets necessary to reflect the complexity of the property market in Birmingham) through to ten submarkets. The results presented here are for an eight submarket partitioning of the data. This division resulted in submarkets that were readily interpretable, conformed to the Valuations Officers understanding of the Birmingham property market and returned hedonic price functions with parameters that conformed to a priori expectations.

Tables 3 and 4 present summary statistics that can be used to compare the characteristics of the properties in the eight submarkets. Figure 2 plots the spatial location of properties in the different submarkets. We use the data and maps to interpret and describe the eight submarkets.

Table 3: Average attribute values for eight submarket division of Birmingham property market (structural characteristics)

Sub-market	Price (£)	Floor Area (m^2)	Garden Area (m^2)	Property Age (years before 1997)	% Detached Houses	% Terraced Houses
1	35,976	106	117	88	0.03	0.76
2	54,151	98	128	74	0.05	0.62
3	93,154	114	361	46	0.38	0.03
4	49,238	93	180	66	0.05	0.33
5	73,017	107	263	59	0.18	0.18
6	51,471	94	197	58	0.05	0.30
7	43,138	91	197	62	0.02	0.37
8	175,391	216	973	71	0.68	0.03
Total	59,160	103	227	65	0.12	0.35

Table 4: Average attribute values for eight submarket division of Birmingham property market (socioeconomic characteristics)

Sub-market	Wealth Factor	Ethnicity Factor	Adult Age-Composition Factor	Family-Composition Factor
1	-0.28	2.32	-0.01	0.69
2	0.08	-0.01	-1.69	-1.13
3	1.27	-0.41	0.19	0.44
4	0.33	-0.14	-0.52	-0.39
5	0.96	-0.29	-0.12	0.29
6	-0.15	-0.52	0.28	-0.03
7	-0.56	-0.43	0.38	0.23
8	0.77	0.04	0.34	-0.27
Total	-0.21	0.04	-0.14	0.03

Submarket 1: Ethnic, inner-city
Submarket 1 is concentrated in the inner-city, forming a distinct ring surrounding Birmingham City centre. Consisting mostly of turn of the century terraced houses, the defining feature of this submarket is the high concentration of residents from the ethnic minorities. Perhaps unsurprisingly, this submarket is also characterised by relative poverty, a wide range of adult ages and a high level of households with children.

Submarket 2: Young, first-time buyers without children
In the main, properties in submarket 2 are concentrated in a band to the south of the city centre, located relatively close to the inner-city with an especially strong concentration around the University and Hospital complex. Properties are similar in size and type to those in submarket 1 but on average command a selling price almost £20,000 greater. The defining feature of this submarket is that it comprises neighbourhoods inhabited by young adults without children.

Submarket 3: Northern Suburbs, Affluent
Properties in this submarket are found in the north and western city suburbs mostly in the sort-after Sutton Coldfield area of the City. Properties in this submarket are amongst the most expensive in Birmingham (average price £93,000), they are large mostly detached (38%) or semi-detached with

expansive gardens. Not surprisingly, the submarket is characterised by very wealthy households, with relatively older, white adults, many with families.

Submarket 4: Northern and Western Suburbs, Standard
Submarket 4 is also located in the northern and western suburbs of Birmingham. Unlike submarket 3, however, few properties are located in the Sutton area and the properties and their gardens tend to be of a somewhat more standard size. Likewise, residents are moderately wealthy though they tend to be relatively young and many are without children.

Submarket 5: Southern Suburbs, Affluent
Properties in submarket 5 are located in the Southern suburbs of Birmingham. Whilst geographically distinct this submarket shows many similarities to submarket 3 boasting relatively large properties with substantial gardens. Residents are wealthy, generally white and many have families. Interestingly, properties in the southern market are significantly cheaper than their northern counterparts. In a similar vein, the residents are somewhat less wealthy and tend to be somewhat younger.

Submarket 6: Southern Suburbs, Poor
Properties in submarket 6 share much of the geographical range of those in submarket 5. In contrast to properties in that market, however, houses are relatively small with small gardens and sell at significantly lower prices. Indeed, many of the properties in this submarket were originally built as council housing. The submarket is characterised by mainly white and relatively poor residents.

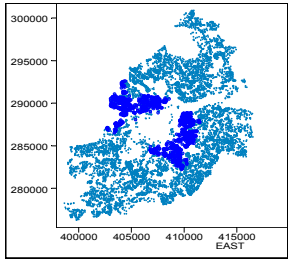
Submarket 7: Northern and Western Suburbs, Poor
Submarket 7 shares the same geographical range as the submarket 4. Indeed, structurally, properties in the two submarkets are relatively similar. Socio-economically, however, the two submarkets are quite distinct. Residents of submarket 7 are the least wealthy in Birmingham and the most exclusively white. Households tend to be relatively young with families. Much of the property in this submarket was constructed as council housing and bears close resemblance to submarket 6, its southern equivalent. However, properties in this submarket sell for significantly lower prices. Indeed, discussion with valuation officers at the VOA confirmed that the southern part of the city is considered a leafier, more desirable location than that occupied by equivalent properties in submarket 7.

Submarket 8: Very Large Properties
Geographically Submarket 8 is the least clearly defined of the eight submarkets. There are two main concentrations of properties in this submarket, one around

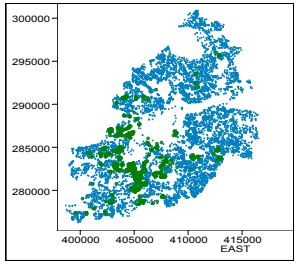
the fringes of Sutton Coldfield park in the north of the city and the other in a swathe to the south of the city located in wards such as Edgbaston and Harbourne. In contrast, structurally and socioeconomically this is the most clearly defined submarket. Properties in this submarket command the very highest prices in Birmingham (average £175,000) reflecting their size (average floor area 216m^2) large gardens (average 973m^2) and desirable location. Not surprisingly, residents tend to be wealthy, somewhat older than the Birmingham average with fewer children.

Figure 2: Locations of Properties in Different Submarkets

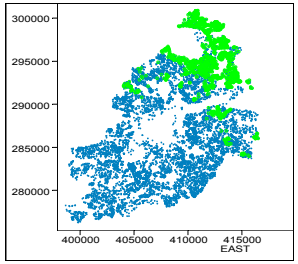
Submarket 1: Poor Ethnic Inner City



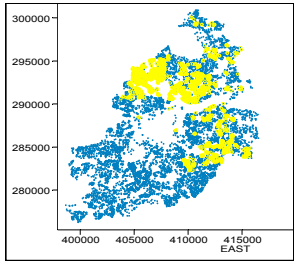
Submarket 2: First-Time Buyers



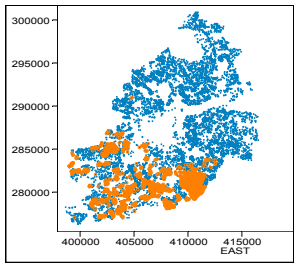
Submarket 3: Affluent Northern Suburbs



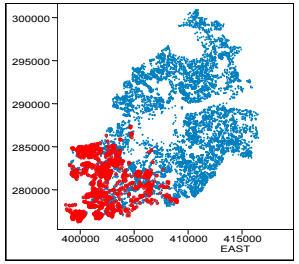
Submarket 4: Standard North & West



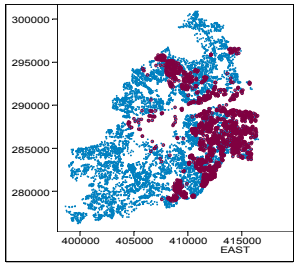
Submarket 5: Affluent Southern Suburbs



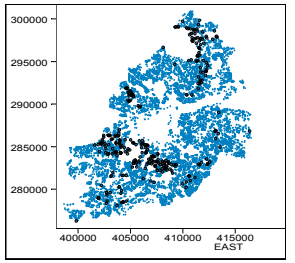
Submarket 6: Poor Southern Suburbs



Submarket 7: Poor North & West



Submarket 8: Very Large Properties



5. Functional Form

The objective of the empirical investigation of the Birmingham data set is to provide an empirical estimate of the hedonic price function $P = P(z)$ (1). To do this it is necessary to define a regression equation that, in its most general form, can be written;

$$g(P_i) = h(z_i) + \varepsilon_i \qquad (2)$$

where P_i is the price of observation i, z_i is a row vector of associated property characteristics and ε_i is an observation specific error term. The two functions $g(\cdot)$ and $h(\cdot)$ determine the exact nature of the relationship between the dependent variable (P_i) and explanatory variables (z_i). To avoid complicating the notation unduly we ignore the fact that a separate regression equation must be estimated for each submarket.

Our first task in estimating the parameters of the hedonic price function is to specify functional forms for $g(\cdot)$ and $h(\cdot)$. Unfortunately, economic theory provides little guidance on the nature of the relationship between property prices and property characteristics. Indeed, the choice of functional form for the hedonic price function has been an issue of some debate.

In general, researchers have opted to define $g(\cdot)$ as the log transformation. We follow that convention here. That is, our empirical model regresses the natural logarithm of property price on some function of the explanatory variables. Using the log of property price has at least two advantages.

- First, the distribution of property prices in a market tends to show considerable right skew. Such data distributions are often associated with heteroskedasticity and/or non-normality of errors, both of which complicate estimation.

- Second, using a log transformation allows for readily interpretable coefficient estimates. For example, the coefficient on a regressor entered in simple linear form indicates the constant percentage response in property price to a unit increase in the regressor

The regression model (2) can be rewritten as;

$$\ln P_i = h(z_i) + \varepsilon_i \qquad (3)$$

Estimating a model such as (3) often requires the imposition of strong assumptions on the function $h(\cdot)$. For example, many studies assume that this is a linear function, giving rise to the familiar model;

$$\ln P_i = z_i\boldsymbol{\gamma} + \varepsilon_i \qquad (4)$$

where $\boldsymbol{\gamma}$ is a column vector of parameters.

Of course, little in economic theory would suggest that such a strong assumption is valid. As a result, considerable attention has been focused on the use of more flexible specifications. In particular, a number of researchers have investigated the use of the Box-Cox flexible functional form (e.g. Cropper, Deck and McConnell, 1988; Cheshire and Sheppard, 1998). Whilst, this approach allows the regression model to more accurately reflect the patterns of association inherent in the data, it also has a number of drawbacks (as discussed by Ramussen and Zuehlke; 1990).

An alternative to increasing the degree of parameterisation of the regression model is to adopt a non-parametric regression approach. Here, the function $h(\cdot)$ is dictated entirely by the data ensuring the regression function is extremely robust to misspecification. Unfortunately, non-parametric estimation of $h(\cdot)$ is only realistic when there are only a small number of regressors in z_i. When there are many regressors, non-parametric response coefficients may be very imprecise.

An intermediate strategy is to employ a semiparametric form such as that proposed by Robinson (1988). Here part of the model is specified parametrically whilst the rest is estimated using non-parametric techniques. Robinson's model is of the form

$$\ln P_i = z_i\boldsymbol{\beta} + q(\boldsymbol{x}_i) + \varepsilon_i \qquad (5)$$

where z_i is a k-vector of regressors associated with a k-vector of parameters $\boldsymbol{\beta}$, whilst $\boldsymbol{x}_i$ is a p-vector of regressors whose influence on property prices is determined by the unknown function $q(\cdot)$.

Robinson shows that the model in Equation 5 can be rewritten as;

$$\ln P_i - E[\ln P \mid \boldsymbol{x}_i] = (z_i - E[z \mid \boldsymbol{x}_i])\boldsymbol{\beta} + \varepsilon_i \qquad (6)$$

suggesting that $\boldsymbol{\beta}$ can be estimated in a two-step procedure;

- First, the unknown conditional means $E[\ln P \mid \boldsymbol{x}_i]$ and $E[z \mid \boldsymbol{x}_i]$ are estimated using a non-parametric estimation technique.

- Second, the estimates are substituted in place of the unknown functions in Equation (6) and ordinary regression techniques employed to estimate $\boldsymbol{\beta}$.

Indeed, Robinson shows that the resulting parameter estimates are asymptotically equivalent to those that would be derived if the true functional form of $q(\cdot)$ were known and could be used in the estimation.

Robinson's model was pioneered in the hedonic literature by Anglin and Gencay (1996) and has recently been employed by Gibbons and Machin (2002) and Gibbons (2002).

In this case, the semiparametric specification has a number of advantages.

- First, within any one submarket, the structural characteristics of a property are likely to be most important in determining a properties market price. The hedonic price model would be considerably more robust if these variables could be included in the unknown function $q(\cdot)$. Referring to the table of regressors in Appendix A see that the majority of these characteristics are defined as dummy variables. As Anglin and Gencay (1996) point out the inclusion of these dummy variables in $q(\cdot)$ would effect the scale but not the curvature of that function. A reasonable approximation, therefore, would be to include these dummy variables in the linear part of the model. Of the remaining structural variables, those defining a property's floor area, garden area and age, provide a reasonably accurate picture of a properties structure (the age of a property proxying for both quality characteristics and architectural design).

- Second, the variables of interest in this research project are those describing a property's exposure to noise pollution. Including these in the linear part of the model allows for ease of interpretation of parameter estimates, simplifies the calculation of implicit prices and facilitates comparison of estimates with other studies.

Accordingly the vector $\boldsymbol{x}$ entering the unknown function $q(\cdot)$ consists of the log of a property's floor area, the log of a property's garden area and the property's age. Further, property prices in the UK have been reasonably volatile over recent decades. To account for price movements over the course of 1997 a continuous variable indicating the date of the property sale is included in the function $q(\cdot)$. The researchers believe that $\boldsymbol{x}$ is of sufficient dimension to capture

many of the important determinants of property price whilst not compromising the ability of non-parametric regression to return accurate estimates of the function $q(\cdot)$. The choice of variables to be included in the nonparametric function $q(\cdot)$ is summarised in Table 5.

Table 5: Variables included in nonparametric part of the hedonic price regression

Variable	Description
Area	Natural logarithm of property floor area in m^2
Garden	Natural logarithm of garden size in m^2
Age	Age of property in decades from 1997
Sale Date	Date of sale in days from 1st January 1997

Following Anglin and Gencay (1996) $E[\ln P \mid \boldsymbol{x}_i]$ and $E[z \mid \boldsymbol{x}_i]$ are estimated using non-parametric kernel regression.

The kernel estimator of the density of the random vector $\boldsymbol{x}$ evaluated at $\boldsymbol{x}_i$ is given by;

$$\hat{f}_H(\boldsymbol{x}_i) = \frac{1}{N}\sum_{j=1}^{N} K_H(\boldsymbol{x}_j - \boldsymbol{x}_i) \qquad (7)$$

where $K_H(\boldsymbol{x}_j - \boldsymbol{x}_i) = \det(H)^{-1}.K\left(H^{-1}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right)$ for some multivariate kernel function $K(u)$ and for a given P x P vector of bandwidths, H.

In effect, the kernel density estimator counts the number of observations in the dataset in close proximity to $\boldsymbol{x}_i$. The density at $\boldsymbol{x}_i$ is approximated by dividing this count by the number of observations in the dataset. Whether observations $\boldsymbol{x}_j$ are considered close to $\boldsymbol{x}_i$ is determined by the bandwidth matrix H. The larger the elements of the bandwidth matrix, the more observations are drawn into the count. Further the weight allotted to each observation in the count is determined by the kernel function $K(u)$. The kernel function must be symmetric, continuously differentiable and integrates to unity. Moreover, most commonly used kernel functions allot greater weight to observations in close proximity to $\boldsymbol{x}_i$ than those further away.

Here we use a matrix of bandwidths determined by $\boldsymbol{S}$, the sample covariance matrix of $\boldsymbol{x}$, such that;

$$\boldsymbol{H} = h\boldsymbol{S}^{\frac{1}{2}} \tag{8}$$

for some positive scalar h. In this case, the argument to the kernel function can be written;

$$\boldsymbol{u} = h^{-1}\,\boldsymbol{S}^{-1/2}\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right) \tag{9}$$

Further, we employ a multivariate Gaussian Kernel of the form;

$$K(\boldsymbol{u}) = (2\pi)^{-P/2}\exp\left(-\frac{1}{2}\boldsymbol{u}'\boldsymbol{u}\right) \tag{10}$$

As such the kernel density estimator of equation (7) can be written in the specific form;

$$\hat{f}_h(\boldsymbol{x}_i) = \frac{1}{N}\sum_{j=1}^{N} h^{-P}\det(\boldsymbol{S})^{-1/2}\,(2\pi)^{-P/2}\exp\left(-\frac{1}{2}\boldsymbol{u}'\boldsymbol{u}\right) \tag{11}$$

To generalise notation, let r represent the element whose conditional expectation we wish to estimate. In our case, therefore, r denotes any element of the $\boldsymbol{z}$ vector or the log of property price. Then the conditional expectations we wish to estimate are given by;

$$E[r \mid \boldsymbol{x}_i] = \frac{\int r\, f(r, \boldsymbol{x}_i)\,dy}{\int f(r, \boldsymbol{x}_i)\,dy} \tag{12}$$

Nadaraya-Watson kernel regression estimates (12) by replacing the numerator and denominator with their equivalent kernel density according to;

$$\hat{E}_h[r \mid \boldsymbol{x}_i] = \frac{\sum_{j=1}^{N} r\, K_h\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right)}{\hat{f}_h(\boldsymbol{x}_i)} \tag{13}$$

A central issue in nonparametric estimation is the choice of bandwidth, h. The bandwidth parameter determines the degree of smoothing of the function

$\hat{E}_h[r \mid \boldsymbol{x}_i]$. Too large a value for h induces bias and too small a value results in imprecise estimates.

Once again, following Anglin and Gencay (1998) we select bandwidths using a data driven technique known as cross-validation. As they point out, a seemingly natural way to select h is to choose the bandwidth that minimises the sum of squared residuals from the regression equation;

$$MSE = n^{-1}\sum_{i=1}^{N}\left(\ln P_i - \boldsymbol{z}_i\boldsymbol{\beta} - q_h(\boldsymbol{x}_i)\right)^2 \tag{14}$$

where the estimator of $q_h(\boldsymbol{x}_i)$ is given by;

$$\hat{q}_h(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{N}\left(\ln P_i - \boldsymbol{z}_i\hat{\boldsymbol{\beta}}\right)K_h\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right)}{\hat{f}_h(\boldsymbol{x}_i)} \tag{15}$$

Unfortunately, this procedure falls down because the objective function, MSE, reduces to zero for any h smaller than the closest two data points in the sample. For such values of h the conditional mean function given by $q_h(\boldsymbol{x}_i)$ puts all weight on the ith observation such that $q_h(\boldsymbol{x}_i)$ perfectly predicts $\ln P_i$.

Accordingly, the criterion function in (15) cannot be used to decide upon the optimal bandwidth. Rather researchers employ the cross-validation statistic;

$$MSE_{CV} = n^{-1}\sum_{i=1}^{N}\left(\ln P_i - \boldsymbol{z}_i\boldsymbol{\beta} - q_{h,i}(\boldsymbol{x}_i)\right)^2 \tag{16}$$

The cross-validation statistic avoids the problems of the raw MSE statistic by employing a conditional mean function $q_{h,i}(\boldsymbol{x}_i)$ that is calculated by leaving out the i^{th} observation;

$$\hat{q}_{h,i}(\boldsymbol{x}_i) = \frac{\sum_{j\neq i}^{N}\left(\ln P_i - \boldsymbol{z}_i\hat{\boldsymbol{\beta}}\right)K_h\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right)}{\sum_{j\neq i}^{N} K_h\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right)} \tag{17}$$

The cross-validation procedure, therefore, is to carry out a grid search for optimal h. The regression equation (6) is re-estimated numerous times using

different values of h. For each value of h the cross-validation statistic is estimated using (16) and the h providing the minimum value for this statistic is chosen as the optimal bandwidth.

Here we improve on the estimation procedure of Anglin and Gencay (1998) by using an adaptive kernel estimator. The motivation behind the adaptive kernel estimator is to improve estimation of the conditional expectation functions $E[\ln P \mid \boldsymbol{x}_i]$ and $E[\boldsymbol{z} \mid \boldsymbol{x}_i]$ by allowing the bandwidth to vary with the density of $\boldsymbol{x}$. Thus where data is relatively sparse the adaptive kernel uses a relatively wide bandwidth, whilst when data is abundant the bandwidth is commensurately reduced.

Adaptive kernel estimation requires a two-stage estimation procedure. First a pilot bandwidth, h_p, is employed to estimate the density of $\boldsymbol{x}$; $\hat{f}_{h_p}(\boldsymbol{x})$. Using this estimated density we calculate;

$$\lambda_i = \left(\frac{\hat{f}_{h_p}(\boldsymbol{x}_i)}{\eta} \right)^{-\rho} \tag{18}$$

where η is a normalisation factor given by $\ln \eta = \sum_j \ln \hat{f}_{h_p}(\boldsymbol{x}_j) \big/ n$ and ρ is a parameter taking a value between 0 and 1 chosen by the researcher. Here we select a value for ρ of 0.25.

In the second stage, the adaptive kernel generalises (13) by using a new observation specific bandwidth parameter $h_j = h\lambda_j$ to estimate the conditional expectations.

Given the cross-validated, adaptive kernel estimates of $E[\ln P \mid \boldsymbol{x}_i]$ and $E[\boldsymbol{z} \mid \boldsymbol{x}_i]$, we are still left with the task of estimating the semiparametric model;

$$\ln P_i - E[\ln P \mid \boldsymbol{x}_i] = (\boldsymbol{z}_i - E[\boldsymbol{z} \mid \boldsymbol{x}_i])\boldsymbol{\beta} + \varepsilon_i \tag{6}$$

To maintain clarity, let us introduce some new notation. Let tilde indicate differences from nonparametric expectations, such that;

$$\tilde{y}_i = \ln \tilde{P}_i = \ln P_i - E[\ln P \mid \boldsymbol{x}_i]$$

and

$$\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - E[\boldsymbol{z} \mid \boldsymbol{x}_i].$$

Consequently, Equation (6) simplifies to;

$$\tilde{y}_i = \tilde{\boldsymbol{z}}_i \boldsymbol{\beta} + \varepsilon_i \tag{19}$$

One possibility is to estimate (19) using ordinary least squares (OLS) according to the familiar formula;

$$\boldsymbol{\beta}^{OLS} = \left(\tilde{\boldsymbol{Z}}' \tilde{\boldsymbol{Z}} \right)^{-1} \tilde{\boldsymbol{Z}}' \tilde{\boldsymbol{Y}} \tag{20}$$

where $\tilde{\boldsymbol{Z}}$ is the N x k matrix of data formed by stacking the $\tilde{\boldsymbol{z}}_i$ vectors and $\tilde{\boldsymbol{Y}}$ is the N x 1 vector with elements $\tilde{y}_i$.

Mention has already been made of the suspicion that the road noise variable is measured with error. If this is the case then the $\boldsymbol{\beta}^{OLS}$ will be biased and inconsistent (for a more detailed exposition see, for example, Davidson and MacKinnon; 1993). Typically, the parameter on the mismeasured variable will be biased towards zero and the other coefficients will be biased in unknown directions.

The most general technique for handling such situations is the method of *instrumental variables* (IV). The fundamental ingredient of any IV procedure is a matrix of instrumental variables. We shall denote this N x l matrix of variables by $\boldsymbol{M}$. Crucially, each variable in $\boldsymbol{M}$ must be independent of the measurement error in the road noise. Further, $\boldsymbol{M}$ must contain at least as many variables as $\boldsymbol{Z}$ such that $l \geq k$. In our case we construct $\boldsymbol{M}$ by dropping the road noise variable from $\boldsymbol{Z}$ and adding ten more variables (distance and inverse distance from properties to four different types of road and views of road surface from the front and back of properties). By combining the $\boldsymbol{Z}$ matrix that contains a mismeasured variable with the $\boldsymbol{M}$ matrix that does not, the IV estimator produces consistent parameter estimates. The IV estimator is given by;

$$\boldsymbol{\beta}^{IV} = \left(\tilde{\boldsymbol{Z}}' \tilde{\boldsymbol{M}} \left(\tilde{\boldsymbol{M}}' \tilde{\boldsymbol{M}} \right)^{-1} \tilde{\boldsymbol{M}}' \tilde{\boldsymbol{Z}} \right)^{-1} \tilde{\boldsymbol{Z}}' \tilde{\boldsymbol{M}} \left(\tilde{\boldsymbol{M}}' \tilde{\boldsymbol{M}} \right)^{-1} \tilde{\boldsymbol{M}}' \tilde{\boldsymbol{Y}} \tag{21}$$

where $\tilde{\boldsymbol{M}}$ is the matrix of deviations of $\boldsymbol{M}$ from their non-parametrically estimated local means.

6. Spatial Correlation

Up to this point, our statistical analysis of the hedonic price functions in the separate submarkets has ignored the spatial organisation of the data. In effect, we have assumed that the observations of property sales are independent such that we can glean no information on the selling price of a property from the selling price of other properties. Of course, this is hardly likely to be the case. Properties that are located near to each other in space are also likely to share common environmental, accessibility, neighbourhood and perhaps even structural characteristics. Even once we account for the values of known covariates, omitted variables are likely to induce spatial dependence among the errors.

If hedonic residuals are spatially correlated, the parameter estimates from an OLS or IV regression will be inefficient and will produce biased estimates of the standard errors of the parameter estimates. In the case where the residuals are positively spatially correlated, as is to be expected with hedonic property price regressions, OLS or IV will underestimate the population residual variance and the resulting *t*-statistics will be biased upwards. Whilst OLS or IV parameter estimates remain unbiased, ignoring spatial autocorrelation may lead to erroneously high significance being attached to the influence of property attributes on selling prices.

Over recent years, the existence of spatial autocorrelation has received a great deal of attention in the hedonic literature (e.g. Dubin, 1992; Can, 1992; Pace and Gilley, 1997; Basu and Thibodeau, 1998; Bell and Bockstael, 2000). In the main, researchers have focused on the spatial error dependence model. In our case this can be expressed as;

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{Z}}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} \qquad (22)$$

where

$$\boldsymbol{\varepsilon} = \rho\boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{u} \qquad (23)$$

where $\widetilde{\boldsymbol{Y}}$, $\widetilde{\boldsymbol{Z}}$ are defined as before, $\boldsymbol{\beta}^0$ is the [K x 1] vector of "true" parameters (estimated by a consistent estimator such as IV) and $\boldsymbol{\varepsilon}$ is the [$N \times 1$] vector of random error terms with mean zero. The nature of the spatial error dependence is defined by equation (22). Here $\boldsymbol{W}$ is an [$N \times N$] weighting matrix, ρ is the error dependence parameter to be estimated and $\boldsymbol{u}$ is the usual [$N \times 1$] vector of random error terms with expected value zero and variance-covariance matrix $\sigma^2\boldsymbol{I}$. Rearranging (23) we find that;

$$\boldsymbol{\varepsilon} = \left(I - \rho\boldsymbol{W}\right)^{-1}\boldsymbol{u} \qquad (24)$$

which indicates that the error terms $\boldsymbol{\varepsilon}$ have a non-spherical variance-covariance matrix $\sigma^2\left(\boldsymbol{I} - \rho\boldsymbol{W}\right)^{-1}\left(\boldsymbol{I} - \rho\boldsymbol{W}'\right)^{-1}$. Further, the error in the spatial error dependence model can be seen to be made up of two parts; a purely random element and an element containing a weighted sum of the errors on nearby properties. The association between one property and another is contained in the weighting matrix, $\boldsymbol{W}$. The diagonal elements of the weighting matrix are zero since, clearly, the error for an observation cannot be used to explain itself. The off-diagonal elements of the matrix represent the potential spatial dependence between observations. Thus if the ij^{th} element of the weighting matrix, w_{ij}, is zero, we are assuming that there is no correlation in the errors of the i^{th} and j^{th} observations. Conversely if w_{ij} takes on a non-zero value we are assuming that there is correlation in the errors of these two observations.

The researcher must stipulate the nature of dependence between observations by defining the weights matrix in advance of estimation. Here we experimented with a variety of weights matrices but the final specification of (23) used a binary weights matrix in which it was assumed that properties separated by more than 300 metres were unrelated. The w_{ij}^{th} element of $\boldsymbol{W}$**,** therefore, was initially set to one if the i^{th} and j^{th} property were located within 300m of each other, otherwise that element was set to zero.

Following normal procedure, $\boldsymbol{W}$ was row standardised such that each row's elements were made to sum to one. When $\boldsymbol{W}$ is row standardised, the product $\boldsymbol{W}\boldsymbol{\varepsilon}$ equals $\sum_j w_{ij}\varepsilon_j$, and has an intuitive interpretation; it is simply a vector of weighted averages of the errors of neighbouring observations. As Bell and Bockstael (2000) point out, row standardisation is undertaken to simplify estimation of the model. There is usually no underlying economic story supporting the procedure. Moreover, the spatial dependence parameter ρ estimated on a row standardised weights matrix must be interpreted with caution. In particular, ρ in this case is not directly equivalent to an autocorrelation coefficient.

The characteristics of the weights matrices constructed for the property sales observations in the eight submarkets are detailed in Table 6.

Even with a relatively restrictive 300 metre cut-off, the majority of properties are associated with other properties in the same submarket. In submarket 1, for example, only 7 properties out of the 1,395 observations were further than 300 metres from another property in the sample. On average in this submarket, each property was located within 100 metres of 21 other properties in the sample, with at least one observation within 300m of 52 other properties in the sample. Notice that the number of associations in submarket 8 is somewhat lower than in the other submarkets. One explanation of this observation is that properties in

the affluent suburbs are more greatly dispersed than those in the other submarkets.

Table 6: Characteristics of the spatial weights matrices

Submarket	Characteristics of the Spatial Weights Matrix			
	Obs.	Average Associations per Obs.	Maximum Associations	Number with no Associations
1. Ethnic Inner City	1395	21.1	52	7
2. First Time Buyers	1091	21	84	17
3. Affluent, North	1268	12.9	42	15
4. Standard, North	1859	17.1	48	19
5. Affluent, South	1535	16	42	18
6. Standard, South	1338	11.9	31	16
7. White Poor	2066	14	41	30
8. Very Affluent	303	2.1	7	65

The spatial error dependence model can be estimated using maximum likelihood (ML) techniques in which the $\boldsymbol{u}$ vector is assumed to follow a multivariate normal distribution. However, for large samples this may be computationally prohibitive. Instead we follow Bell and Bockstael (2000) and use the generalised moments (GM) estimator developed by Kelejian and Prucha (1999). As Bell and Bockstael (2000) describe, whilst this estimator may not be as efficient as the ML estimator it possesses two advantages. First, the calculation of the estimator is fairly straightforward even with extremely large samples. And second, the GM estimator is consistent even when the error terms $\boldsymbol{u}$ are not normal.

The GM estimator is based on our assumption that the error terms $\boldsymbol{u}$ are distributed $IID\left(0, \sigma^2\right)$. As Kelejian and Prucha (1999) show, this assumption allows us to construct the following three moment conditions;

$$E\left[\frac{1}{N}\boldsymbol{u}'\boldsymbol{u}\right] = \sigma^2$$

$$E\left[\frac{1}{N}\boldsymbol{u}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{u}\right] = \frac{\sigma^2}{N}Tr\left(\boldsymbol{W}'\boldsymbol{W}\right) \qquad (25)$$

$$E\left[\frac{1}{N}\boldsymbol{u}'\boldsymbol{W}'\boldsymbol{u}\right] = 0$$

where the third equality results from the fact that the diagonal elements of $\boldsymbol{W}$ are set to zero.

Of course, the error term $\boldsymbol{u}$ is unobservable from a regression $\tilde{\boldsymbol{Y}}$ on $\tilde{\boldsymbol{Z}}$. Rather, we must rewrite the moment conditions in (25) in terms of $\boldsymbol{\varepsilon}$. Using (24) we get;

$$E\left[\frac{1}{N}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)'\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)\boldsymbol{\varepsilon}\right] = \sigma^2$$

$$E\left[\frac{1}{N}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)'\boldsymbol{W}'\boldsymbol{W}\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)\boldsymbol{\varepsilon}\right] = \frac{\sigma^2}{N}Tr\left(\boldsymbol{W}'\boldsymbol{W}\right) \qquad (26)$$

$$E\left[\frac{1}{N}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)'\boldsymbol{W}'\left(\boldsymbol{I}-\rho\boldsymbol{W}\right)\boldsymbol{\varepsilon}\right] = 0$$

Under our assumptions, the IV estimator (21) provides consistent estimates of the error terms $\boldsymbol{\varepsilon}$ that we label $\hat{\boldsymbol{\varepsilon}}$. To simplify notation we follow Bell and Bockstael (2000) and denote $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{W}\boldsymbol{\varepsilon}$ and $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{W}\boldsymbol{W}\boldsymbol{\varepsilon}$. Thus from (26) we can build the following three-equation system;

$$G_N\left[\rho, \rho^2, \sigma^2\right]' - g_N = v_N\left(\rho, \sigma^2\right) \qquad (27)$$

where the data vectors G_N and g_N are defined as;

$$G_N = \begin{bmatrix} \frac{2}{N}\hat{\boldsymbol{\varepsilon}}'\hat{\hat{\boldsymbol{\varepsilon}}} & \frac{-1}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\hat{\boldsymbol{\varepsilon}}} & 1 \\ \frac{2}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\hat{\boldsymbol{\varepsilon}}} & \frac{2}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\hat{\boldsymbol{\varepsilon}}} & \frac{1}{N}Tr\left(\boldsymbol{W}'\boldsymbol{W}\right) \\ \frac{2}{N}\left(\hat{\boldsymbol{\varepsilon}}'\hat{\hat{\boldsymbol{\varepsilon}}}+\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\boldsymbol{\varepsilon}}\right) & \frac{-1}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\hat{\boldsymbol{\varepsilon}}} & 0 \end{bmatrix} \quad \text{and} \quad g_N = \begin{bmatrix} \frac{1}{N}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ \frac{1}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\boldsymbol{\varepsilon}} \\ \frac{1}{N}\hat{\hat{\boldsymbol{\varepsilon}}}'\hat{\hat{\boldsymbol{\varepsilon}}} \end{bmatrix}$$

and $\nu_N\left(\rho,\sigma^2\right)$ is a [1 × 3] vector of residuals dependent on the parameters ρ and σ^2.

The system of equations in (27) can be solved by nonlinear least squares (NLS) in which the parameter estimates $\hat{\rho}$ and $\hat{\sigma}^2$ are defined as those values that minimise the sum of square residuals; $\nu_N\left(\rho,\sigma^2\right)'\nu_N\left(\rho,\sigma^2\right)$.

Armed with a consistent estimate of the spatial correlation parameter, $\hat{\rho}$, the semiparametric IV model (21) can be re-estimated using feasible generalised least squares (FGLS) according to;

$$\boldsymbol{\beta}^{SP} = \left(\widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\left(\widetilde{\boldsymbol{M}}'(I-\hat{\rho}\boldsymbol{W})'(I-\hat{\rho}\boldsymbol{W})\widetilde{\boldsymbol{M}}\right)^{-1}\widetilde{\boldsymbol{M}}'\widetilde{\boldsymbol{Z}}\right)^{-1}\widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\left(\widetilde{\boldsymbol{M}}'(I-\hat{\rho}\boldsymbol{W})'(I-\hat{\rho}\boldsymbol{W})\widetilde{\boldsymbol{M}}\right)^{-1}\widetilde{\boldsymbol{M}}'\widetilde{\boldsymbol{Y}} \tag{28}$$

where $\boldsymbol{\beta}^{SP}$ is a semiparametric IV estimator accounting for spatial error dependence.[8] The estimator in (28) is calculated using code written by the consultants in the Gauss programming language. The calculations are made feasible even in relatively large sample sizes through the use of sparse matrix commands that take advantage of the relatively large number of zero elements in the weights matrix $\boldsymbol{W}$.

[8] New residuals could be estimated using $\boldsymbol{\beta}^{SP}$ and the new solution for ρ recovered in order to iterate the FGLS estimator. However, this is not relevant for large samples and this approach is not followed here.

7. Results

To provide an indication of the degree of explanatory power achieved in the econometric analysis, observe Table 7. This presents the results of three auxiliary regressions carried out for each submarket. The first column reports the R^2 statistic for a simple OLS regression of natural log of property price, ln(***P***), against ***X***; the four variables to be contained in the nonparametric part of the final model (see Table 5) and a constant. Since one of the motivating factors behind including these variables (floor and garden area, proeprty age and sale date) in the nonparametric part of the model, we would hope to see these four variables explaining a large part of the variation in the dependent variable. The R^2 statistic measures the proportion of the total variation in the dependent variable explained through the included regressors.

For some of the submarkets, notably submarkets 3 and 5, the contention that the four variables play an important role in explaining property prices is well supported by the high R^2 statistics. However, in other submarkets, these variables seem to have low explanatory power. This observation may well stem from the clustering procedure that is itself based, in part, on these variables. Within submarkets, such as submarket 7, the clustering process may have introduced relative homogeneity with respect to property floor area and garden size thereby reducing the importance of these variables in explaining differences in property prices within that submarket.

The second column of Table 7 presents a simple linear OLS regression of all the variables used in the econometric analysis, that is the ***X*** and ***Z*** matrices. These models now contain around 64 variables compared to the 4 variables used to generate the R^2 scores in the first column. As we would expect, in all submarkets, the explanatory power of the model improves dramatically. The third column of Table 7 shows the same analysis but now uses the semiparametric model in which the influence of the four variables in the ***X*** matrix are accounted for using nonparametric regression techniques. Increasing the flexibility of the functional form by using the semiparametric estimator again results in significant gains in the explanatory power of the model. For 7 of the 8 submarkets the explanatory power of the model increases between 6% and 9%. In one case, submarket 2, the increase in explanatory power is a massive 20%.

Though these results are presented for OLS models and not the IV models corrected for spatial dependence (where the interpretation of an R^2 statistic is somewhat less clear) they tend to support the contention that the semiparametric model significantly increases the power of the model to describe variation in property prices.

Table 7: Explanatory power of OLS specifications

Submarket	R^2 Statistics		
	OLS Nonparametric Variables only	OLS All Variables	OLS Semiparametric
1. Ethnic Inner City	.401	.503	.571
2. First Time Buyers	.254	.459	.660
3. Affluent, North	.468	.689	.765
4. Middle, North	.261	.496	.556
5. Affluent, South	.474	.677	.749
6. Middle, South	.174	.571	.635
7. White Poor	.122	.444	.530
8. Very Affluent	.379	.630	.710

Table 8 presents results for the detection and estimation of spatial error dependence. The Table presents the results of two tests[9] and the estimated spatial dependence coefficient for each of the submarkets calculated using semiparametric IV regression residuals at the optimal (cross-validated) band-width.

The first test statistic is Moran's I statistic (Cliff and Ord, 1972). This test is predicated on normal errors and tests the null hypothesis that there is no spatial dependence between error terms (that is, $\rho = 0$). The test statistic is asymptotically distributed as a standard normal variate. Clearly, the probability of the null being true is very low in each of the submarkets. The second test statistic is that proposed by Kelejian and Robinson (1992). This test is valid even with nonnormal errors. The test statistic is chi-squared distributed with degrees of freedom given by the number of parameters in the model. Even with the robust test the results are conclusive, the null hypothesis of no spatial error dependence is rejected with over 99% confidence in each case. The spatial autocorrelation coefficient estimated for each submarket is reported in the final column of Table 8.

[9] Computational details can be found in Anselin and Hudak (1992).

Table 8: Spatial error dependence test statistics and correlation coefficient

Submarket	Test Statistics				SAR Coef. (ρ)
	Moran's I		Kelejian-Robinson		
	Stat	Prob	Stat	Prob	
1. Ethnic Inner City	194.30	.000	.107	.000	.438
2. First Time Buyers	1,392.59	.000	.228	.000	.510
3. Affluent, North	102.37	.001	.056	.000	.217
4. Standard, North	360.53	.000	.100	.000	.345
5. Affluent, South	337.84	.000	.119	.000	.352
6. Standard, South	231.55	.000	.121	.000	.307
7. White Poor	397.87	.000	.121	.000	.359
8. Very Affluent	225.90	.000	.191	.000	.175

Table 9 highlights some of the results from the full model. The top part of the table picks out a handful of the parameter estimates for the eight submarkets. The final three rows of the table list the number of parameters in each submarket regression, *K*, the optimal bandwidth selected through cross-validation, *h*, and the number of observations in each submarket, *N*. Notice that *K* is differs across the submarket models, since some variables (especially the dummy variables for property type or beacon group) showed no variation within certain submarkets.

In all eight submarket models the dependent variable is the natural logarithm of property price. Thus the parameter estimate for a variable such as the 'number of bedrooms' can be interpreted as the percentage change in the price of a property from the addition of an extra bedroom, all else equal. Indeed, the 'number of bedrooms' variable tends to behave as expected; all significant parameter estimates are positively signed and these indicate that an extra bed-room increases a property's price by between 2.5% and 5%. Popular perception might indicate that this variable should be more important in determining property prices. Of course, in the models presented here, the overall size of the property is controlled for by including 'floor area' as a regressor in the non-parametric part of the model. As such the parameter estimates presented in Table 9, indicate the percentage increase in the price of a property for each extra bedroom, *holding total floor area constant*. Hence more bedrooms should really

be interpreted as ‘more but smaller bedrooms’ and the relative unimportance of this variable seems more acceptable.

The variable indicating the number of toilets in a property has no significant impact on property prices, though, on the whole, the parameters tend to indicate a positive relationship between WCs and property value. Again, this lack of significance may be due to the fact that, in general, the number of bedrooms in a property will be highly correlated with the property’s overall size for which we have already controlled. Also, the use of a continuous variable (i.e one providing a count of the number of toilets in a property), may not have been the most appropriate choice of functional form once the data had been partitioned into the relatively homogenous submarket groupings. Rather, a dummy variable specification may have proved more successful.

The dummy variable indicating the presence of a garage is positive in all submarkets indicating that having a garage can add from 2% (submarket 2) to 15% (submarket 8) to the value of a property. In six of the eight submarkets the garage parameter is statistically significant. Notably, the presence of a garage is significant in all of the suburban submarkets but is insignificant in the two submarkets (1 and 2) located mainly in the inner city. This might reflect, either paucity in the data with very few properties in these areas possessing garages or that in such areas less people have access to or requirement of their own vehicle. The parameters estimated on the ‘house floors’ and ‘detached house’ variables are unequivocal. A property is valued more highly the less floors it has (all else equal) and being detached adds between 7% and 18% to the selling price of a property depending on submarket.

Table 9: Selected parameter estimates from the semiparametric instrumental variables estimator accounting for spatial error dependence

Variable	Submarket							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Bedrooms	0.026*	-0.002	0.032**	0.036***	0.024**	0.005	-0.024	0.002
WCs	0.017	0.012	0.011	0.022	-0.004	-0.006	0.013	0.001
Garage	0.033	0.020	0.036*	0.034***	0.047***	0.069***	0.033***	0.135**
House Floors	-0.147***	-0.083**	-0.129*	-0.093***	-0.133***	-0.159***	-0.068**	-0.097
Detached House	0.119*	0.163***	0.112***	0.103***	0.110***	0.062**	0.178***	0.073
Primary School	0.0087	0.120***	0.161***	0.203***	0.042*	0.146***	0.105***	0.148
View of Water	-0.013	0.001	-0.0001	-0.0002	-0.004*	0.014***	0.006**	-0.001
View of Park	0.0001	0.001**	0.0002	-0.0003	0.0001	-0.0001	0.0000	0.0003
Wealth	0.170***	0.142***	0.225***	0.142***	0.195***	0.160***	0.145***	0.207***
Ethnicity	0.031	-0.166***	-0.163***	-0.103***	-0.093***	-0.118**	-0.040***	-0.291***
K	**57**	**60**	**61**	**60**	**61**	**60**	**60**	**60**
h	**.39**	**.23**	**.36**	**.33**	**.31**	**.35**	**.31**	**.60**
N	**1,395**	**1,091**	**1,268**	**1,861**	**1,536**	**1,338**	**2,097**	**303**

* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence

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, 2001) suggesting that primary school quality and proximity are a very important determinant of property price. Only in Submarket 1, the ethnic inner city, and Submarket 8, the very affluent, is the primary school variable statistically insignificant. The latter probably indicates the lack of parents with young children amongst the very affluent, whilst the former might indicate that the 'primary school' variable is inappropriate to describe educational priorities amongst the mostly Asian members of Submarket 1. In particular, the variable takes no account of the presence of independent Muslim schools in the inner city Birmingham region. Unfortunately, the two view variables do not seem to add a great deal to the analysis. In all but a handful of cases the parameters on 'views of water' and 'views of parks' are insignificant.

The variables describing the socioeconomic characteristics of neighbourhoods (constructed in the factor analysis) tend to be important in explaining variation in property prices. In Table 9 we present parameter estimates for the Wealth and Ethnicity factors. Unsurprisingly, the increasing wealth of the inhabitants of an area tends to manifest itself in higher property prices. Whilst there is an issue of the direction of causality in this relationship we assume here that it is the presence of more affluent neighbours that increases the selling price of a property. In seven of the eight submarkets, the increasing presence of residents from ethnic minorities tends to decrease property prices. In contrast, the parameter on the ethnicity variable in Submarket 1 (the ethnic inner city), is positive; suggesting that for this submarket the increasing presence of residents from ethnic minorities increases selling prices of properties. Combined these two observations tend to suggest a preference for ethnic homogeneity amongst the residents of the City of Birmingham.

Table 10 presents the results for the noise pollution variables. These are included in the hedonic price function in a piecewise linear fashion. That is, noise pollution is assumed to have no impact on property prices until it exceeds a threshold level of 55dB. This threshold is often taken as the "background" noise level in urban environments. Since the dependent variable is the log of property price, the coefficients represent the Noise Sensitivity Depreciation Index (NSDI). In other words, the coefficient gives the constant percentage response in property price to a one decibel absolute increase in noise pollution over 55dB.

Table 10: Noise pollution parameter estimates from the semiparametric instrumental variables estimator accounting for spatial error dependence

Submarket	Noise Variable								
	Road			Rail			Air		
	N	Coef	Imp Pr	N	Coef	Imp Pr	N	Coef	Imp Pr
1	273	-.0043*	-149.13	45	-.0107**	-365.43	0		
2	283	-.0158***	-825.98	57	-.0139***	-725.86	0		
3	268	-.0035*	-320.27	23	-.0043	-391.54	32	-.0066	-602.00
4	676	-.0034**	-160.60	72	-.0024	-112.65	45	-.0179	-856.53
5	355	-.0067***	-4761	66	-.0046	-322.60	0		
6	299	-.0050**	-248.80	31	-.0109	-537.11	0		
7	500	-.0042**	-174.29	76	-.0085***	-357.79	375	-.0149***	-627.41
8	117	-.0115**	-1,885.84	9	-.0338*	-5,560.79	2	-.0682	-11,236.00

* Significant at 10% level of confidence
** Significant at 5% level of confidence
*** Significant at 1% level of confidence

As expected each submarket returns a somewhat different estimate for the coefficient on noise pollution. These differences will result from differences in the prevalence of noise pollution and the preferences of households in each submarket.

Let us begin by examining the findings presented in Table 10. Here the results for each source of noise pollution (road, rail and air traffic) for each submarket are given in three columns. In each case, the first column is titled *N* and indicates the number of observations in that submarket registering a level of noise pollution (from that source) above the 55dB baseline. Notice how there are many more observations of properties suffering from road noise than there are rail noise. Indeed, the relative paucity of properties exposed to rail traffic noise suggests that it will be relatively more difficult to find a statistically significant relationship between property prices and rail traffic noise pollution. The same could be said of aircraft noise which is concentrated, to a large extent, in Submarket 7 located in the East of the city near Birmingham International Airport. Indeed, in four of the submarkets there are no properties exposed to aircraft noise. The second column for each noise source in Table 10 provides coefficient estimates and the third column implicit prices.

Let us begin by discussing the parameter estimates. Notice first, that in accordance with prior expectations, all parameters on the three types of noise pollution in each of the eight submarkets have negative coefficients.

Focusing on the road noise variables, observe that the parameters range from a value of -.0034 in Submarket 4 to a maximum of -.0158 in submarket 2. In other words, a one decibel increase in road traffic noise can wipe off between 0.3% and 1.6% of the selling price of a property, depending on submarket. Encouragingly, six of the eight submarkets have coefficients that are significant at the 95% level of confidence whilst the other two are significant at the 90% level of confidence. Reassuringly, these parameter estimates cover the range reported from studies in other markets (see Bateman *et al*., 2001 for a review).

Turning to the rail noise coefficients, observe that on the whole the parameters are of a similar magnitude to those found to characterise road noise. Indeed, whilst the exact statistical tests have not been performed, the differences do not appear large enough to conclude that the two noise sources have a significantly different impact on property prices (though possible exceptions might be Submarket 1 and Submarket 7). Whilst all rail noise coefficients are negative, only four of the eight submarkets have parameters that are significantly different from zero at the 90% level of confidence. This result is presaged by the relative paucity of observations of properties exposed to rail traffic noise pollution. In a similar vein, the anomalously large parameter estimate in Submarket 8 is more

likely to be explained by the fact that this is based on only 9 observations rather than a substantive difference in the impact of rail noise in this submarket.

The air noise coefficients perform the least convincingly with only one of four submarkets returning a statistically significant coefficient. Once again, the anomalously large coefficient on aircraft noise in submarket 8 probably reflects paucity of data.

In general, however, the results presented in Table 10 are very pleasing. The coefficients are all correctly signed and mostly have plausible magnitudes.

The implicit price of noise (i.e. the extra that must be paid for an identical property boasting one unit less noise pollution) is given by the partial derivative of the hedonic price function according to;

$$p_{z_k}(z_k; \mathbf{z}_{-k}) = \frac{\partial P(\mathbf{z})}{\partial z_k} \tag{29}$$

where z_k is a noise variable.

Since the empirical hedonic price function estimated here is of semi-log form, the implicit price for noise can be calculated according to the specific equation;

$$\begin{aligned} p_{z_k}(z_k; \mathbf{z}_{-k}) &= \exp\left(E[\ln P \mid \mathbf{x}_i] + (z_i - E[\mathbf{z} \mid \mathbf{x}_i])\hat{\boldsymbol{\beta}}_{z_k}\right)\hat{\boldsymbol{\beta}}_{z_k} \\ &= \exp\left(\ln \hat{P}_i\right)\hat{\boldsymbol{\beta}}_{z_k} \end{aligned} \tag{30}$$

where $\hat{\boldsymbol{\beta}}_{z_k}$ is the parameter estimated on the noise variable z_k.

Average implicit prices in each submarket have been calculated for the three noise variables and are presented in Table 10.

8. Conclusions

This paper has attempted to provide a "state of the art" empirical analysis of hedonic housing price data. The data set used in this estimation is perhaps the richest of its kind yet to be constructed for any property market. Issues of market segmentation have been addressed through the clustering of properties into eight submarkets identified according to similarities in their geographical location, property types and the socioeconomic composition of neighbourhoods.

The analytical approach taken is one that introduces considerable flexibility into the specification of the empirical hedonic price function. In particular, the influence of key structural characteristics on properties is modelled nonparametrically, whilst the influence of other property characteristics are captured using a traditional linear parametric form. Estimation of the parameters of the hedonic price function for each submarket is achieved through the following steps;

1. **Construct Data Matrices**: The researcher selects the characteristics that will be included in the hedonic price model. From this list the researcher decides which variables will be included in the nonparametric and which in the parametric parts of the model. The variables in the nonparametric part of the model are grouped into the matrix $\boldsymbol{X}$ with typical row $\boldsymbol{x}_i$. In our case this consists of the four variables, log of floor area, log of garden size, age in decades and date of sale.

   Variables in the parametric part of the model are grouped into the matrix $\boldsymbol{Z}$ with typical row $\boldsymbol{z}_i$. Since it is assumed that the road noise variable is measured with error a further matrix $\boldsymbol{M}$ is constructed. $\boldsymbol{M}$ consists of all the columns of $\boldsymbol{Z}$ bar the road noise variable but including the additional variables indicating straight line distance to different types of roads, the inverse of these distances and the area of road surface visible from the front and back of the property.

2. **Bandwidth Selection**: Select the maximum and minimum value for the bandwidth of the nonparametric kernel. The cross-validation procedure will grid search across this range seeking the bandwidth that minimises the cross-validation statistic. Set the bandwidth, h, to the minimum value in this range.

3. **Adaptive Kernel**: Using the bandwidth, h, calculate the density of $\boldsymbol{x}$, at each observation in the data set according to (11). Then, using (18) calculate a new observation specific bandwidth h_i that is adapted to the density of the data in the region of $\boldsymbol{x}_i$.

4. **Nadarya-Watson Kernel Regression**: Using Nadarya-Watson kernel regression, (12) and the bandwidth h_i, estimate the expectations of $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{M}$ conditional on the variables $\boldsymbol{X}$. Strip these expectations from the data to form the matrices $\widetilde{\boldsymbol{Y}}$, $\widetilde{\boldsymbol{Z}}$ and $\widetilde{\boldsymbol{M}}$ representing differences from nonparametric means.

5. **Semiparametric Instrumental Variables Regression**: Using IV regression (21) on the data matrices $\widetilde{\boldsymbol{Y}}$, $\widetilde{\boldsymbol{Z}}$ and the instrument matrix $\widetilde{\boldsymbol{M}}$, estimate the parameters $\boldsymbol{\beta}^{IV}$, which, according to our assumptions, should be a consistent estimator of the true parameters $\boldsymbol{\beta}^0$.

6. **Spatial Dependence Parameter**: Using the IV regression error terms estimate the spatial dependence parameter, ρ, using the General Method of Moments estimator defined by (27).

7. **Semiparametric Instrumental Variables Regression corrected for Spatial Error Dependence**: Use the estimated spatial dependence parameter, $\hat{\rho}$, to estimate the parameters $\boldsymbol{\beta}^{SP}$ according to (28).

8. **Cross-Validation**: Using $\boldsymbol{\beta}^{SP}$ calculate the cross-validation statistic according to (16). Increment the bandwidth h by a small amount and repeat steps 3 to 7 until have grid-searched across the whole range of values selected in step 2. Select parameter estimates of $\boldsymbol{\beta}^{SP}$ that minimise the cross-validation statistic.

In general, the results of this semiparametric model are very pleasing. The main objective of the empirical research in this paper was to return estimates of the implicit price of noise from road, rail and air traffic. Pleasingly, the model generates coefficients on the noise variables that are all correctly signed and most have plausible magnitudes. Of the 20 noise parameters estimated in the IV model 13 are significant at the 10% level of confidence.

References

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

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

Anselin, L., and Hudak, S., (1992). "Spatial econometrics in practice: A review of software options", *Regional Science and Urban Economics*, 2, pp 509 – 536.

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

Basu, S. and Thibodeau, T.G., (1998). "Analysis of spatial autocorrelation in house prices", *Journal of Real Estate Finance and Economics*, 17(1), pp 61-85.

Bateman, I.J., Day, B.H., Lake, I., and Lovett, A.A., (2001). *The Effect of Road Traffic on Residential Property Value: A Literature Review and Hedonic Pricing Study*, Edinburgh: Scottish Executive and The Stationary Office.

Bell, K. and Bockstael, N.E. (2000). "Applying the generalised method of moments approach to spatial problems involving micro-level data", *Review of Economics and Statistics*, 82(1), pp 72-82.

Butler, R.V. (1980). "Cross-sectional variation in the hedonic relationship for urban housing markets", *Journal of Regional Science*, 20, pp 439-453.

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

Cheshire, P. and Sheppard, S. (1998). "Estimating the demand for housing, land and neighbourhood characteristics", *Oxford Bulletin of Economics and Statistics*, 60(3), pp 357-382.

Cliff, A., and Ord, J.K., (1972). "Testing for spatial autocorrelation among regression residuals", *Geographical Analysis*, 4, pp 267-284.

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

Davidson, R., and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*, Oxford University Press: New York.

Day, B.H., (2001). "The theory of hedonic markets: Obtaining welfare measures for changes in environmental quality using hedonic market data", *Report to EU Working Group on Noise*, CSERGE, University College London, UK.

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

Dubin, R.A., (1992). "Spatial autocorrelation and neighbourthood quality", *Regional Science and Urban Economics*, 22, pp 433-452

Gibbons, S., (2002). "Paying for good neighbours? Neighbourhood deprivation and the community benefits of education", *Centre for the Economics of Learning Discussion Paper*, 17.

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

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

Kelejian, H.H., and Prucha, I.R., (1999). "A generalised moments estimator for the autoregressive parameter in a spatial model", *International Economic Review*, 40(2), pp 509-533.

Kelejian, H.H., and Robinson, D.P., (1992). "Spatial autocorrelation: A new computationally simple test with an application to per capita county police expenditures", *Regional Science and Urban Economics*, 22, pp 317-331.

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

Lindeman, R.H., Merenda, P.F., and Gold, R., (1980). "Chapter 8: Factor Analysis", in *Introduction to Bivariate and Multivariate Analysis*, Scott, Foresman and Co.

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

Openshaw, S. (1995). *Census users' handbook*. Cambridge: Pearson Professional Limited.

Ordnance Survey (1996). *Digital Map Data and Customised Services*. Southampton: Ordnance Survey.

Pace, R.K., and Gilley, O.W., (1997). "Using the spatial configuration of the data to improve estimation", *Journal of Real Estate Finance and Economics*, 14(3), pp 333-340.

Posse, C. (2001). "Hierarchical model-based clustering for large datasets", *Journal of Computational and Graphical Statistics*, 10(3), pp464-86.

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

Robinson, P.M., (1988). "Root-*N*-consistent semiparametric regression", *Econometrica*, 56, pp 931-954.

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

Smith V.K. and Huang, J-C., (1995). "Can markets value air quality? A meta-analysis of hedonic property values", *J. Pol. Econ.*, 103, pp 209-27.

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

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.

Straszheim, M.R. (1974). "Hedonic estimation of housing market prices: A further comment", *Review of Economics and Statistics*, 56, pp 404-06.

Straszheim, M.R. (1975). *An Econometric Analysis of the Urban Housing Market*, New York: National Bureau of Economic Research.

Appendix A: Variables used in the Regression Analysis

Table A1: Structural variables included in the Hedonic Price Models

Variable	Code	Description and *a priori* Expectations
Floor Area (m^2)	Area	Larger properties will command higher prices
Garden Area (m^2)	Garden	Properties with larger gardens will command higher prices
Number of Bedrooms	Bedrooms	Properties with more bedrooms will tend to command higher prices.
Number of WCs	WCs	Properties with more WCs will tend to command higher prices.
Number of Storeys	Storeys	Given that two properties have the same floor area it is expected that those with less storeys will be preferred to those with more storeys.
Garage	Garage	Properties with a garage will tend to command higher prices.
Central Heating	Central Heating	Properties with central heating will tend to command higher prices.
Age of property (decades)	Age	The relationship between property age and property price is not entirely clear. Older properties may be desired for their "character" and "original features", more modern properties for their state of repair and more up-to-date facilities. What is clear, however, is that property age proxies for a number of property characteristics not least of which will be the architectural design of the house.
Property Type (Dummy Variables)	Detached House Semi-Detached House End Terrace House Terrace House Detached Bungalow Semi-Detached Bungalow End Terrace Bungalow Terrace Bungalow	In the models, semi-detached houses are taken as the baseline property type since all submarkets contain properties of this type. The coefficients estimated on the other property type dummy variables, reflect the relative difference in price between that property type and a semi-detached house with exactly the same characteristics. In general, it is expected that houses will fetch more than bungalows. Moreover, properties will increase in value from terraces through end terraces and semi-detached properties through to detached properties.
Beacon Group (Dummy Variables)	BG 1 (Unrenovated cottage pre 1919) BG 2 (Renovated cottage pre 1919) BG 3 (Small "industrial" pre 1919) BG 4 (Medium "industrial" pre 1919) BG 5 (Large terrace pre 1919) BG 8 (Small "villa" pre 1919)	In the models, BG 21 (standard houses built between the war) is taken as the baseline beacon group since all submarkets contain properties of this type. The coefficients estimated on the other beacon group dummy variables, reflect the relative difference in price between that properties of that beacon group and a property in beacon group 21 with similar characteristics. The beacon group data collected from the VOA provides a detailed categorisation of properties according to their age, size, architectural type and quality. As such, we would expect these dummy variables to be important descriptors that add significantly to the explanatory power of the model.

Variable	Code	Description and *a priori* Expectations
	BG 9 (Large "villas" pre 1919)	
	BG 10 (Large detached pre 1919)	
	BG 19 (Houses 1908 to 1930)	
	BG 20 (Subsidy houses 1920s & 30s)	
	BG 21 (Standard houses 1919 to 1945)	
	BG 24 (Large houses 1919 to 1945)	
	BG 25 (Individual houses 1919 to 1945)	
	BG 30 (Standard houses 1945 to 1953)	
	BG 31 (Standard houses post 1953)	
	BG 32 (Large houses post 1953)	
	BG 35 (Individual houses post 1945)	
	BG 36 ("Town Houses" post 1950)	

Table A2: Neighbourhood variables included in the Hedonic Price Models

Variable	Code	Description and *a priori* Expectations
Wealth Factor	Wealth	Increasing values for this factor indicate the increasing wealth of a properties neighbourhood. Properties in wealthier neighbourhoods are expected to command higher prices in the market.
	Wealth Squared	
Ethnicity Factor	Ethnicity	Increasing values for this factor indicate the increasing presence of members of ethnic minorities in a neighbourhood. The (perhaps unfortunate) expectation is that increasing ethnicity will reduce property prices..
	Ethnicity Squared	
Age Composition Factor	Age Composition	Increasing values for this factor indicate the increasing age of adults in a properties neighbourhood. It is expected that properties in neighbourhoods with generally older residents will command higher prices in the market.
	Age Composition Squared	
Family Composition Factor	Family Composition	Increasing values for this factor indicate the increasing presence of households with children in a neighbourhood. Properties in neighbourhoods with more families are expected to command lower prices in the market.
	Family Composition Squared	

Table A3: Locational variables included in the Hedonic Price Models

Variable	Code	Description and *a priori* Expectations
Proximity to City Centre (mins)	CBD	This variable measures the travel time by car from a property to the city centre. Since it is expected that this relationship may be non-linear both linear and squared terms are included. One likely result is that property prices will fall moving away from the city centre but at a declining rate.
	CBD squared	
Proximity to and Size of Local Centres	Local Centre	This variable uses a weighted average of inverse walking distances to general stores to measure the proximity to local centres that accounts for the size of the local centre. Expectations are similar to those for proximity to the city centre.
	Local Centre squared	
Proximity to a Railway Station (inverse mins)	Railway Station	This variable measures the inverse of the walking distance from a property to the nearest railway station. It is expected that property values will be higher near to a transport network node such that a positive coefficient is anticipated.
Proximity to a Park (inverse mins)	Park	This variable measures the inverse of the walking distance from a property to the nearest park. It is expected that property values will be higher near to a recreational areas such that a positive coefficient is anticipated.
Proximity to A-Type Industrial Processes (inv. m)	Industry A	This variable measures the inverse of the straight line distance between each property and the nearest large scale industrial plant. Since such plants are assumed disamenities it is anticipated that this variable will have a negatively signed coefficient.
Proximity to B-Type Industrial Processes (inv. m)	Industry B	This variable measures the inverse of the straight line distance between each property and the nearest medium scale industrial plant. Again it is anticipated that this variable will have a negatively signed coefficient.
Proximity to Land Fill sites (inv. m)	Land Fill	This variable measures the inverse of the straight line distance between each property and the land fill site. Again, it is anticipated that this variable will have a negatively signed coefficient.
Proximity and Quality of Primary Schools	Primary School	Using government published statistics, this variable provides a distance weighted average of the performance of nearby primary schools. Since parents may be attracted to locations near better primary schools, it is anticipated that this variable will have a positive coefficient.
Region (dummy variables)	Central Region	These dummy variables provide a high level spatial categorisation of properties by geographical location. Since different submarkets are located in different spatial regions, the baseline region varies across submarket models.
	Northern Region	
	Western Region	
	Eastern Region	
	Sutton Coldfield Region	
	Southern Region	

Table A4: Environmental variables included in the Hedonic Price Models

Variable	Code	Description and *a priori* Expectations
Road Traffic Noise (dB)	Road Noise*	These variables measure decibels of noise above 55dB from different sources of noise pollution. The 55dB cut off reflects the fact that noise levels below this level are indistinguishable from "background" noise in an urban environment.
Railway Traffic Noise (dB)	Rail Noise	
Aircraft Traffic Noise (dB)	Air Noise	Since noise pollution is a disamenity. We would expect properties exposed to greater levels of noise to command lower prices.