ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Day, Brett; Bateman, Ian; Lake, Iain

Working Paper

Ommitted locational variates in hedonic analysis: A semiparametric approach using spatial statistics

CSERGE Working Paper EDM, No. 04-04

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

Suggested Citation: Day, Brett; Bateman, Ian; Lake, Iain (2004) : Ommitted locational variates in hedonic analysis: A semiparametric approach using spatial statistics, CSERGE Working Paper EDM, No. 04-04, 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/10419/80299

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

OMITTED LOCATIONAL VARIATES IN HEDONIC ANALYSIS:

A SEMIPARAMETRIC APPROACH USING SPATIAL STATISTICS

by

Brett Day,
Ian Bateman and Iain Lake

CSERGE Working Paper EDM 04-04

OMITTED LOCATIONAL VARIATES IN HEDONIC ANALYSIS: A SEMIPARAMETRIC APPROACH USING SPATIAL STATISTICS

by

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

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

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

Author contact details:
Brett Day: email – brett.day@uea.ac.uk
Tel: 0044 (0) 1603 592064, Fax: 0044 (0) 1603 593739

Acknowledgements

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

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

A frequent assumption of hedonic price estimation using property market data is that spatial autocorrelation of regression residuals is a feature of the error generating process. Under this assumption, *spatial error dependence* models that impose a specific spatial structure on the error generating process provide efficient parameter estimates. In this paper we argue that spatial autocorrelation is induced by spatial features influencing property prices that are not observed by the researcher. Whilst many of these features comprise the subtle nuances of location that might adequately be handled by modelling the error process, others may be substantive spatial features whose absence from the model is likely to induce omitted variable bias in the parameter estimates. Accordingly we propose an alternative estimation strategy. We use spatial statistics to determine the nature of spatial dependence in regression residuals. Subsequently we adopt a semiparametric *smooth spatial effects* estimator to account for omitted locational covariates over the spatial scale indicated by the spatial statistics. The parameter estimates from this model are found to differ significantly from those of a spatial error dependence model.

Key words: Hedonics, omitted variables, spatial error dependence, smooth spatial effects, Moran's *I*

1. INTRODUCTION

The recent history of research into the hedonic analysis of property markets has witnessed a widespread recognition of the importance of spatial processes. In the theoretical literature, models have been developed in which households' choices of residential location may depend explicitly on the sets of people that choose to live in each location (e.g. Epple and Platt, 1998; Epple and Seig, 1999; Nesheim, 2002). These models predict that in equilibrium, households will sort themselves across the urban area such that the characteristics of households living in the same neighbourhood are likely to be more similar to each other than they are to the population as a whole. This equilibrium is characterised by a hedonic price function that maintains price differentials between locations in the urban area (Nesheim, 2002).

Likewise, in the empirical literature, there has been a growing acknowledgement that the econometric methods used to estimate hedonic price functions from property market data should explicitly concern themselves with the spatial organisation of the data (e.g. Dubin, 1988, 1992, 1998; Can, 1992; Pace and Gilley, 1997; Can and Megbolugbe, 1997; Basu and Thibodeau, 1998; Pavlov, 2000; Bell and Bockstael, 2000; Leggett and Bockstael, 2000; Gawande and Jenkins-Smith, 2001; Ihlanfeldt and Taylor, 2001, Gibbons and Machin, 2003; Gibbons, 2003).

In particular, empirical researchers are concerned with the fact that the selling prices of properties will be positively correlated over space. In addition to the price differentials generated by the sorting processes identified in the theoretical literature, there are numerous reasons why researchers might expect prices to be spatially correlated. For example, urban areas often develop piecemeal over time. Local neighbourhoods tend to be constructed at the same time and by the same developers. Consequently, properties within neighbourhoods are likely to exhibit structural similarities not only in terms of their age but also in terms of their size, layout and interior and exterior design features. Moreover, properties in the same neighbourhood also share the same physical surroundings. As such they will have comparable access to locational amenities (e.g. schools, shops, parks, transport links etc.) and exposure to disamenities (e.g. industrial sites, landfills, air pollution, noise pollution etc.). If households value proximity to (distance from) these amenities (disamenities), then the selling prices of properties will be correlated over space.

The efforts of empirical researchers to incorporate spatial considerations into their analyses have been manifold. For example, in order to attend to the theoretical prediction that property prices may vary according to the socioeconomic composition of neighbourhoods, researchers invariably include measures of neighbourhood socioeconomics in their specification of the hedonic price function. Alternatively, hedonic price functions can be specified such that the marginal prices of property attributes are allowed to vary according to neighbourhood characteristics (e.g. Can, 1992, Can and Megbolugbe, 1997). In a similar vein, some researchers have sort to identify sets of socioeconomically homogeneous neighbourhoods and estimate separate hedonic price functions for properties falling into each set (e.g. Day, 2003; Day et al., 2003; Goodman and Thibodeau, 2003). We employ this latter estimation strategy in the empirical work presented in this paper.

Furthermore, researchers have employed ever more sophisticated data sets that provide details of many of the structural characteristics of properties and make use of geographical information systems (GIS) to construct variables that paint a comprehensive picture of each properties' access to amenities and exposure to disamenities (e.g. Lake et al., 2000). The data set used in this study and described in Section 2 of this paper is an example of just such a data set.

Despite these advances in data collation, it seems unlikely that any data set will be sufficiently comprehensive that it captures every aspect of property construction and location that might induce correlation in prices over space.

Typically, researchers address this problem by including locational constants that crudely describe each property's location in the urban area (e.g. properties may be categorised according to postal region or perhaps administrative or political subdivisions of the urban area). Even so, there is no guarantee that these locational constants are effective proxies for variations in the unobserved covariates. In particular, it seems unlikely that the unmeasured spatial processes will operate on the exact spatial scale as the regions defined by the locational constants. Similarly, it seems implausible to expect that these spatial processes will obey the rigid boundaries imposed by the locational constants. For example, it is more likely that a property located at the edge of its allotted region will hold more in common with properties lying just over the boundary in the adjacent region, than it will with properties on the far side of its own region (Dubin, 1992). Alternatively, some researchers include as regressors polynomial expressions in the latitude and longitude of each property (e.g. Dubin, 1992; Pace and Gilley, 1997). Whilst allowing for continuous variation in prices over the urban area, this approach will only effectively capture large-scale spatial variation in prices.

The fact remains, that any empirical specification of the regressors in a hedonic price function is unlikely to be sufficiently comprehensive to remove all spatial effects from the data. Of course, one can test this hypothesis by examining regression residuals for spatial autocorrelation. Evidence of positive spatial autocorrelation in regression residuals is an indication of spatial processes that are not captured by the specification of the hedonic price function. As described in Section 3 of this paper these tests require the researcher to specify *a priori* the area over which spatial autocorrelation in the regression residuals is thought to operate. However, there is no established procedure for determining this distance. Can (1992) and Bell and Bockstael (2000), for example, simply try a variety of distances and find evidence of spatially correlated residuals in all cases.

Alternatively, a more thorough appreciation of the nature of spatial dependence in regression residuals can be obtained through construction of the spatial correlogram. In the hedonic analysis of property markets, Dubin (1988, 1992, 1998) constructs spatial correlograms for residuals by taking the average correlation in residuals at progressively larger separation intervals or distance classes. In this paper we propose a more sophisticated approach inspired by the paper of Ellner and Seifu (2002). Here we employ a test of spatial autocorrelation of regression residuals known as Moran's *I* statistic (Cliff and Ord, 1972). We calculate Moran's *I* statistic for residuals at progressively larger separation intervals. Since the distribution of *I* under a null hypothesis of no spatial autocorrelation is known, it is possible to establish statistically the separation interval at which correlation of the residuals is no longer a feature of the data.

Of course, having identified spatial autocorrelation in regression residuals, the researcher is faced by the troublesome task of deciding how to proceed. As described in Section 4 of this paper, there are, in essence, three routes that may be followed. One approach is to assume that one has data on all relevant determinants of property prices and that spatial autocorrelation of the residuals is merely an artefact of a mis-specified model. Under this assumption the prognosis is that respecifying the model will solve the problem. A second approach is to assume that the true model is the model at hand but that autocorrelation among the disturbances is due to spatial dependence in the process generating the nuisance. Again the proscribed course of action is to model that nuisance process and thereby alleviate the symptoms of spatial autocorrelation of residuals. Models of this type we describe as *Spatial Error Dependence* models. SED models have become increasingly popular in applied work in the hedonic analysis of property markets, chiefly because of advances in the ease with which models of this type can be estimated (Kelejian and Prucha, 1999; Bell and Bockstael, 2000).

The final approach and that championed here is to accept that there are spatial features influencing property prices that are not observed by the researcher. Whilst many of these features might be the subtle nuances of location that might adequately be handled by modelling of the nuisance process, others may be substantive spatial features whose

absence from the model is likely to induce missing variable bias in the parameter estimates. For example, properties located close to an abattoir are likely to exhibit considerably deflated market prices. If proximity to abattoirs is not included as a regressor in the estimated hedonic price function then one might conclude that the model is misspecified and that the parameter estimates are unreliable.

In a non-spatial setting the presence of omitted variables presents an almost insurmountable obstacle to the researcher. However, as pointed out by Gibbons and Machin (2003) and Gibbons (2003), where the omitted variables can reasonably be expected to be features of geographical space, a course of action suggests itself. That course of action is to account for the missing covariates by spatially smoothing the data using a nonparametric kernel regression procedure. That is, for each property the influence of its particular location on its price can be estimated as the distance-weighted average of the prices of other properties in its neighbourhood. The hedonic price function can then be estimated by linear regression using the deviations of observed prices and regressors from their expected values at each location. Gibbons and Machin (2003) call this a *Smooth Spatial Effects* (SSE) estimator.

A question that remains is over what spatial area the data should be smoothed. As Gibbons and Machin (2003) point out, this amounts to deciding upon the bandwidth for the kernel used to smooth the data. A larger bandwidth will account for spatial processes operating over a wider area, a smaller bandwidth will account for more localised phenomena. Gibbons and Machin (2003) choose a bandwidth motivated by the concern that spatially smoothing the data over too small an area will impact upon the parameter estimate for the variable that forms the focus of their study (namely, proximity to primary schools). Here we adopt an alternative procedure suggested for use in another context by Ellner and Seifu (2002).

Construction of the spatial correlogram for the regression residuals provides a statistical indication of the area over which spatial correlation is a feature of the data. We assume that our regression model lacks covariates that operate so as to influence property prices over this spatial scale. Our choice of spatial smoothing bandwidth is motivated by the desire to remove the impacts of these missing covariates. The procedure outlined by Ellner and Seifu (2002) involves repeated estimation of the SSE model using progressively larger bandwidths. At each iteration, Moran's *I* statistic is calculated to assess the degree of autocorrelation in the residuals over the spatial scale identified by the correlogram. The optimal bandwidth is selected as that bandwidth at which the computed value of *I* matches its expectation under the hypothesis of uncorrelated residuals. Ellner and Seifu term this the *Residual Spatial Autocorrelation* (RSA) criterion.

The rest of this paper is organised as follows. In Section 2 we introduce the data set that forms the focus of our empirical application. In Section 3 we describe Moran's *I* statistic as a measure of spatial autocorrelation in regression residuals. We also describe the use of Moran's *I* in the construction of the spatial correlogram and apply this procedure to the data. In Section 4 we briefly describe models used to account for the spatial autocorrelation of residuals and introduce the smooth spatial effects estimator. In Section 5 we apply the RSA criterion of Ellner and Seifu (2002) to the data in order to choose the optimal region over which to spatially smooth. We compare the recommendations of this procedure with that of cross-validation; an alternative procedure frequently used to select bandwidths. Finally, we apply statistical tests to determine whether the parameters of the SSE differ significantly from a SED model that does not account for omitted spatial covariates.

2. THE 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[1]. These records indicated selling prices, dates of sales and full property address for each residential property transaction.

The Valuation Office Agency (VOA) provided property characteristics data. The VOA is an executive agency of the Inland Revenue, one of whose main functions is to value property for taxation purposes. In order to perform this function, the VOA maintains a database describing the structural characteristics of every residential property in England.[2] Amongst other details, the VOA provided data on the number of bedrooms and bathrooms in each property, total floor area, the property's age, whether the property was a bungalow or house (flats are not included in the analysis), whether the property was detached, semi-detached, in a terrace or at the end of a terrace, whether the property had central heating and access to off-road parking. Furthermore, the VOA classifies properties according to age and style of construction into one of around 30 property types called Beacon Groups. This information was also recorded as it provides a useful additional indication of property quality that cannot be determined from size and age alone.

Addresses were geolocated using a GIS. Subsequently GIS datasets were used to provide details of the garden area and aspect of each property and to calculate straight line distances, car travel times and walking distances from each property to (dis)amenities including schools, shops[3], railway stations and industrial sites.

When considering the accessibility of properties to shops, any measure based on proximity to only one facility has disadvantages. For example, a property 200m from ten shops is likely to be perceived as having better accessibility than another property 200m from one shop. As a result, measures for access to shops were constructed using a weighted sum of distances to all shops. A similar procedure was used when considering accessibility to primary schools. Recent research suggests that selection procedures for primary school intake that favour local residents can considerably inflate house prices around high performing schools (Gibbons and Machin, 2003).[4] 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

---

[1] 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.

[2] 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.

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

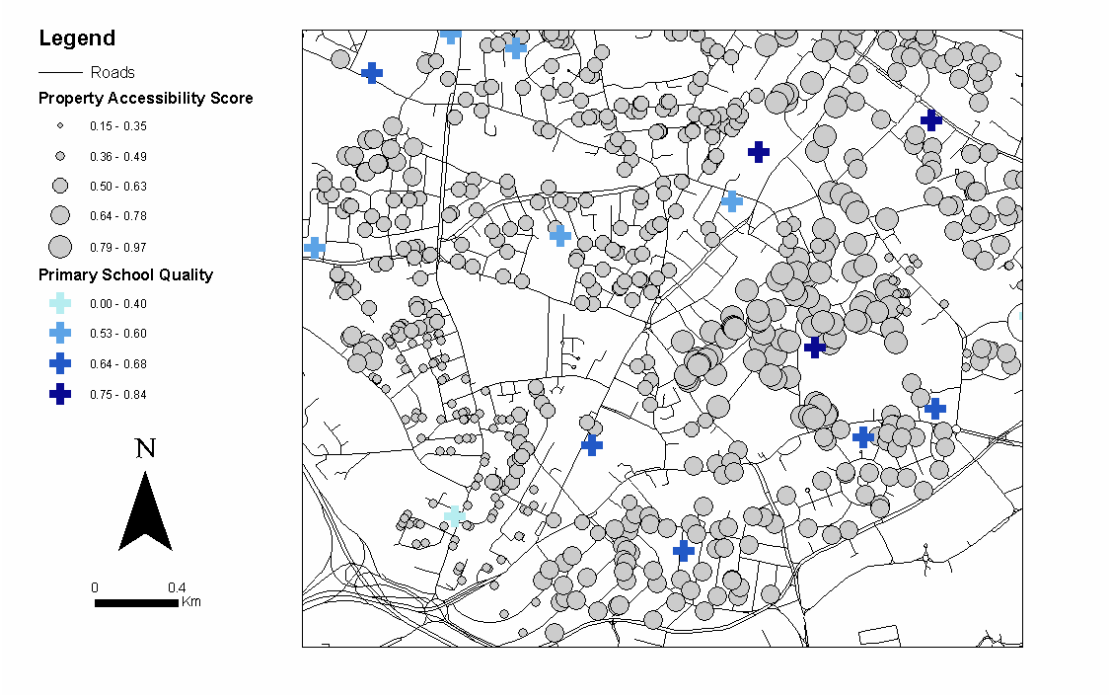
[4] As Gibbons and Machin (2001) argue, the issue is thought less important for secondary schools that typically draw from much wider catchments. Also, high educational achievement at primary school level may be a pre-requisite for admission to selective secondary schools. For example, the five selective 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)

Science, Mathematics and English (the level expected of 11 year olds).[5] A primary school accessibility index was constructed using (2) with the weight α_j set to this measure of school quality and δ = 1. Figure 1 presents the primary school quality/accessibility variable is depicted for a region of the study area.

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

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

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

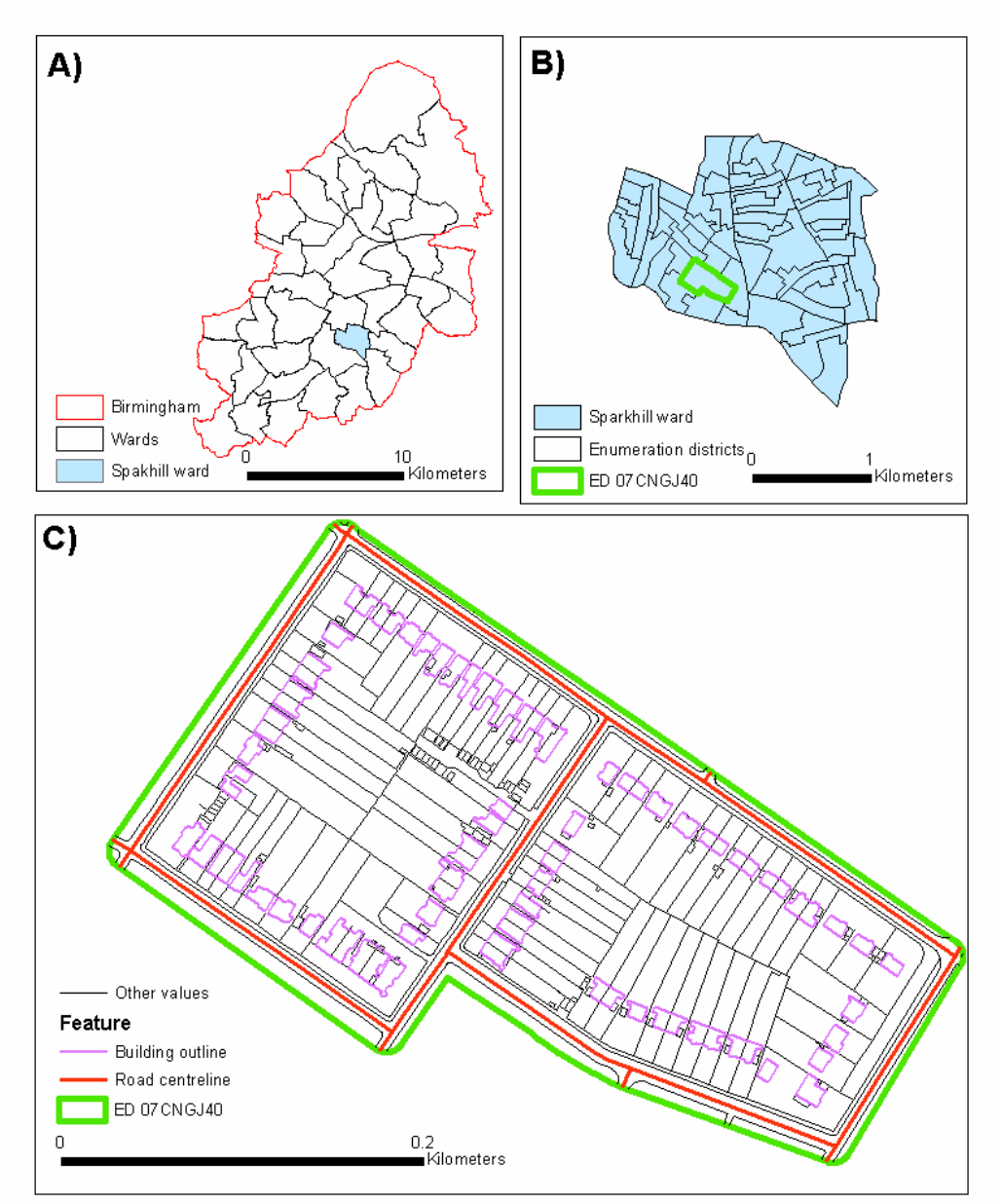


Data on the socio-economic composition of property neighbourhoods were drawn from the 1991 UK census provided by the Office for National Statistics (ONS). For the purposes of this research we recognise two levels of neighbourhoods. The smallest area over which census data is provided by the ONS is an enumeration district (ED). Birmingham is divided into 1,940 EDs, with each ED containing an average of 191 households. EDs are gathered into larger scale political units known as wards. Birmingham contains 39 wards such that each ward comprises an average of 50 EDs and 9,500 households. The organisation of these spatial units are shown in Figure 2.

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

The census provides a myriad of information on the socioeconomic characteristics of the population living in each ED. As described in Day et al. (2004) these variables were subjected to a factor analysis that identified six major dimensions of difference or similarity in the socioeconomic composition of the households inhabiting each ED. The scores for these six factors are included as regressors in the empirical application described below. These scores can be interpreted as capturing (1) increasing age of the population of an ED, (2) increasing proportion of households with children in an ED, (3) increasing poverty of households in an ED, (4) increasing proportion of Asian households in and ED, (5) increasing proportion of black households in and ED and (6) increasing levels of skills of inhabitants of an ED (as defined by educational levels, employment status and occupation).

Figure 2: Hierarchy of administrative areas in Birmingham



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

Table 1: Data Descriptions

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

Variable	Mean	Std. Dev.	Min	Max
End Terrace Bungalow	0.000	0.022	0	1
Terrace Bungalow	0.000	0.017	0	1
Detached House	0.116	0.320	0	1
Semi-Detached House	0.396	0.489	0	1
End Terrace House	0.115	0.319	0	1
Terrace House	0.352	0.478	0	1
Beacon Group (proportion)				
1. Unrenovated cottage pre 1919	0.000	0.019	0	1
2. Renovated cottage pre 1919	0.001	0.027	0	1
3. Small “industrial” pre 1919	0.040	0.195	0	1
4. Medium “industrial” pre 1919	0.226	0.418	0	1
5. Large terrace pre 1919	0.006	0.078	0	1
8. Small “villa” pre 1919	0.020	0.138	0	1
9. Large “villas” pre 1919	0.009	0.093	0	1
10. Large detached pre 1919	0.003	0.058	0	1
19. Houses 1908 to 1930	0.011	0.103	0	1
20. Subsidy houses 1920s & 30s	0.140	0.347	0	1
21. Standard houses 1919-45	0.257	0.437	0	1
24. Large houses 1919-45	0.016	0.124	0	1
25. Individual houses 1919-45	0.000	0.022	0	1
30. Standard houses 1945-53	0.045	0.207	0	1
31. Standard houses post 1953	0.190	0.392	0	1
32. Large houses post 1953	0.032	0.177	0	1
35. Individual houses post 1945	0.001	0.038	0	1
36. “Town Houses” post 1950	0.004	0.062	0	1
Sale Date (proportion)				
1st Quarter (Jan. to Mar.)	0.214	0.410	0	1

Variable	Mean	Std. Dev.	Min	Max
2nd Quarter (Apr. to June)	0.247	0.431	0	1
3rd Quarter (July to Sept.)	0.287	0.452	0	1
4th Quarter (Oct. to Dec.)	0.252	0.434	0	1
Neighbourhood Characteristics				
Poverty Factor	-0.375	0.855	-1.934	2.363
Skills Factor	0.180	1.000	-1.398	4.198
Age Factor	0.055	0.807	-3.216	3.143
Family Factor	-0.029	0.842	-3.198	3.791
Asian Factor	-0.045	0.942	-1.131	5.152
Black Factor	-0.240	0.750	-2.016	8.214
Locational Characteristics				
Proximity to City Centre (mins)	1,313	478	208	3,187
Proximity and Quantity of Shops	2.276	1.273	0.07	9.56
Proximity and Quality of Primary Schools	0.602	0.177	0.15	0.97
Walking time to Rail Station (mins)	1,846	1,013	21.05	5,525
Walking time to a Park (mins)	900	558	3.17	3,425
Driving time to Airport (mins)	2,388	655	602	4,385
Proximity to A-Type Industrial Processes (m)	2,464	1,820	21.94	10,204
Proximity to B-Type Industrial Processes (m)	814	527	10	3,333
Proximity to Land Fill sites (m)	946	608	10	3,472
Environmental Characteristics				
Views of Water (weighted m^2)	0.480	7.543	0	348
Views of Parkland (weighted m^2)	6.290	36.831	0	664
Road Traffic Noise (dB)	49.8	9.4	31.6	75.8
Rail Traffic Noise (dB)	36.8	12.6	0	74.7
Aircraft Noise (dB)	4.8	16.0	0	69

3. ASSESSING SPATIAL AUTOCORRELATION IN REGRESSION RESIDUALS

In this section, we describe a procedure designed to assess the nature of spatial autocorrelation present in regression residuals. This procedure is illustrated through the estimation of hedonic price functions using the data set described in the last section. Drawing on the analysis in Day et al. (2004), the data is partitioned into seven clusters of properties. Each cluster of properties is defined by the similarity of the socioeconomic composition of the neighbourhoods in which those properties are located. For each cluster we estimate a simple linear regression;

$$\ln \boldsymbol{P}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \qquad\qquad j = 1, 2, ..., M$$

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

Since it adds nothing to the discussion, let us simplify notation by dropping the cluster index, *j*. Further, to allow a more generic discussion let us replace the regressand $\ln P$ with the nonspecific vector of dependent variables ***y***, giving;

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

Our null hypothesis is the absence of spatial autocorrelation in the regression residuals. That is we assume that;

$$E\left[\boldsymbol{\varepsilon}\right] = \boldsymbol{0} \qquad \text{and} \qquad E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N\right) \tag{2}$$

In effect, the null is to assume that the regression residuals are distributed randomly across space. That is to say, any observed value of the regression residual could occur at any location with equal likelihood. In this case, ordinary least squares (LS) will return consistent and efficient estimates of the parameters of the model.

The alternative hypothesis is that the regression residuals exhibit spatial autocorrelation. To test for such autocorrelation, the researcher must stipulate the nature of the possible spatial dependence by specifying an $N \times N$ weighting matrix, ***W***. The diagonal elements of the weighting matrix are zero since, clearly, we are not concerned with testing the correlation of residuals with themselves. The off-diagonal elements of the matrix stipulate 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 residuals 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. One commonly followed convention is to assume that observations separated by greater than some distance, *d*, are unrelated. So, if the i^{th} and j^{th} observation are separated by less than *d*, the w_{ij}^{th} element of ***W*** is initially set to a value of one, otherwise that element is set to zero. As we shall discuss shortly, the choice of *d* is of considerable importance.

A number of test statistics have been devised to test for spatial autocorrelation in regression residuals. These include the extension to Moran's *I* statistic (Moran, 1950) proposed by Cliff and Ord (1972), tests based on the Lagrange multiplier principle (e.g. Burridge, 1980; Anselin 1988) and a specification robust approach suggested by Kelejian and Robinson

(1992). Here we adopt the approach of Ellner and Siefu (2002) and employ Moran's *I* statistic as our test of spatial autocorrelation. As Hepple (1998) describes, Moran's *I* provides a general-purpose test capable of detecting most forms of spatial pattern.

Let $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the regression residuals when $\hat{\boldsymbol{\beta}}$ is the LS estimator of $\boldsymbol{\beta}$, then Moran's *I* statistic is given by;

$$I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_j w_{ij}}{\sum_{i=1}^{N}\hat{\boldsymbol{\varepsilon}}_i^2}\left(\frac{N}{S_0}\right) = \frac{\hat{\boldsymbol{\varepsilon}}'\boldsymbol{W}\hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}\left(\frac{N}{S_0}\right) \tag{3}$$

where *N* is the number of observations and $S_0 = \sum_i \sum_j w_{ij}$, the sum of all the elements in the weights matrix.

The numerator in Equation (3) is a cross-products (covariance) term, while the denominator is a variance term. As such *I* behaves as a product-moment correlation, varying on the interval [-1,1], with 1 indicating perfect positive correlation of residuals and -1 indicating perfect negative correlation of residuals. The significance of non-zero *I* can be judged by comparison with the distribution of *I* under the null hypothesis of residuals that are randomly distributed over space.[6] Cliff and Ord (1972, 1973) showed that in large samples, this distribution was approximately normal and developed formulas for its mean and variance (see Anselin and Hudak, 1992). The statistical and analytical power of the test has been confirmed by numerous Monte Carlo studies (e.g. Bartels and Hordijk, 1977; Brandsma and Ketellapper, 1979; Anselin and Rey, 1991). Whilst Hepple (1998) has developed the exact distribution of the *I* statistic, here we continue to use the Cliff-Ord normal approximation due to the comparative simplicity of its calculation.

Our major concern in this section is the choice of an optimal value *d* to use in testing for spatial autocorrelation. That is, we wish to define a statistical procedure that indicates the area over which spatial autocorrelation of residuals is a feature of the data. As discussed in the introduction we assume that our regression model lacks covariates that operate so as to influence property prices over this spatial scale.

To a greater extent, researchers in the hedonic literature have not concerned themselves with the choice of *d*. Indeed in testing for spatial autocorrelation, or for that matter modelling spatial autocorrelation, *d* is generally chosen in some *ad hoc* manner. For example, Bell and Bockstael (2000) choose a value of 600m since this is the average size of housing developments in the area. Likewise, Ihlanfeldt and Taylor (2001) choose a distance of 3 miles since this is the smallest distance to guarantee that all observations have at least one neighbour and because "this distance seems sufficiently large to allow for almost any type of spatial dependence".

Here we make use of the *correlogram*, more familiar to economists for its application in times series econometrics. We calculate Moran's *I* for a series of lag distances (or distance

[6] If we believe that the residuals are randomly distributed over space then the value of *I* in Equation (2) is only a single value out of a possible *N*! values that could be found if the residuals were randomly reallocated over observations and *I* recalculated. Indeed, if we were to graph the density of the *N*! possible values of *I* we would produce a distribution from which a standard error could be obtained. If the particular value of *I* is found to be a rare occurrence under randomisation then it can be inferred that some pattern of spatial autocorrelation exists in the data.

classes) from each point by specifying a weighting matrix that assigns a value of one to pairs of observations separated by a distance that falls within that class, and a value of zero otherwise. The resulting *spatial* correlogram illustrates the degree of autocorrelation at each lag distance. Dubin (1988, 1992, 1998) follows a similar procedure to construct spatial correlograms for residuals from hedonic price regressions for property market data. Here however, we adopt a more sophisticated approach inspired by the paper of Ellner and Seifu (2002). We plot on the same correlogram the expected value and the 95% confidence intervals of the distribution of Moran's *I* under the null hypothesis of random distribution of residuals over space (using the formulas of Cliff and Ord, 1973). We take *d* as being the distance class at which the correlogram falls within the 95% confidence interval of random spatial distribution of residuals.

Our procedure differs from that of Ellner and Seifu (2002) in that they do not calculate a correlogram. Rather they plot the value of *I* for weights matrices defined by progressively larger values of *d*. We prefer our approach since the presence of substantial autocorrelation at small values of *d* may dominate the value of Moran's *I* statistic when calculated for more inclusive values of *d*. Thus the value of Moran's *I* statistic may remain significant at larger values of *d* even if the more distant observations bought into the calculation by extending *d* are not actually correlated.

In the application described here we estimate two specifications of the regression model in (1). In the first, the regressor matrix, $\boldsymbol{X}$, includes the multiplicity of structural, neighbourhood, environmental and locational variables listed in Table 1. The correlograms for this specification are plotted in Figure 3. In the second specification we include a set of locational constants. The constants indicate in which of the 39 wards each property is located (see Figure 2). As discussed in the introduction, these wide-area locational constants constitute a crude attempt to capture spatial variation in property prices that is not accounted for by the other regressors included in the hedonic analysis. The correlograms for this specification are plotted in Figure 3.

The correlograms are calculated for 100m distance classes. In Figures 2 and 3 the value of Moran's *I* (and it's expectation and 95% confidence band under random distribution of errors) for a distance class is plotted at the upper limit of the class. The value for *d*, therefore, is taken as the upper boundary of the largest distance class to fall outside the 95% confidence bands such that the *I* statistics for successive distance classes fall consistently with these confidence bands. For example, in Figure 6 the last vertex of the correlogram for Cluster 6 to fall outside the 95% confidence bands is that for the 500m to 600m distance class. Subsequent distance classes return *I* statistics that are not significantly differently from what might be expected under random distribution of residuals. In this case, *d* is taken to be 600m. Not all cases are as clear cut. The correlogram for Cluster 5 in Figure 3 dips into the 95% confidence bands for the 300m to 400m distance class but subsequent classes return *I* statistics evidencing statistically significant spatial correlation. In this case *d* is taken to be greater than 1000m (the highest value plotted on the correlograms).

Some details of the various regressions for the two specifications and values for *d* are reported in Table 2 (full regression results are reported in Appendices A and B at the end of this paper). It is immediately clear from these statistics, that including the locational constants considerably improves the specification of the model. For all seven partitions of the data the adjusted R^2 statistic is seen to increase with the inclusion of the locational constants (ranging from a minimum increase of 1.4% to a maximum of 4%, with an average across all seven clusters of 2.5%). The final two columns of Table 2 report an *F*-test of the significance of the locational constants. In all cases, the locational constants prove to be highly significant. These findings must be treated with caution as the *F*-test is only appropriate if there is no spatial autocorrelation in the residuals.

That spatial autocorrelation is present in the regression residuals is immediately evident from the correlograms in Figures 3 and 4. For all clusters in both specifications of the model the *I*

statistics indicate significant autocorrelation of the residuals over at least the first distance band (0 to 100m). Furthermore, comparing the correlograms in Figure 3 with those in Figure 4 underscores the importance of including wide-area locational constants. In the models with no locational constants, spatial autocorrelation remains an important feature of the data even for remote distance classes. In Clusters 2, 5 and 7 for example, there is significant spatial correlation for the largest distance class plotted on the correlograms (900m to 1km).

The introduction of the wide-area locational constants does much to improve matters. Indeed, examination of the correlograms in Figure 4 reveals that the introduction of wide-area locational constants virtually eliminates autocorrelaion for distance classes greater than 300m. However, for all clusters the correlograms reveal significant evidence of more localised autocorrelation of the regression residuals.

Table 2: Statistics from hedonic regressions for each cluster with and without spatial constants

Cluster	*N*	Regressions without locational constants			Regressions with locational constants			F-test of spatial constants	
		K	*Adj. R^2*	*d* (metres)	*K*	*Adj. R^2*	*d* (metres)	*F*-stat (df)	*p*-value
Cluster 1	2261	64	0.685	900	96	0.709	200	6.69 (32, 2165)	<.0001
Cluster 2	1258	63	0.777	>1000	90	0.817	300	10.66 (27, 1168)	<.0001
Cluster 3	2173	63	0.776	500	96	0.791	300	5.44 (33, 2077)	<.0001
Cluster 4	895	61	0.771	200	93	0.785	100	2.64 (32, 802)	<.0001
Cluster 5	2018	63	0.751	>1000	97	0.779	200	8.27 (34, 1921)	<.0001
Cluster 6	1207	60	0.810	800	85	0.836	200	8.34 (25, 1122)	<.0001
Cluster 7	970	62	0.787	500	82	0.813	100	7.28 (20, 888)	<.0001

Figure 3: Spatial correlograms for residuals from regressions not including spatial constants

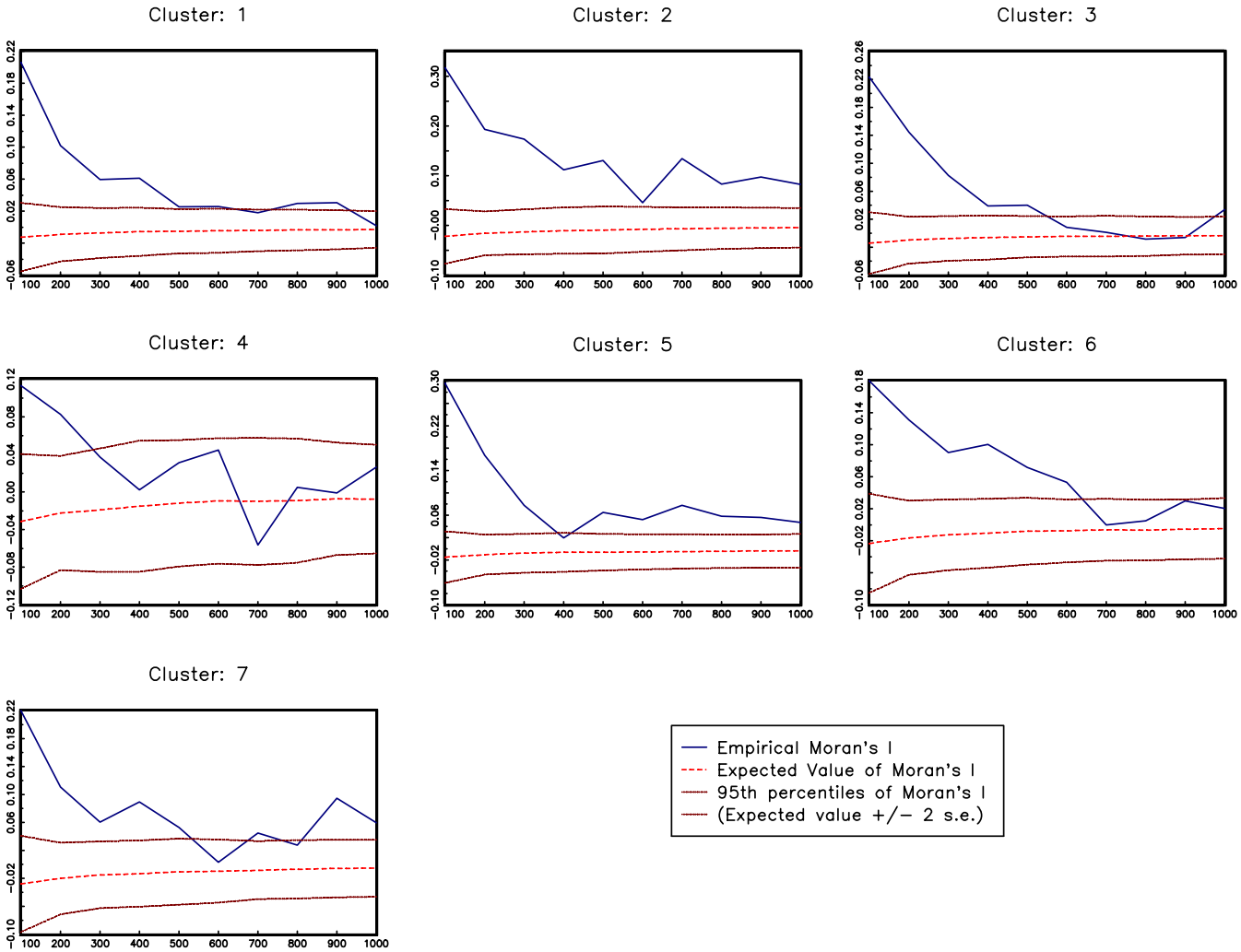
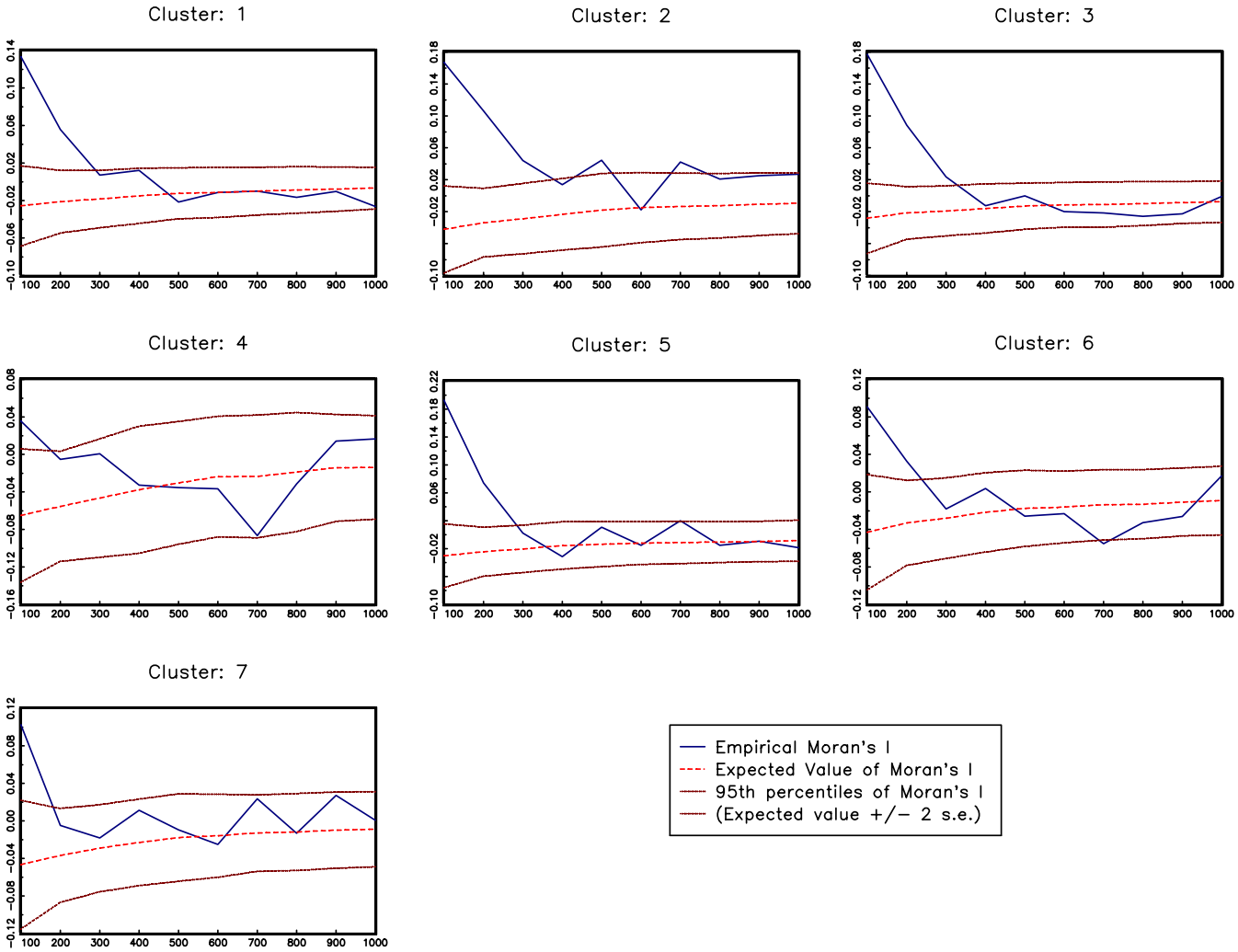


Figure 4: Spatial correlograms for residuals from regressions including spatial constants



4. LINEAR REGRESSION WITH SPATIALLY CORRELATED RESIDUALS

The correlograms in Figure 4 reveal that the regression residuals are spatially correlated over a region of up to 300m. As such we can reject the model described by Equations (1) and (2). As described in the introduction three broad approaches to dealing with spatial autocorrelation have been proposed in the literature. We shall briefly review these in this section.

The first approach is to assume that one has data on all relevant determinants of property prices and that spatial autocorrelation of the residuals is merely an artefact of misspecifcation of the functional form of the hedonic price equation. For example, Can (1990, 1992) argues that the model in (1) is misspecified because parameter estimates are not constant over the urban landscape. Rather they are assumed to drift over space as a function of a set of regressors describing characteristics of different locations. As such, Can partitions the regressors into two sets, the $N \times K_1$ matrix $\boldsymbol{Z}_1$ and the $N \times K_2$ matrix $\boldsymbol{Z}_2$. In Can's specification $\boldsymbol{Z}_1$ comprises variables describing the socioeconomic composition of neighbourhoods whilst $\boldsymbol{Z}_2$ comprises variables describing the structural characteristics of properties. Can assumes that the parameters estimated on the $\boldsymbol{Z}_2$ regressors are not constant but vary according to the values taken by the regressors in $\boldsymbol{Z}_1$. This assumption results in what Can describes as the *spatial expansion* specification;

$$y_i = \alpha + \sum_k z_{1ki}\beta_k + \sum_l \sum_k \left(\gamma_{k0} + \gamma_{kl} z_{1ki}\right) z_{2li} + \varepsilon_i \qquad i = 1, 2, ..., N \qquad (4)$$

where i indexes property observations, y_i is i^{th} element of $\boldsymbol{y}$ (e.g. the price of the i^{th} property), z_{1ki} is the i^{th} observation of the k^{th} variable in the $\boldsymbol{Z}_1$ matrix, z_{2li} is the i^{th} observation of the l^{th} variable in the $\boldsymbol{Z}_2$ matrix and α, β and γ are the parameters to be estimated.

A natural extension of the spatial expansion specification is proposed by Pavlov (2000). Again, Pavlov assumes that the coefficients of a linear hedonic function vary across the urban space. However, rather than specifying a functional relationship between the spatially varying coefficients and a set of locational variables, Pavlov allows the value of the coefficients to be determined by the data. The space varying coefficients are made functions of locations according to the *space-varying coefficients* (*SVC*) specification;

$$y_i = \alpha_i\left(c_{1i}, c_{2i}\right) + \sum_k \beta_{ki}\left(c_{1i}, c_{2i}\right) x_{ki} + \varepsilon_i \qquad i = 1, 2, ..., N \qquad (5)$$

where α_i is a space varying constant specific to the location of the i^{th} observation as defined by its coordinates $\boldsymbol{c}_i = \left(c_{1i}, c_{2i}\right)$. Likewise, β_{ki} is a space-varying coefficient specific to the location of the i^{th} observation. An estimate of the coefficients at any particular location is made using weighted least squares. Only the m nearest observations to the location of interest receive non-zero weights in this regression. Greater weight is attributed to observations more proximal to the location of interest according to the Epanechnikov weighting scheme (Epanechnikov, 1969). The *SVC* method allows for both the intercept and the slope parameters of the hedonic price function to differ by location. However, this ability to handle spatial processes in the data comes at a cost. As Pavlov (2000) points out, the space-varying coefficients method lacks a theoretical inferential framework. Since the parameters of the hedonic vary continuously over space it is not possible to judge the statistical significance of any particular regressor in determining property prices.

Another respecification of the hedonic model that specifically accounts for spatial processes is the *spatial autoregressive* model. This model has been studied variously by Anselin (1989), Can (1992), Can and Megbolugbe (1997) and Gawande and Jenkins-Smith (2001). In the spatial autoregressive model the price of a property is deemed to be determined, in part, by the prices of neighbouring properties according to;

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

where $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are defined as previously, $\boldsymbol{W}$ is the $N \times N$ spatial weighting matrix, $\boldsymbol{y}$ is the $N \times 1$ vector of property prices, and ρ is the spatial autoregressive coefficient. Can (1990) argues that this specification has some merits since it mimics the actual workings of the property market in which estate agents appraise the value of a property according to both its own attributes and the price history of houses in the neighbourhood.

A second approach is to assume that the true model is the model at hand but that autocorrelation among the disturbances is due to spatial dependence in the process generating the nuisances. We call approaches that make this assumption *spatial error dependence* (SED) models. The SED approach has become increasingly popular in applied work, chiefly because of advances in the ease with which models of this type can be estimated (Kelejian and Prucha, 1999; Bell and Bockstael, 2000).

The consequence of a spatially dependent nuisance process is that the observations contain less information than if they had been independent. Indeed the statistical properties that are attributed to an estimator such as LS when errors are i.i.d. do not hold in this case. Nonetheless, the parameter estimates from the application of LS will not be biased, merely inefficient. In this case, the proscribed course of action is to model the nuisance process so as to obtain approximately the same quantity of information as provided by an independent set of observations.

For example, we might assume that the autocorrelation follows the first order Markovian scheme;

$$\boldsymbol{\varepsilon} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{u} \tag{7}$$

or equivalently;

$$\boldsymbol{\varepsilon} = \left(\boldsymbol{I}_N - \lambda \boldsymbol{W}\right)^{-1} \boldsymbol{u} \tag{8}$$

where λ is the error dependence parameter and $\boldsymbol{u}$ is the usual $N \times 1$ vector of random error terms with expected value zero and variance-covariance matrix $\sigma^2\boldsymbol{I}$. Notice that $\lambda = 0$ implies $\boldsymbol{\varepsilon} = \boldsymbol{u}$ and there is no spatial dependence in the data. This particular model has been studied by various authors including Pace and Gilley (1997), Kelejian and Prucha (1999), Bell and Bockstael (2000) and Leggett and Bockstael (2000). Along similar lines, Dubin (1988, 1992, 1998) and Basu and Thibodeau (1998) develop explicit models of the nuisance process and estimate the parameters of the nuisance process and the regression coefficients simultaneously using maximum likelihood.

Of course, SED models impose considerable structure on the processes determining spatial correlation in regression residuals. For example, they assume isotrophy. That is they assume the same model of error dependence can be applied over all space. Furthermore, spatial autocorrelation of the regression residuals is induced by locational features influencing property prices that are not observed by the researcher. SED models assume

that these comprise the subtle nuances of location that might adequately be handled by modelling the nuisance process. Alternatively, the omitted spatial covariates may be substantive features whose absence from the model is likely to induce missing variable bias in the parameter estimates.

In a non-spatial setting the presence of omitted variables presents an almost insurmountable obstacle to the researcher. However, as pointed out by Gibbons and Machin (2001, 2003) and Gibbons (2002, 2003), where the omitted variables can reasonably be expected to be features of geographical space, a course of action suggests itself. Gibbons and Machin (2003) propose the *smooth spatial effects* (SSE) estimator which they specify as;

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + q(\boldsymbol{c}_i) + \boldsymbol{\varepsilon}_i \qquad i = 1, 2, ..., N \tag{9}$$

where $\boldsymbol{x}_i$ is the vector of observed regressors for the i^{th} observation, $\boldsymbol{c}_i = (c_{1i}, c_{2i})$ is the coordinates vector establishing the location of the i^{th} observation in space and $q(\cdot)$ is some unknown function. In effect, the specification in (9) replaces the unobserved spatial covariates with an element that is a function of location. The influence of these unobserved covariates on property prices is determined by the unknown function $q(\cdot)$. Since the influence of $q(\cdot)$ on property prices is handled nonparametrically the SSE presents an extremely flexible approach to dealing with unobserved spatial covariates.

Equation (9) is a specific example of a more general class of semiparametric models known as *partially linear models*. In that context, Robinson (1988) shows that (9) can be rewritten as;

$$y_i - E[y \mid \boldsymbol{c}_i] = (\boldsymbol{x}_i - E[\boldsymbol{x} \mid \boldsymbol{c}_i])\boldsymbol{\beta} + \varepsilon_i \tag{10}$$

suggesting that □ can be estimated in a two-step procedure;

- First, the unknown conditional means $E[y \mid \boldsymbol{c}_i]$ and $E[\boldsymbol{x} \mid \boldsymbol{c}_i]$ are estimated using a nonparametric estimation technique.
- Second, the estimates are substituted in place of the unknown functions in Equation (10) and ordinary regression techniques employed to estimate □.

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. That is, estimating Equation (10) is asymptotically equivalent to knowing both the values taken by the missing spatial covariates and knowing how these covariates impact on property prices.

In the hedonic literature, Robinson's model has been employed in a slightly different context by Anglin and Gençay (1996). Both Gibbons and Machin (2003) and Anglin and Gencay (1996) employ the Nadaraya-Watson nonparametric estimator to determine the quantities $E[y \mid \boldsymbol{c}_i]$ and $E[\boldsymbol{x} \mid \boldsymbol{c}_i]$. Notice that these quantities are simply the expected values of *y* and ***x*** at a particular location. In effect, the Nadaraya-Watson estimator calculates these expectations by taking the weighted average of the values of observations close to that location. Whether an observation is considered close to the location is determined by the bandwidth parameter *b*. The larger the value taken by *b*, the more observations are drawn into the calculation of the average. Further, the weight allotted to each observation in the calculation of the local average is determined by the kernel function. The kernel function must be symmetric, continuously differentiable and integrate to unity. Moreover, most

commonly used kernel functions allot greater weight to observations that are in close proximity to the location than to those that are further away.

An alternative to the Nadaraya-Watson approach is to employ *local linear* estimators which offer significant gains especially at the boundaries of the data and when the data is not equally spaced (see Fan, 1992 or Hastie and Loader, 1993, for more detailed discussion). Furthermore, as we shall discuss shortly, our estimation strategy requires repeated nonparametric estimation of the quantities $E\left[y \mid \boldsymbol{c}_i \right]$ and $E\left[\boldsymbol{x} \mid \boldsymbol{c}_i \right]$. Since nonparametric regression can be extremely time-consuming and computer-intensive, we employ fast implementation techniques as described in Fan and Marron (1994), Wand (1994) and Bowman and Azzalini (2003).

In particular, we begin by summarising the density of observations over space by linearly binning onto the verteces of a regular spatial grid. In this application the margins of the cells of the grid are set to 150m. Likewise we summarize the values of *y* and each of the variables present in ***x*** by calculating their linearly weighted averages at each of the verteces of the grid. Furthermore, to take advantage of computational savings offered by the use of the fast Fourier transform (FFT) we choose to use a bivariate Gaussian kernel function. Given a choice of smoothing bandwidth *b*, the expected values of *y* and ***x*** are calculated at each vertex of the grid using local linear regression. Finally, the values at each particular property location are recovered by linearly interpolating from the values of the four most proximate verteces of the grid. Since the data set is relatively large, binning the data and employing FFT-based calculations was found to be many times quicker than employing a naïve implementation of local linear regression.

5. CHOICE OF SPATIAL SMOOTHING PARAMETER

A question that remains is the choice of smoothing bandwidth, *b*. A larger bandwidth will account for spatial processes operating over a wider area, a smaller bandwidth will account for more localised phenomena. For example, observe Figures 5 and 6. These provide plots of $E[y \mid \boldsymbol{c}_i]$, that is the expected value of $\ln P$ at a given location, for two different smoothing bandwidths. Notice first that in both cases, there is considerable variation in $E[y \mid \boldsymbol{c}_i]$ across the urban area. In particular, notice the substantial peak in the north-east section of the plots (that is, towards the back right of the cube in the figures). These peaks correspond to the desirable north-eastern suburbs of the City of Birmingham. Notice further that using a larger bandwidth as in Figure 5, results in a simpler, less convoluted surface than using a smaller bandwidth as in Figure 6. Notice that using a smaller bandwidth, as in Figure 6, brings to light possibly important local features of the data that had previously been masked by the use of the larger bandwidth.

Gibbons and Machin (2003) choose a bandwidth motivated by the concern that spatially smoothing the data over too small an area will impact upon the parameter estimate for the variable that forms the focus of their study (namely, proximity to primary schools). Here we select a bandwidth using the *Residual Spatial Autocorrelation* (RSA) criterion suggested for use in a slightly different context by Ellner and Seifu (2002).

The logic behind Ellner and Seifu's procedure is very simple. In section 3 we discussed how spatial statistics could be used to assess optimal *d*; that is, the area over which spatial autocorrelation of residuals is a feature of the data. Having established that the residuals show evidence of spatial autocorrelation, we conclude that our regression model lacks covariates that operate so as to influence property prices over the spatial scale given by *d*. Consequently, the RSA procedure is to search across different smoothing bandwidths, *b*, and for each bandwidth calculate the degree of spatial autocorrelation of the residuals over an area *d*. Acceptable smoothing bandwidths are those for which we can reject the hypothesis of spatial autocorrelation in the residuals.

Figure 7 plots the value of Moran's *I* statistic for values of *b* at 25m intervals between 300m and 1,800m for each cluster. Also plotted in Figure 7 are the expected values of Moran's *I* and the 95% confidence intervals for the statistic under the assumption of randomly distributed residuals. The optimal spatial smoothing bandwidth is chosen as that at which Moran's *I* statistic is approximately equal to its expected value. This bandwidth is reported in Table 3 along with the upper and lower values for *b* at which it is still possible to reject the hypothesis of spatially autocorrelated residuals.

Figure 5: Spatial smoothing with 1800m bandwidth (plotted on a 600m grid from the South-West)

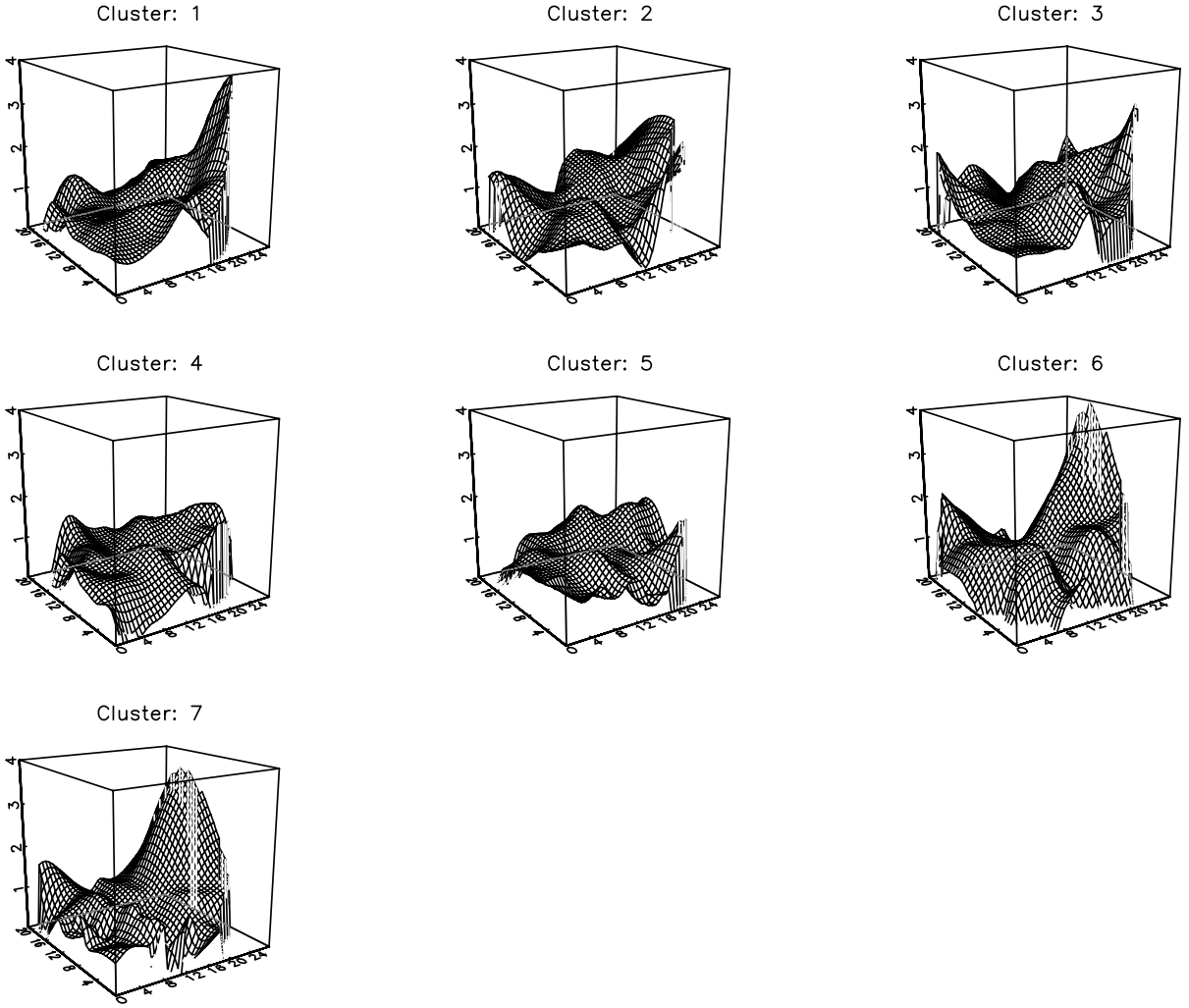


Figure 6: Spatial smoothing with 1200m bandwidth (plotted on a 600m grid viewed from the South-West)

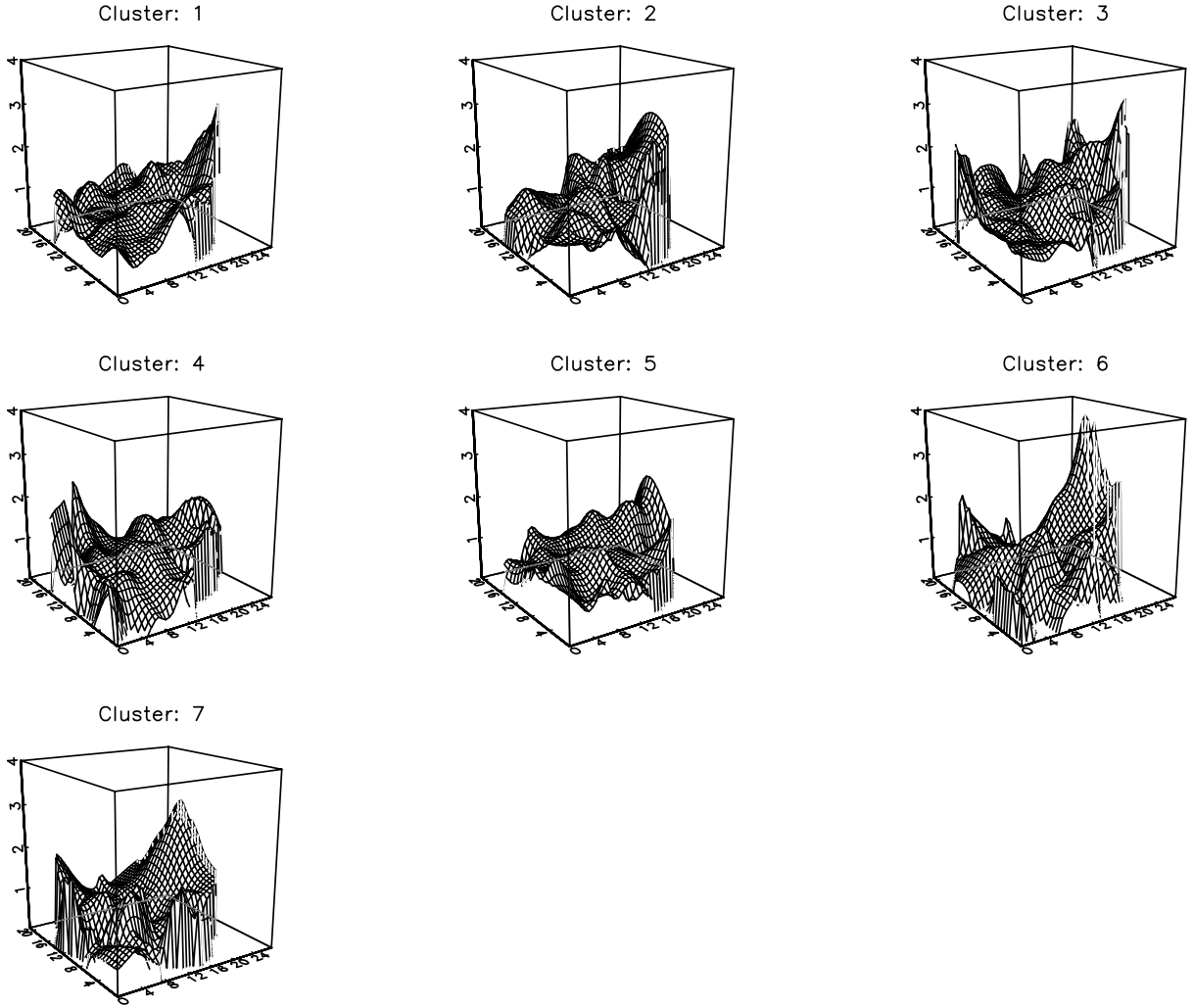


Table 3: Bandwidth choice using RSA and cross-validation criteria

Cluster	RSA			Cross-Validation
	Optimal	Lower Bound	Upper Bound	Optimal
Cluster 1	550	425	725	1050
Cluster 2	675	525	875	525
Cluster 3	600	450	750	575
Cluster 4	650	375	1800	1150
Cluster 5	450	300	550	775
Cluster 6	625	450	950	600
Cluster 7	400	<300	850	1025

A commonly applied alternative for choosing bandwidths is cross-validation. For example, this procedure was applied by Anglin and Gencay (1998) in choosing the degree of smoothing in their semiparametric estimator for a hedonic price model. 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 equation;

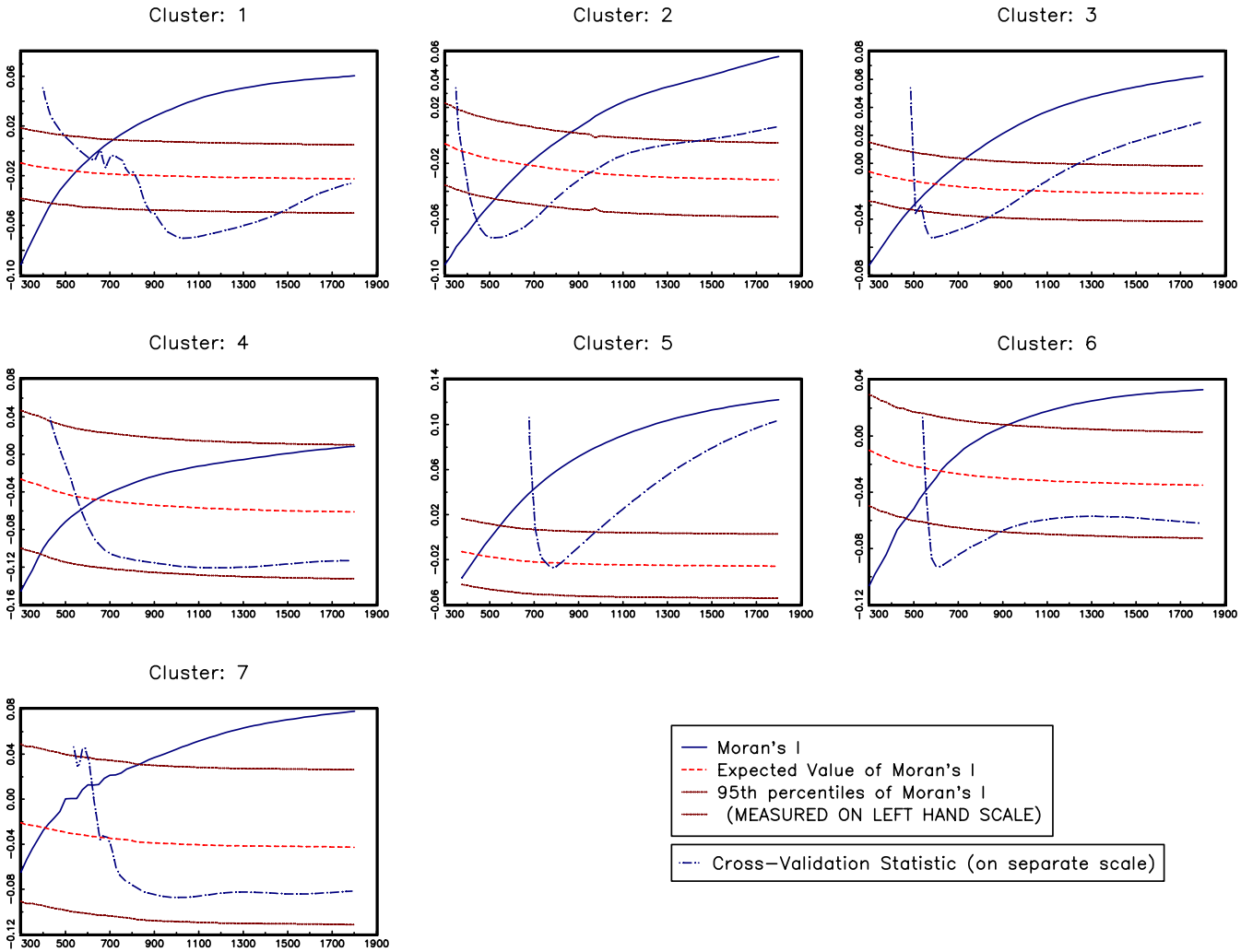
$$MSE = n^{-1}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - q_b\left(\boldsymbol{c}_i\right)\right)^2 \tag{11}$$

where *MSE* stands for Mean Square Error. Of course we don't know the true value of $q_b\left(\boldsymbol{c}_i\right)$ but we can estimate it by applying local linear regression to the quantities $\left(y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)$, where $\hat{\boldsymbol{\beta}}$ are the SSE estimates of the parameter values using a bandwidth of *b*. Unfortunately, there is a problem with such a procedure; for any *b* smaller than the closest two data points in the sample, the MSE reduces to zero. For such values of *b* the conditional mean function given by $q_b\left(\boldsymbol{c}_i\right)$ puts all weight on the i^{th} observation such that $q_b\left(\boldsymbol{c}_i\right)$ perfectly predicts y_i.

Accordingly, the criterion function in (11) cannot be used to decide upon the optimal bandwidth. Rather researchers employ the cross-validation statistic;

$$MSE_{CV} = n^{-1}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - q_{b,i}\left(\boldsymbol{c}_i\right)\right)^2 \tag{12}$$

Figure 7: Moran's *I* and cross-validation statistics for various spatial smoothing bandwidths by cluster



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

The cross-validation procedure, therefore, is to carry out a grid search for optimal *b*. Equation (10) is re-estimated numerous times using different values of *b*. For each value of *b* the cross-validation statistic is estimated using (12) and the *b* providing the minimum value for this statistic is chosen as the optimal bandwidth. The cross-validation statistics for bandwidths at 25m intervals between 300m and 1,800m for each cluster are plotted in Figure 7 and the optimal values recorded in the final column of Table 3. The cross-validated bandwidth can be demonstrated to be asymptotically optimal with respect to MSE (Härdle and Marron, 1985).

Notice that in four of the clusters (2, 3, 4 and 6), the bandwidth selected using cross-validation falls within the 95% confidence bounds of Moran's *I*. That is, if we were to smooth the data in these clusters using the optimal cross-validation bandwidth we would find that we could reject the hypothesis of autocorrelation in the residuals over an area of radius *d*. In the remaining three clusters (1, 5 and 7) cross-validation indicates a higher value for *b* than would be chosen by selecting a bandwidth according to the RSA criterion.

We test to see whether choosing the bandwidth through cross-validation rather than by the RSA criterion makes a difference. As pointed out by Gibbons (2001) sensitivity to bandwidth choice can be tested by the usual Hausman test for equivalence of parameters in alternative estimators. Denote by $\hat{\beta}^w$ the estimator using the wider bandwidth that is consistent under both the null and the alternative hypotheses, and by $\hat{\beta}^n$ the estimator using the narrower bandwidth that is fully efficient under the null but inconsistent if the null is not true. The Hausman statistic is given by;

$$\tau_H = \left(\hat{\beta}^n - \hat{\beta}^w\right) Var\left(\hat{\beta}^n - \hat{\beta}^w\right)^{-1} \left(\hat{\beta}^n - \hat{\beta}^w\right) \qquad (13)$$

As Hausman (1978) shows, under the null hypothesis, the middle term in (13) (the variance matrix of the vector of differences between the parameters of the two estimators) asymptotically reduces to $Asy.Var\left(\hat{\beta}^n\right) - Asy.Var\left(\hat{\beta}^w\right)$. Of course, to make use of the Hausman result one must be able to consistently estimate the asymptotic variance matrices of the two sets of parameter estimates under the null. Unfortunately, in the presence of spatial autocorrelation of unknown form such an estimate is unavailable. Consequently, we apply a bootstrap procedure to estimate $Var\left(\hat{\beta}^n - \hat{\beta}^w\right)$. We sample with replacement from the unsmoothed data and re-estimate the SSE model using the bandwidths implied by first the RSA criterion and then cross-validation. For each bootstrap sample we calculate the difference between the two vectors of parameters. The desired variance matrix is estimated by calculating the empirical variance matrix of the differences resulting from 1,000 replications of the bootstrap procedure.[7]

The Hausman test statistics reported in Table 4 are based on a subset of the regression parameters. We do not include the parameters for locational constants in the tests since these prove to be somewhat unstable in the SSE model where much of their influence is

[7] Since we are estimating a variance matrix and not the tails of a distribution (as is usually the case with bootstrap procedures) we do not require a very large number of replications. Even with 1,000 replications the bootstrap took nearly 12 hours to run for each cluster on a PC with a 2.8 GHz Intel Pentium 4 processor with 512 Mb of RAM. The bootstrap would have been unfeasible without the application of the fast local linear regression procedures described in Section 4.

obviated by spatial smoothing of the data. Furthermore, the model specification includes numerous sets of dummy variables detailing categorical descriptors of property attributes (e.g. numbers of bathrooms, bedrooms, storeys etc.). When particular categories in these dummy variable sets are poorly represented in the data, it may prove impossible to estimate all the parameters of the model for every bootstrap sample. The tests are based on all parameters that are successfully estimated for each iteration of the bootstrap.

Table 4: Comparison of bandwidth choice using RSA and cross-validation criteria

Cluster	Hausman Test		
	Statistic	df	p-value
Cluster 1	65.72	52	0.067
Cluster 2	46.74	48	0.525
Cluster 3	60.03	52	0.208
Cluster 4	31.09	46	0.955
Cluster 5	75.93	50	0.010
Cluster 6	40.31	48	0.777
Cluster 7	29.94	47	0.975

The Hausman test reveals that significant differences in the parameters can be discerned in only two of the clusters; Cluster 1 (at greater than 90% confidence) and Cluster 5 (at greater than 95% confidence). In general then, our data suggests that choosing a bandwidth using the RSA criterion does not result in parameter estimates that differ significantly from those estimated using a bandwidth selected using cross-validation. Nonetheless, we contend that the RSA criterion provides an intuitive criterion by which bandwidths can be selected and, through the elimination of spatial autocorrelation, permits statistical inference and testing to proceed using standard econometric tools whilst imposing little assumed structure on the model of the hedonic price function.

6. TESTING FOR OMITTED SPATIAL COVARIATES

The analysis of the previous sections has determined that the regression residuals from LS estimation exhibit patterns of spatial autocorrelation. The final comparison we wish to make is between the SSE (selecting bandwidth using the RSA criterion) and the SED models. Under the assumptions of the SED model, LS returns unbiased parameters

Our null hypothesis is that the spatial autocorrelation can adequately be described as a feature of the error generating process as is assumed by the SED model. The alternative hypothesis is that spatial autocorrelation in the residuals indicates locational covariates whose omission from the model is the source of omitted variable bias. If the null were true then we would not expect any substantive differences between the parameters from an LS estimator and the SSE estimator. Alternatively, if the SSE model captures the influence of substantive features of the spatial environment omitted from the regressor data, then we may expect to witness statistically significant differences between the parameters of the two models.[8]

First observe the unadjusted R^2 statistics for the two models reported in Table 5.[9]

Table 5: Unadjusted R^2 statistics for the LS and SSE models

Cluster	R^2 LS	R^2 SSE
Cluster 1	0.721	0.760
Cluster 2	0.830	0.864
Cluster 3	0.800	0.827
Cluster 4	0.807	0.838
Cluster 5	0.790	0.834
Cluster 6	0.847	0.868
Cluster 7	0.829	0.853

In all cases there is a considerable increase in explained variation with the SSE estimator when compared to the OLS estimator. On average the unadjusted R^2 statistic increases by 3.1%, ranging from a low of 2.1% in cluster 6 to a maximum of 4.4% in cluster 5. However, as Anglin and Gencay (1996) observe, it would be more appropriate to compare the adjusted R^2 statistics for the two models, but this comparison is impossible to make as the effective degrees of freedom of the semiparametric SSE model is not known. Consequently, we perform a number of statistical tests to compare the two estimators.

Our testing strategy is to compare the SSE model to the LS model under the null hypothesis that the LS model is correctly specified though there may remain spatial autocorrelation in the nuisance process.

Following, Robinson (1988) and Anglin and Gencay (1996) we first apply the Hausman test.

---

[8] Appendix B provides full listings of parameter estimates for the LS estimator whilst full listings for the SSE estimator can be found in Appendix C.

[9] Following Anglin and Gençay (1996) we calculate the R^2 statistic as $R^2 = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}/\boldsymbol{y}'\boldsymbol{y}$ where each element of $\hat{\boldsymbol{y}}$ is given by $\hat{y}_i = \hat{E}[y \mid \boldsymbol{c}_i] + \left(\boldsymbol{x}_i - \hat{E}[\boldsymbol{x} \mid \boldsymbol{c}_i]\right)\hat{\boldsymbol{\beta}}$ (where a circumflex denotes an estimated quantity).

$$\tau_H = \left(\hat{\beta}^{SSE} - \hat{\beta}^{LS}\right) Var\left(\hat{\beta}^{SSE} - \hat{\beta}^{LS}\right)^{-1} \left(\hat{\beta}^{SSE} - \hat{\beta}^{LS}\right) \quad (14)$$

Once again, we are unable to use Hausman's expression for the variance of the difference in the two parameter vectors since we do not have an easy way of computing the asymptotic variance matrix of the LS estimator when the residuals are spatially correlated. Instead we employ a bootstrap procedure as outlined in Section 5. The test statistic, τ_H is asymptotically distributed chi-squared with degrees of freedom equal to the number of parameters estimated by both models.

A second test is that proposed by Whang and Andrews (1993). Their test is based on the vector of sample moments;

$$\boldsymbol{r} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - E[y \mid \boldsymbol{c}_i] - (\boldsymbol{x}_i - E[\boldsymbol{x} \mid \boldsymbol{c}_i])'\hat{\beta}\right)(\boldsymbol{x}_i - E[\boldsymbol{x} \mid \boldsymbol{c}_i]) \quad (15)$$

Clearly, if $\hat{\beta}$ in (11) is replaced by $\hat{\beta}^{SSE}$ then ***r*** will be a vector of zeros since (15) is simply the set of normal equations for the SSE estimator. Of course, under the null $\hat{\beta}^{LS}$ should be approximately equal to $\hat{\beta}^{SSE}$. As such the Whang and Andrews test requires that $\hat{\beta}$ in (15) be replaced by $\hat{\beta}^{LS}$. If the null holds then the moments in (15) should still approximate a vector of zeros. The test statistic is given by;

$$\tau_{WA} = \boldsymbol{r}'\hat{\Phi}^{-1}\boldsymbol{r} \quad (16)$$

where $\hat{\Phi}$ is a consistent estimator of the variance matrix of ***r*** under the null. Whang and Andrews (1993) show that τ_{WA} is asymptotically distributed chi-squared with degrees of freedom equal to the number of parameters estimated by the SSE model. Furthermore, Whang and Andrews (1993) give formulas for $\hat{\Phi}$ when the residuals are correlated. Here we prefer to bootstrap $\hat{\Phi}$ by resampling with replacement from the original data 1,000 times, re-estimating the LS and SSE models and calculating ***r*** for each bootstrap sample. Our bootstrap estimate of $\hat{\Phi}$ is the empirical variance matrix of the 1,000 bootstrap estimates of ***r***.

As discussed in Section 5, the Hausman and Whang and Andrews test statistics are based on a subset of the regression parameters. Again we do not include the parameters for locational constants in the tests, nor can we include parameters that are not estimated for every bootstrap sample.

The final test applied here is that of Li and Wang (1998). Similar to the Whang and Andrews test, the Li and Wang test statistic is based upon the residual from a "mixed" regression;

$$u_i = y_i - \hat{\beta}_0^{LS} - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^{SSE} \quad (17)$$

where $\hat{\beta}_0^{LS}$ is the estimated constant from the LS regression. Their test statistic is based upon a standardised kernel estimator of the moment condition $E\left[u_i \, E\left[u_i \mid \boldsymbol{c}_i \right] f\left(\boldsymbol{c}_i \right) \right]$, where $f\left(\boldsymbol{c}_i \right)$ is the spatial density of the observations.

The test statistic is given by;

$$\tau_{LW} = \frac{\sum_{i=1}^{N} \sum_{j=1,\, j \neq i}^{N} u_i \, u_j \, K_b\left(\boldsymbol{c}_i - \boldsymbol{c}_j \right)}{\left(\sum_{i=1}^{N} \sum_{j=1,\, j \neq i}^{N} 2 u_i^2 \, u_j^2 \, K_b^2\left(\boldsymbol{c}_i - \boldsymbol{c}_j \right) \right)^{1/2}} \tag{18}$$

where $K_b\left(\boldsymbol{c}_i - \boldsymbol{c}_j \right)$ is a kernel function. Li and Wang (1998) show that under the null the test statistic is distributed $N(0,1)$.

The results of these tests are presented in Table 6. All three tests support the same conclusion. In clusters 4, 6 and 7 we cannot reject the hypothesis that the LS model is correctly specified when compared to an SSE alternative. In these cases, one is safe to assume that the SED model provides be an effective model of the spatial processes in operation. This result is supported by the evidence of the correlograms in Figure 4 which show that evidence for spatial autocorrelation was weakest in these three clusters.

Table 6: Tests comparing LS null against SSE alternative

Cluster	Hausman		Whang & Andrews		Li & Wang	
	Stat (df)	*p*-value	Stat (df)	*p*-value	Stat	*p*-value
Cluster 1	85.47 (52)	0.002	112.74 (51)	0.000	4.984	0.000
Cluster 2	91.77 (48)	0.000	136.80 (48)	0.000	4.177	0.000
Cluster 3	91.32 (52)	0.001	95.81 (52)	0.000	1.476	0.070
Cluster 4	34.37 (46)	0.896	51.00 (44)	0.284	-1.463	0.928
Cluster 5	98.41 (50)	0.000	143.62 (50)	0.000	5.505	0.000
Cluster 6	37.52 (48)	0.862	58.46 (48)	0.143	-1.418	0.922
Cluster 7	43.42 (47)	0.621	58.03 (47)	0.130	0.252	0.401

In contrast, for clusters 1, 2, 3 and 5 we can unequivocally reject the null. In these cases all three tests support the conclusion that the LS model is incorrectly specified and that the SSE model returns significantly different parameter estimates. In short, these clusters show strong evidence of the presence of omitted spatial covariates. Applying a SED estimator to these data would provide biased estimates of the model parameters.

7. CONCLUSIONS

Spatial autocorrelation of regression residuals is a common feature of many econometric models. For example, in this paper we find strong evidence for spatial autocorrelation in the regression residuals from a hedonic model that examines differences in property prices in the City of Birmingham in the United Kingdom.

To gain a more thorough appreciation of the nature of spatial dependence in regression residuals we propose construction of a spatial correlogram plotting Moran's *I* statistic for residuals at progressively larger separation intervals. Since the distribution of *I* under a null hypothesis of no spatial autocorrelation is known, it is possible to establish statistically the separation interval at which correlation of the residuals is no longer a feature of the data. In this case, we find that this separation interval differs between various subsets of the data, ranging from 100m to 300m.

Over recent years, a substantial literature has arisen concerning itself with how best to estimate econometric models blighted by spatial autocorrelation. In general, the preferred approach has been to assume that spatial correlation in regression residuals is the consequence of some modelable process generating the nuisances. Maximum likelihood and general method of moments estimators have been proposed for such SED models.

Of course, spatial autocorrelation of the regression residuals is induced by spatial features influencing property prices that are not observed by the researcher. The SED models assume that spatial autocorrelation is a consequence of an amalgam of the many subtle nuances of location and that this amalgam might adequately be regarded as a nuisance process.

However, the possibility exists that the researcher fails to observe substantive spatial features whose absence from the model is likely to induce missing variable bias in the parameter estimates. Fortunately, where the omitted variables are expected to be features of geographical space, a course of action suggests itself. In particular, we employ the SSE estimator of Gibbons and Machin (2003) and Gibbons (2003). The SSE accounts for missing spatial covariates by nonparametrically smoothing the data over a proscribed area.

In our application we spatially smooth the data using local linear regression. This approach offers significant gains over the Nadaraya-Watson smoother, especially at the boundaries of the data and when the data is not equally spaced. Furthermore, we adopt Ellner and Seifu's (2002) RSA criterion in order to select the spatial smoothing bandwidth. The RSA criterion selects that spatial bandwidth which eliminates spatial autocorrelation from the regression residuals. We compare this to the bandwidth selected through the minimisation of the cross-validation statistic, a selection criterion which has been shown to have some asymptotic optimality features (Härdle and Marron, 1985). In most cases we find that we cannot reject the hypothesis that the parameters from the SSE model using bandwidths suggested by the RSA criterion and cross-validation are equal. We contend that the spatial smoothing bandwidth for the SSE model should be selected using the RSA procedure. In particular, this procedure provides an intuitive criterion by which bandwidths can be selected and through the elimination of spatial autocorrelation, permits statistical inference and testing to proceed using standard econometric tools whilst imposing little assumed structure on the model of the hedonic price function.

Finally, we have applied statistical tests to determine whether the parameters of the SSE estimator differ significantly from those of a SED estimator. In cases where the correlogram of the regression residuals indicates that spatial autocorrelation is an important feature of the data, we find that we can clearly reject the hypothesis that the two estimators return the same parameter estimates. In these cases, we have strong evidence for the presence of substantive omitted spatial covariates, such that the application of an SED estimator would provide biased estimates of the model parameters.

References

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

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

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

Anselin, L., and Rey, S., (1991). "Properties of tests for spatial dependence in linear regression models", *Geographical Analysis*, 23, pp 112-131.

Bartels, C. P. A., and Hordijk, L., (1977). "On the power of the generalized Moran contiguity coefficient in testing for spatial autocorrelation among regression disturbances", *Regional Science and Urban Economics*, 7, pp 83-101.

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.

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.

Bowman, A. W., and Azzalini, A., (2003). "Computational aspects of nonparametric smoothing with illustrations from the sm library", *Computational Statistics and Data Analysis*, 42, pp. 545-560.

Brandsma, A. S., and Ketellapper, R. H., (1979). "Further evidence on alternative procedures for testing of spatial autocorrelation among regression residuals", in *Exploratory and Explanatory Statistical Analysis of Spatial Data*, C. P. A. Bartels and R. H. Ketellapper (eds.), Martinius Nijhoff, Boston, MA.

Burridge, P., (1980). "Testing for spatial autocorrelation among regression residuals", *Environment and Planning A*, 13, pp 795-800.

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

Can, A., and Megbolugbe, I., (1997). "Specification and estimation of hedonic housing price models", *Journal of Real Estate Finance and Economics*, 14, pp 203-222.

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

Cliff, A., and Ord, J. K., (1973). *Spatial Autocorrelation*, Pion Limited: London.

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

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

Day, B. H., Bateman, I. J., and Lake, I., (2004). "Nonlinearity in hedonic price equations: An estimation strategy using model-based clustering", *CSERGE Working Paper*, EDM 04-02, University of East Anglia, 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., (1988). "Estimation of regression coefficients in the presence of spatially autocorrelated error terms", *Review of Economics and Statistics*, 70, pp 466-474.

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

Dubin, R. A., (1998). "Predicting house prices using multiple listings data", *Journal of Real Estate Finance and Economics*, 17(1), pp 35-59.

Ellner, S. P., and Seifu, Y., (2002). "Using spatial statistics to select model complexity", *Journal of Computational and Graphical Statistics*, 11(2), pp. 348-369.

Epanechnikov, V., (1969). "Nonparametric estimates of a multivariate probability density", *Theory of Probability and its Applications,* 14, pp 153-158.

Epple, D., and Platt, G., (1998). "Equilibrium and Local Redistribution in an Urban Economy when Households Differ in Preferences and Incomes", *Journal of Urban Economics*, 43(1), pp. 23-51

Epple, D., and Sieg, H., (1999). "Estimating Equilibrium Models of Local Jurisdictions", *Journal of Political Economy*, 99(4). Pp.828-858.

Fan, J., (1992). "Design-adaptive nonparameteric regression", *Journal of the American Statistical Association*, 87, pp. 998-1004.

Fan, J., and Marron, J. S., (1994). "Fast implementations of nonparametric curve estimators", *Journal of Computational and Graphical Statistics*, 3(1), pp. 35-56.

Gawande, K., and Jenkins-Smith, H., (2001). "Nuclear waste transport and residential property values: An hedonic pricing model", *Journal of Environmental Economics and Management*, 42(2), 207-233.

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., (2003). "Paying for good neighbours? Estimating the value of an educated community", *Urban Studies*, 40(1).

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

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

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

Härdle, W., and Marron, J. S., (1985). "Optimal bandwidth selection in nonparametric regression function estimation", *Annals of Statistics*, 13, pp. 1465-1481.

Hastie, T. J., and Loader, C., (1993). "Local regression: Automatic kernel carpentry", *Statistical Science*, 8, pp 120-143.

Hausman, J., (1978). "Specification tests in econometrics". *Econometrica*, 46(3), pp. 1251-1271.

Hepple, L. W., (1998). "Exact testing for spatial autocorrelation among regression residuals", *Environment and Planning, A*, 30, pp 85-108.

Ihlanfeldt, K. R. and Taylor, L. O., (2001). “Externality effects of small-scale hazardous waste sites: evidence from urban commercial property markets” *Environmental Policy Working Paper Series*, #2001-002, Georgia State University, Atlanta.

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

Li, Q., and S. Wang (1998). “A simple consistent bootstrap test for a parametric regression function”, *Journal of Econometrics*, 87, pp 145-165.

Leggett, C. G., and Bockstael, N. E., (2000). “Evidence of the effects of water quality on residential land prices”, *Journal of Environmental Economics and Management*, 39(2), pp. 121-144.

Moran, P. A. P., (1950). “Notes on continuous stochastic phenomena”, *Biometrika*, 37, pp.17-23.

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

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

Pavlov, A. D., (2000). “Space-varying regression coefficients: A semi-parametric approach applied to real estate markets”, *Real Estate Economics*, 28(2), pp 249-283.

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

Wand, M. P., (1994). “Fast implementations of nonparametric curve estimators”, *Journal of Computational and Graphical Statistics*, 3(4), pp. 433-445.

Whang, Y-J, and D. W. K. Andrews (1993). “Tests of specification for parametric and semiparametric models”, *Journal of Econometrics*, 57, pp 277-318.

Appendix A: Parameters of hedonic price regressions excluding locational constants (LS)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Constant	8.4851***	7.8679***	8.1762***	8.3128***	7.6377***	7.9238***	8.2710***
Structural Characteristics:							
Floor Area (log)	0.3804***	0.4254***	0.4623***	0.4199***	0.5455***	0.4195***	0.4221***
Garden Area (log)	0.0828***	0.1681***	0.1043***	0.0911***	0.0894***	0.1498***	0.1318***
Garage	0.0503***	0.0579***	0.0527***	0.0430**	0.0457***	0.0451**	0.0313
Central Heating	0.0239	0.0013	0.0752***	-0.033	0.1386***	0.0909**	-0.0819**
Age	-0.005	0.0001	-0.0127*	-0.008	0.0035	0.0021	-0.0098
WCs							
One	b	b	b	b	b	b	b
Two	0.0221	-0.0397**	0.0454***	-0.0147	0.0334**	-0.0317	-0.0214
Three	0.0386	0.1677	-0.0138	0.0452	0.1056**	-0.0661	0.0139
Four	.	0.6197*	-0.2030*	.	0.4440*	.	.
Five	.	0.1736	.	.	.	.	.
Bedrooms							
One	0.0062	-0.0855	0.033	0.2299*	0.0658	0.1125	0.2077
Two	0.0067	-0.005	0.0056	-0.0297	0.0154	-0.0196	0.0599**
Three	b	b	b	b	b	b	b
Four	0.0319	0.0039	0.0008	0.0219	0.0336	0.0797***	0.0556
Five	0.0077	-0.0303	0.0905*	0.1743**	0.1161**	0.2038***	0.0619

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Six	-0.2325**	-0.0362	-0.0106	0.3477	0.3703***	0.1266	0.034
Seven	0.0924	-0.4692***	0.5032**	-0.0941	-0.4015**	0.163	0.4468*
Eight	.	-0.1253	.	0.4468*	-0.0352	.	0.3906
Nine	.	.	-0.1555	.	.	.	.
Storeys							
One	-0.0273	-0.2675	-0.0095	0.1076*	0.1931***	0.046	0.0904
Two	b	b	b	b	b	b	b
Three	-0.0582	-0.2184***	-0.0880**	-0.055	-0.0975***	-0.0336	0.0179
Four	-0.1976*	-0.9372***	-0.4144***	-0.1468	-0.1842	-0.306	-0.6345***
Five	.	.	.	-0.4005*	-0.29	.	-0.7839**
Construction Type							
Detached Bungalow	0.1623	0.6566*	0.1512***	0.0712	.	.	.
Semi-Detached Bungalow	0.1663	.	.	.	0.0407	-0.0427	-0.1108
End Terrace Bungalow	.	.	.	.	.	.	.
Terrace Bungalow	0.0305	.	0.0082	.	.	.	0.0862
Detached House	0.1594***	0.1717***	0.1227***	0.1479***	0.1089***	0.0536*	0.0682
Semi-Detached House	b	b	b	b	b	b	b
End Terrace House	-0.0891***	-0.1017***	-0.0331*	-0.0436*	-0.0668***	-0.0024	-0.1009***
Terrace House	-0.0720***	-0.0355	-0.0634***	-0.0367	-0.0840***	-0.0642**	-0.1032***
Beacon Group							

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
1. Unrenovated cottage pre 1919	-0.2055	.	.	0.0791	-0.3116	0.3689	.
2. Renovated cottage pre 1919	.	0.0174	0.3770*	.	0.6351***	0.1352	0.3305
3. Small “industrial” pre 1919	-0.1129***	0.0886	0.1503***	0.1013*	-0.0799*	-0.1381**	-0.1991***
4. Medium “industrial” pre 1919	-0.0153	0.0186	0.0098	0.0209	-0.0707**	-0.0562	-0.0763
5. Large terrace pre 1919	-0.0142	0.0804	0.0043	-0.0029	-0.0375	-0.0203	-0.0861
8. Small “villa” pre 1919	-0.0482	0.1066*	-0.0196	0.0595	-0.0452	-0.1389	0.1326*
9. Large “villas” pre 1919	0.0346	0.1315*	0.039	0.1847*	-0.0535	0.2349***	0.1346
10. Large detached pre 1919	0.3787*	0.018	0.2219***	0.0779	0.2661*	-0.0541	-0.3021
19. Houses 1908 to 1930	0.0965**	0.0218	-0.0121	0.093	0.014	0.1491**	0.0644
20. Subsidy houses 1920s & 30s	-0.0596***	-0.0533	-0.0647***	-0.0899***	-0.0161	-0.0101	-0.0847*
21. Standard houses 1919-45	b	b	b	b	b	b	b
24. Large houses 1919-45	0.2196***	0.0732	0.2029***	0.1632**	0.1537***	0.2395***	0.088
25. Individual houses 1919-45	0.2847	.	0.1857	.	-0.2497	0.0835	.
30. Standard houses 1945-53	-0.0626**	-0.1043**	-0.1269***	-0.0027	-0.0527*	-0.1485***	-0.0686

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
31. Standard houses post 1953	0.0101	0.0857*	-0.0049	-0.0013	0.0461	0.0622	0.0098
32. Large houses post 1953	0.2506***	0.2504***	0.1400***	0.1629**	0.1100**	0.1340**	0.1578*
35. Individual houses post 1945	0.5985***	0.411	-0.1745	0.0926	0.0893	0.0523	.
36. “Town Houses” post 1950	-0.1964***	-0.2505	-0.1535	.	-0.0365	-0.2151**	-0.2501*
Sale Date							
1st Quarter (Jan. to Mar.)	-0.0552***	-0.0351*	-0.0588***	-0.0357*	-0.0358**	-0.0686***	-0.0684***
2nd Quarter (Apr. to June)	-0.0186	-0.0064	-0.019	-0.0545***	-0.0078	0.0396**	-0.0246
3rd Quarter (July to Sept.)	b	b	b	b	b	b	b
4th Quarter (Oct. to Dec.)	-0.0084	0.0049	0.0034	-0.0284	0.0203	0.0399**	-0.0221
Neighbourhood Characteristics							
Poverty Factor	-0.0816***	-0.0445***	-0.0637***	-0.1070***	-0.0186	-0.1373***	-0.0825***
Skills Factor	0.1009***	0.1026***	0.0901***	0.0741***	0.1376***	0.0652***	0.1060***
Age Factor	0.0181**	0.0336***	0.0188**	0.0246**	0.0154*	0.0788***	0.024
Family Factor	-0.0458***	-0.0967***	-0.0349***	-0.0457***	-0.0212**	-0.0565***	-0.0984***
Asian Factor	-0.0216**	0.001	-0.0167	-0.022	-0.0652***	0.0423***	0.0779***
Black Factor	-0.0321***	-0.0245*	-0.0793***	-0.0497**	-0.0797***	0.0039	0.0139
Locational Characteristics							
Proximity to City Centre	0	0	0	0	0	0.0001*	0

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Proximity and Quantity of Shops	0.0280***	-0.0252***	-0.0359***	-0.0014	0.011	0.0166	-0.0133
Proximity and Quality of Primary Schools	0.1402***	0.1734***	0.1188***	0.1412***	0.1274***	0.0826	0.1290**
Walking time to Rail Station	0	0	0.0000**	0	0	0	0
Walking time to a Park	0	0.0000**	0.0000**	0	0	0.0000*	0
Driving time to Airport	0	0.0000***	0	0.0000*	0	0.0000*	0.0000*
Proximity to A-Type Industrial Processes	0	0	0.0000*	0.0000***	0	0.0000**	0.0000**
Proximity to B-Type Industrial Processes	0.0000**	0.0000**	0	0	-0.0001***	0	0
Proximity to Land Fill sites	0.0000**	0.0000*	0.0000***	0.0001***	0.0000***	0	0.0000*
Environmental Characteristics							
Views of Water	0.0058**	-0.0012	-0.0002	-0.0005	-0.0011	-0.0034	0.0005
Views of Parkland	0	-0.0003*	-0.0002	0.0002	0	0.0004	0
Road Traffic Noise	-0.0004	0.0014	-0.0019*	-0.0044***	-0.0038***	-0.0028*	-0.0013
Rail Traffic Noise	-0.0029	-0.0071	-0.0057	-0.0140***	-0.0045	-0.0052	-0.0142**
Aircraft Noise	-0.0596	-0.1672	-0.0018	-0.0662	.	.	-0.0121
K	64	63	63	61	63	60	62
N	2261	1258	2173	895	2018	1207	970
R^2	0.694	0.788	0.783	0.789	0.759	0.819	0.801

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Adj.R^2	0.685	0.777	0.776	0.771	0.751	0.810	0.787
s	0.222	0.2394	0.2207	0.2015	0.227	0.2441	0.2587

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

Appendix B: Parameters of hedonic price regressions including locational constants (LS)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Constant	8.9067***	8.2327***	8.7047***	8.4688***	8.3764***	8.4239***	8.9403***
Structural Characteristics:							
Floor Area (log)	0.3827***	0.3991***	0.4383***	0.3612***	0.4879***	0.3864***	0.3670***
Garden Area (log)	0.0838***	0.1662***	0.0973***	0.1005***	0.0940***	0.1393***	0.1446***
Garage	0.0448***	0.0579***	0.0550***	0.0350*	0.0524***	0.0607***	0.0369
Central Heating	0.0464*	0.0653*	0.0577**	-0.0283	0.1032***	0.0828**	-0.0716*
Age	-0.0148**	-0.0067	-0.0096	-0.0058	-0.0204***	-0.0091	-0.0106
WCs							
One	b	b	b	b	b	b	b
Two	0.0243*	-0.0399**	0.0315**	-0.0059	0.0297**	-0.022	-0.0244
Three	0.0198	0.2056**	-0.0112	-0.022	0.1304***	0.0222	0.0075
Four	.	0.8627***	-0.2295*	.	0.4666**	.	.
Five	.	0.372	.	.	.	.	.
Bedrooms							
One	0.0727	0.0675	0.0414	0.2473**	0.0351	0.3394	0.1957
Two	0.007	-0.0013	0.0127	-0.0299	0.0152	-0.0062	0.0560**
Three	b	b	b	b	b	b	b
Four	0.0278	0.0029	0.0165	0.0279	0.0407*	0.0677**	0.047
Five	0.0452	-0.0609	0.1349***	0.1474**	0.1459***	0.1758***	0.044

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Six	-0.1925*	-0.0346	0.0692	0.4663*	0.3435***	0.1821**	-0.0534
Seven	0.1105	-0.3969***	0.6348***	-0.2241	-0.5027***	0.2726*	0.4141
Eight	.	-0.0483	.	0.4797**	-0.0843	.	0.2517
Nine	.	.	0.0221	.	.	.	.
Storeys							
One	-0.07	-0.4751	-0.037	0.0449	0.1903***	0.2255	0.1281
Two	b	b	b	b	b	b	b
Three	-0.0481	-0.2195***	-0.1069***	-0.0672	-0.1115***	-0.0166	0.0183
Four	-0.2106*	-0.8875***	-0.4576***	-0.1522	-0.1909*	-0.1956	-0.4995**
Five	.	.	.	-0.3947*	-0.3526	.	-0.5758*
Construction Type							
Detached Bungalow	0.2031	0.8569***	0.1771	0.1213	.	-0.1787	.
Semi-Detached Bungalow	0.1699	.	0.0075	.	-0.0383	.	-0.0694
End Terrace Bungalow	-0.0122	.	.	.	.	.	.
Terrace Bungalow	.	.	.	.	.	.	0.0827
Detached House	0.1396***	0.1477***	0.1220***	0.1386***	0.1087***	0.0721***	0.0884**
Semi-Detached House	b	b	b	b	b	b	b
End Terrace House	-0.0887***	-0.0981***	-0.0440**	-0.0493*	-0.0780***	-0.0309	-0.1012***
Terrace House	-0.0795***	-0.0418*	-0.0647***	-0.0407*	-0.0917***	-0.0763***	-0.0833***
Beacon Group							

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
1. Unrenovated cottage pre 1919	-0.1371	.	.	0.0228	-0.3349	-0.0215	.
2. Renovated cottage pre 1919	.	0.1031	0.3737*	.	0.6604***	0.2239	0.2426
3. Small “industrial” pre 1919	-0.1158***	0.0321	0.1031**	0.0414	-0.0325	-0.0665	-0.1759***
4. Medium “industrial” pre 1919	-0.0225	0.0257	-0.0485*	-0.0053	-0.0259	0.0092	-0.0545
5. Large terrace pre 1919	0.0134	0.0614	-0.0662	-0.1176	0.0702	0.029	-0.1293
8. Small “villa” pre 1919	0.0092	0.0927	-0.0518	0.0416	0.0324	-0.0365	0.2134***
9. Large “villas” pre 1919	0.036	0.0955	0.0222	0.1769*	-0.0274	0.1872**	0.1927**
10. Large detached pre 1919	0.4524**	-0.1677	0.2266***	0.0403	0.4721***	-0.1772	-0.4598*
19. Houses 1908 to 1930	0.1012**	0.072	-0.0585	0.0704	0.0748	0.1327**	0.1309*
20. Subsidy houses 1920s & 30s	-0.0805***	-0.0919***	-0.0880***	-0.1278***	-0.0292	0.0157	-0.0545
21. Standard houses 1919-45	b	b	b	b	b	b	b
24. Large houses 1919-45	0.2597***	0.1120**	0.2021***	0.1768**	0.1352***	0.2615***	0.1892**
25. Individual houses 1919-45	0.1885	.	0.0152	.	-0.3234	0.1769	.
30. Standard houses 1945-53	-0.0851***	-0.1605***	-0.1436***	-0.0127	-0.0845***	-0.0851*	-0.0498

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
31. Standard houses post 1953	-0.0491	-0.0169	-0.0204	-0.0261	-0.0543	0.0304	-0.011
32. Large houses post 1953	0.1536***	0.1201*	0.1202***	0.1581**	0.0157	0.1359**	0.1013
35. Individual houses post 1945	0.5353***	0.4387*	-0.214	0.0809	0.0312	0.0838	.
36. “Town Houses” post 1950	-0.2377***	-0.1489	-0.1329	.	-0.2121	-0.2559***	-0.1947
Sale Date							
1st Quarter (Jan. to Mar.)	-0.0508***	-0.0313*	-0.0564***	-0.0400**	-0.0407***	-0.0716***	-0.0675***
2nd Quarter (Apr. to June)	-0.0231*	-0.017	-0.0224*	-0.0495***	-0.009	0.0246	-0.0262
3rd Quarter (July to Sept.)	b	b	b	b	b	b	b
4th Quarter (Oct. to Dec.)	-0.0039	-0.0094	0.0023	-0.0171	0.015	0.0269	-0.0101
Neighbourhood Characteristics							
Poverty Factor	-0.0871***	-0.0736***	-0.0648***	-0.1284***	-0.0471***	-0.1260***	-0.0483**
Skills Factor	0.0628***	0.0305***	0.0524***	0.0401**	0.0662***	0.0055	0.0404**
Age Factor	0.0209***	0.0303**	0.014	0.0264*	0.0098	0.0587***	0.0404**
Family Factor	-0.0075	-0.0489***	-0.0205*	-0.0255	-0.0106	-0.0248	-0.0506***
Asian Factor	0.0137	-0.0482***	-0.0368***	-0.0697**	-0.0118	0.0314**	0.0438**
Black Factor	-0.0254**	0.0046	-0.0517***	-0.0314	-0.0520***	0.0269**	-0.011
Locational Characteristics							
Proximity to City Centre	0.0001**	-0.0001	-0.0001**	-0.0001	0	0.0001*	-0.0001*

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Proximity and Quantity of Shops	0.0230***	-0.0134	-0.0347***	-0.0023	0.0126	0.0271**	-0.0363***
Proximity and Quality of Primary Schools	0.0961**	0.1593***	0.1089***	0.1766***	0.0942**	0.0134	0.0404
Walking time to Rail Station	0	0	0	0.0001***	0.0000**	0	0
Walking time to a Park	0	0	0.0000**	0	0	0	0
Driving time to Airport	-0.0001***	-0.0001	0	0	-0.0001***	-0.0001**	-0.0001
Proximity to A-Type Industrial Processes	0	0.0001***	0	0.0000**	0.0000**	0	0
Proximity to B-Type Industrial Processes	0	-0.0001***	0	-0.0001*	0	-0.0001***	0
Proximity to Land Fill sites	0	0.0000**	0.0000**	0.0001***	0.0000*	0.0000**	0
Wards							
Acock's Green	-0.2896**	-0.0878	-0.2595***	0.0285	-0.1522***	-0.128	0.0541
Aston	.	-0.5673	-0.3934**	.	.	-0.2820**	-0.1593
Bartley Green	-0.1273	.	-0.2058***	-0.1212	-0.1503*	.	.
Billesley	-0.0747	-0.0923	-0.1043**	0.0655	-0.0829*	.	.
Bournville	0.0683	0.2489*	-0.0475	0.078	0.0654	.	.
Brandwood	-0.1095	0.1651	-0.1362**	0.0854	0.0599	0.056	.
Edgbaston	0.0399	0.2047	-0.2538***	0.118	0.1255*	0.5427***	0.3242
Erdington	-0.2274**	0.0281	-0.1146***	0.0413	-0.1196***	-0.0675	.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Fox Hollies	-0.2115*	.	-0.2427***	0.1233*	-0.1544***	-0.1841*	0.1992
Hall Green	-0.1784	0.0166	-0.1187***	0.0472	-0.0839**	0.122	0.1605
Handsworth	.	-0.2885**	-0.2472*	.	-0.2132***	-0.2763**	-0.0937
Harborne	0.0748	0.3169**	0.0817	0.1066	0.3960***	.	.
Hodge Hill	-0.2692**	.	-0.2717***	0.1024	-0.0918**	-0.0024	.
King's Norton	-0.0621	.	-0.0571	-0.0436	-0.0245	.	.
Kingsbury	-0.2660**	.	-0.2635***	0.0237	-0.0245	.	.
Kingstanding	-0.1627	-0.1531	-0.1734***	-0.0853	-0.0694	0.0205	-0.2233
Ladywood	-0.084	0.1058	-0.3156***	0.1636	-0.0301	-0.221	.
Longbridge	-0.1124	0.169	-0.1032	-0.0639	0.0824	.	0.4550***
Moseley	.	0.155	-0.3543***	-0.3153**	-0.0606	0.1903	0.3409**
Nechells	-0.3682***	-0.2002	-0.1422**	0.0916	-0.4287***	-0.7439***	-0.0714
Northfield	-0.0855	.	-0.0881	0.0913	0.1691	.	.
Oscott	-0.2088*	-0.3334**	-0.2240***	-0.2329**	-0.1163**	.	.
Perry Barr	-0.1629	-0.2357*	-0.2433***	-0.1577	-0.1565***	.	.
Quinton	0.0412	0.2236	-0.1776*	-0.0967	0.0335	0.1068	.
Sandwell	-0.1557	-0.3896***	-0.2749***	-0.0042	-0.2002***	-0.0467	-0.1068
Selly Oak	0.0982	0.2363	.	0.2023*	0.1423**	.	.
Shard End	-0.4113***	.	-0.2220***	0.118	-0.1345	-0.4445***	.
Sheldon	-0.2656**	.	-0.2177***	.	0.0045	-0.067	-0.0842

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Small Heath	-0.2866**	0.0406	.	.	-0.2047	-0.1810*	-0.1147
Soho	.	-0.3117**	.	.	.	-0.4080***	-0.2027
Sparkbrook	.	0.0503	.	.	.	-0.1595	-0.0112
Sparkhill	-0.1936	0.0833	-0.058	0.2220*	-0.0819	-0.0868	0.2269
Stockland Green	.	.	-0.2485***	-0.0557	-0.1976***	.	.
Sutton Four Oaks	0.0501	-0.1293	0.0659*	0.1483	0.0454	.	.
Sutton New Hall	b	b	b	b	b	b	b
Sutton Vesey	-0.0638	.	.	0.0936	-0.0472	0.1214	0.2678*
Washwood Heath	-0.3167***	.	-0.3352***	0.0473	-0.1505***	-0.2755**	-0.4276
Weoley	-0.11	0.0734	-0.2655***	0.0939	-0.0187	0.1514	0.1722
Yardley	-0.2692**	-0.0777	-0.2097***	-0.0155	.	-0.1086	0.0141
Environmental Characteristics							
Views of Water	0.0055**	-0.0001	0	0.0029	-0.0009	-0.0008	0.0002
Views of Parkland	0	-0.0002	-0.0002	0	0.0002	0.0003	0
Road Traffic Noise	-0.0004	-0.0002	-0.0024**	-0.0037**	-0.0038***	-0.0035**	-0.0035*
Rail Traffic Noise	-0.0026	-0.0126*	-0.0086**	-0.0089**	-0.0023	-0.0046	-0.0119**
Aircraft Noise	-0.0906*	-0.1413	0.0102	-0.0637	.	.	-0.0109
K	96	90	96	93	97	85	82
N	2261	1258	2173	895	2018	1207	970
R^2	0.721	0.830	0.800	0.807	0.790	0.847	0.829

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Adj.R^2	0.709	0.817	0.791	0.785	0.779	0.836	0.813
s	0.2133	0.2169	0.2135	0.1955	0.2139	0.2267	0.2425

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

Appendix C: Parameters of hedonic price regressions including locational constants (SSE estimator)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Structural Characteristics:							
Floor Area (log)	0.3627***	0.3321***	0.4327***	0.3594***	0.4662***	0.3757***	0.3711***
Garden Area (log)	0.0829***	0.1672***	0.0935***	0.1051***	0.0836***	0.1353***	0.1329***
Garage	0.0402***	0.0752***	0.0450***	0.0152	0.0366***	0.0602***	0.0516**
Central Heating	0.0378	0.0491	0.0540*	-0.0329	0.1259***	0.0734*	-0.0919**
Age	-0.0136*	-0.0065	-0.0094	0.01	-0.0186**	-0.0098	-0.0164
WCs							
One	b	b	b	b	b	b	b
Two	0.0187	-0.0309**	0.0374***	-0.0177	0.0258**	-0.0206	-0.0322
Three	0.0583	0.1399	-0.0199	-0.0446	0.0768*	-0.0203	-0.0577
Four	.	0.9636***	-0.2579**	.	0.192	.	.
Five	.	0.394	.	.	.	.	.
Bedrooms							
One	0.0574	-0.0211	0.0541	0.2716**	0.0733	0.2203	-0.025
Two	0.0094	-0.0128	0.0217	-0.0105	0.0092	-0.0017	0.0533**
Three	b	b	b	b	b	b	b
Four	0.0389*	0.0339	0.0203	0.0278	0.0553**	0.0552**	0.0458
Five	0.0780*	-0.0236	0.0970**	0.1448*	0.1599***	0.1564***	0.0479

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Six	-0.2504**	-0.0125	0.0698	0.1617	0.2793***	0.0808	-0.1179
Seven	0.1054	-0.2475*	0.3495	-0.1789	-0.4030***	0.1444	0.3967
Eight	.	0.0858	.	0.3422	0.0013	.	0.1812
Nine	.	.	-0.1097	.	.	.	.
Storeys							
One	-0.111	0.3490***	-0.0582	0.0566	0.0949	0.0578	0.3675
Two	b	b	b	b	b	b	b
Three	-0.0353	-0.2155***	-0.1097***	-0.0656	-0.0776**	-0.0206	0.0047
Four	-0.165	-0.8355***	-0.3608***	-0.0745	-0.1747	-0.1914	-0.5552***
Five	.	.	.	-0.4045**	-0.234	.	-0.6024*
Construction Type							
Detached Bungalow	0.1694	.	0.2022***	0.0677	0.099	.	-0.2827
Semi-Detached Bungalow	0.1563	-0.5017	.	.	.	0.126	-0.2649
End Terrace Bungalow	-0.0094	.	.	.	.	.	.
Terrace Bungalow	.	.	0.1702	.	.	.	.
Detached House	0.1367***	0.1173***	0.1223***	0.1149***	0.1130***	0.0698***	0.1110***
Semi-Detached House	b	b	b	b	b	b	b
End Terrace House	-0.1051***	-0.0736***	-0.0495***	-0.0524**	-0.0777***	-0.0136	-0.1303***
Terrace House	-0.0944***	-0.0379*	-0.0567***	-0.0585**	-0.1012***	-0.0589**	-0.1154***

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Beacon Group							
1. Unrenovated cottage pre 1919	-0.082	.	.	0.0164	-0.3308	-0.2744	.
2. Renovated cottage pre 1919	.	0.19	0.4129**	.	0.6484***	0.164	0.15
3. Small “industrial” pre 1919	-0.1401***	0.021	0.0944**	-0.0055	-0.0289	-0.1207*	-0.1614**
4. Medium “industrial” pre 1919	-0.0422	0.0117	-0.0502*	-0.0531	-0.0572*	-0.0184	-0.0265
5. Large terrace pre 1919	-0.0189	0.0599	-0.0677	-0.1966	0.0968	0.053	-0.1649
8. Small “villa” pre 1919	0.0048	0.0389	-0.0408	-0.0003	-0.0311	-0.098	0.2514***
9. Large “villas” pre 1919	0.0611	0.0915	0.045	0.1042	0.0016	0.2684***	0.2272**
10. Large detached pre 1919	0.3673*	-0.1821	0.1573*	0.025	0.0887	0.2041	-0.3399
19. Houses 1908 to 1930	0.1376***	0.0639	-0.0587	0.0744	0.0956*	0.1181*	0.2039**
20. Subsidy houses 1920s & 30s	-0.0714***	-0.0495	-0.0538**	-0.1217***	-0.0232	0.0242	-0.0165
21. Standard houses 1919-45	b	b	b	b	b	b	b
24. Large houses 1919-45	0.2578***	0.1332***	0.2098***	0.2744***	0.1103**	0.2848***	0.1516*
25. Individual houses 1919-45	0.2856	.	0.3533	.	-0.0949	0.2182	.
30. Standard houses 1945-	-0.1016***	-0.1186**	-0.1212***	-0.0009	-0.0923***	-0.0569	-0.0764

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
53							
31. Standard houses post 1953	-0.0262	0.0189	-0.0067	0.0188	-0.0661*	0.0036	-0.0748
32. Large houses post 1953	0.1095**	0.1046*	0.1254***	0.1948**	-0.0057	0.0992	0.0498
35. Individual houses post 1945	0.4179***	0.273	-0.4998**	0.1312	0.0148	0.1124	.
36. “Town Houses” post 1950	-0.1819**	-0.0425	-0.0281	.	-0.1648	-0.2992***	-0.2767**
Sale Date							
1st Quarter (Jan. to Mar.)	-0.0534***	-0.0382**	-0.0569***	-0.0414**	-0.0482***	-0.0703***	-0.0705***
2nd Quarter (Apr. to June)	-0.0280**	-0.0226	-0.0209*	-0.0422**	-0.0205*	0.025	-0.022
3rd Quarter (July to Sept.)	b	b	b	b	b	b	b
4th Quarter (Oct. to Dec.)	-0.0006	-0.0083	-0.005	-0.0189	0.0167	0.0279	-0.0162
Neighbourhood Characteristics							
Poverty Factor	-0.0881***	-0.0430**	-0.0646***	-0.1480***	-0.0396**	-0.1096***	-0.0323
Skills Factor	0.0210*	0.0118	0.0230*	0.0301	0.0305**	0.0131	0.0105
Age Factor	0.0133	0.0353**	0.0086	-0.0199	0.0185	0.0537**	0.012
Family Factor	0.0083	-0.0560***	-0.0206	-0.0412	0.0076	-0.0365*	-0.0505**
Asian Factor	-0.0031	-0.0377*	-0.0028	0.0027	-0.0516	0.0272	0.0236
Black Factor	-0.0032	-0.0132	0.0006	0.0179	-0.0450*	0.0157	-0.0025
Locational Characteristics							

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Proximity to City Centre	0.0002**	0	0	-0.0002	0	0.0001	0
Proximity and Quantity of Shops	0.0184	-0.0006	-0.0287**	-0.0237	0.0101	0.0139	-0.0073
Proximity and Quality of Primary Schools	0.1566***	0.0999	0.1460***	0.1481*	0.1498***	0.018	0.1217
Walking time to Rail Station	0	0	0	0	0	-0.0001	0
Walking time to a Park	0	0	0	0	0.0001*	0	0
Driving time to Airport	-0.0002**	-0.0001	-0.0001	0.0002	0	0	-0.0001
Proximity to A-Type Industrial Processes	0	0.0001*	0	0.0001	0	0	0.0002
Proximity to B-Type Industrial Processes	0	0.0001	0	0	0	-0.0001	-0.0001
Proximity to Land Fill sites	0	0.0001	0.0001*	0.0001	0.0001	0	0
Wards							
Acock's Green	1.5582	-1.2035	0.617	-2.6396	-3.2907	2.0884	-30.4681
Aston	.	0.0857	-0.0436	.	.	2.7161	4,647.448
Bartley Green	1.078	.	2.7871	-3.6536	-3.9907	.	.
Billesley	1.432	-1.0255	1.14	-2.67	-3.4591	.	.
Bournville	1.4856	-0.9024	1.1647	-3.0443	-3.3406	.	.
Brandwood	1.3945	-0.9102	1.0962	-2.8444	-3.2232	1.1456	.
Edgbaston	0.2833	-1.3627	1.7834	-1.9179	-3.6108	0.389	-26.2427

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Erdington	1.634	0.6445	0.2054	-0.2892	-0.6302	2.4806	.
Fox Hollies	1.4164	.	1.922	-2.5354	-3.4242	1.9649	-29.8137
Hall Green	1.419	0.02	2.0962	-2.676	-3.335	2.1434	.
Handsworth	.	-0.6367	-0.5662	.	-0.75	2.7422	4,647.558
Harborne	0.9861	-1.0371	2.4572	-3.8086	-3.7257	.	.
Hodge Hill	1.6503	.	0.1127	-2.6101	-1.0371	4.3444	.
King's Norton	1.3624	.	1.1851	-3.1194	-2.7419	.	.
Kingsbury	1.8315	.	0.074	0.2226	-0.578	.	.
Kingstanding	1.3201	0.7698	0.3603	-0.5309	-0.214	3.5735	4,647.727
Ladywood	.	-1.2317	1.7089	-5.5353	-3.5308	2.59	.
Longbridge	1.4468	-0.6193	1.3655	-3.5282	-1.7353	.	-15323.77
Moseley	.	-1.2053	1.4861	-2.836	-3.6227	1.822	-26.4718
Nechells	1.5333	-0.9836	0.0469	-2.9615	-0.8741	2.1878	4,645.471
Northfield	1.4186	.	1.227	-3.6237	-2.2074	.	.
Oscott	1.3238	0.8529	0.6805	-1.0281	-0.0447	.	.
Perry Barr	1.5363	1.1032	0.6431	-1.138	-0.2238	.	.
Quinton	1.2933	-0.9474	2.187	-3.7711	-3.6272	-27.3819	.
Sandwell	2.0835	-0.1035	0.9237	0.0848	-0.2761	3.1684	4,646.502
Selly Oak	1.4658	-0.9677	.	-3.1897	-3.2682	.	.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
Shard End	1.5772	.	0.1567	-2.0472	-17.6935	4.931	.
Sheldon	1.4995	.	0.6312	.	-4.4502	2.0716	-30.7874
Small Heath	1.7	-1.0106	.	.	-0.7566	2.0433	-27.7914
Soho	.	-0.1601	.	.	.	2.6591	4,647.521
Sparkbrook	.	-1.0953	.	.	.	1.7668	-27.3963
Sparkhill	1.7329	-1.0899	2.138	-2.2247	-3.1955	2.149	-26.8326
Stockland Green	.	.	0.3969	0.1092	-0.5616	.	.
Sutton Four Oaks	.	.	0.1087*	0.0299	-0.1252	.	.
Sutton New Hall	b	b	b	b	b	b	b
Sutton Vesey	1.8219	.	.	-0.2182	-0.7973	2.061	4,648.265
Washwood Heath	1.605	.	-0.0751	-2.7345	-0.891	3.4125	-26.2216
Weoley	1.1882	-1.1806	2.255	-3.462	-3.2382	-33.5121	.
Yardley	1.5651	-1.2513	0.6089	-2.6128	.	2.0667	-30.1466
Environmental Characteristics							
Views of Water	0.0044	0.0004	0.0003	0.0021	0.0007	0.0004	0.0002
Views of Parkland	0	-0.0001	-0.0002	0.0001	-0.0001	0.0004	0
Road Traffic Noise	0.0002	0.0015	-0.0033***	-0.0028	-0.0041***	-0.0022	-0.003
Rail Traffic Noise	-0.0039	-0.0125*	-0.0069*	-0.0087**	0.0011	-0.0052	-0.0103*
Aircraft Noise	-0.0094	-0.024	0.0136	-0.1072	.	.	-0.0143

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7
K	93	88	95	92	96	84	79
N	2261	1258	2173	895	2018	1207	970
R^2	0.760	0.864	0.827	0.838	0.834	0.868	0.853
s^2	0.039	0.038	0.039	0.032	0.036	0.044	0.050
b	550	675	600	650	450	625	400
h	200	300	300	100	200	200	100
Moran's I	-0.016	-0.021	-0.015	-0.048	-0.016	-0.025	-0.025
Probability of Moran's I	0.987	0.953	0.932	0.961	0.915	0.956	0.941

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