Lett Spat Resour Sci (2014) 7:85–102
DOI 10.1007/s12076-013-0103-x

ORIGINAL PAPER

Evaluating two model reduction approaches for large scale hedonic models sensitive to omitted variables and multicollinearity

Toke E. Panduro · Bo J. Thorsen

Received: 11 March 2013 / Accepted: 18 September 2013 / Published online: 8 October 2013

Abstract Hedonic models in environmental valuation studies have grown in terms of number of transactions and number of explanatory variables. We focus on the practical challenge of model reduction, when aiming for reliable parsimonious models, sensitive to omitted variable bias and multicollinearity. We evaluate two common model reduction approaches in an empirical case. The first relies on a principal component analysis (PCA) used to construct new orthogonal variables, which are applied in the hedonic model. The second relies on a stepwise model reduction based on the variance inflation index and Akaike's information criteria. Our empirical application focuses on estimating the implicit price of forest proximity in a Danish case area, with a dataset containing 86 relevant variables. We demonstrate that the estimated implicit price for forest proximity, while positive in all models, is clearly sensitive to the choice of approach, as the PCA reduced model produces a parameter estimate double the size of the alternative models. While PCA is an attractive variable reduction approach, it may result in an important loss of information relative to the stepwise reduction information based approach.

Keywords Forest proximity · Spatial autocorrelation · GIS · Principal component analysis

Electronic supplementary material The online version of this article (doi:10.1007/s12076-013-0103-x) contains supplementary material, which is available to authorized users.

T. E. Panduro (✉)
Department of Food and Resource Economics, Faculty of Science, University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg C, Denmark
e-mail: tepp@ifro.ku.dk

B. J. Thorsen
Department of Food and Resource Economics and Center for Macroecology, Evolution and Climate, Faculty of Science, University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg C, Denmark

JEL Classification Q51 · R15 · R21 · R31

1 Introduction

The models in applied hedonic valuation studies of environmental externalities have grown in terms of included transactions and number of explanatory variables. Up to recently, studies have been based on only few a thousand transactions and a limited set of explanatory variables (Dubin and Goodman 1982; Garrod and Willis 1992; Morancho 2003; Anthon et al. 2005), while some more recent publications use several thousand observations and include a considerable amount of explanatory variables (Cavailhès et al. 2009; Mukherjee and Caplan 2011; Kuethe 2012). An extreme case of this trend can be found in the work of Gibbons et al. (2011) with more than one million transactions and 33 explanatory spatial variables. While the present study is no exception from this trend, we limit the analysis to 5,659 transactions, but apply as many as 86 available variables which are relevant to the hedonic model. As typical in environmental valuation hedonic studies, we focus on the implicit price of a specific variable, in this case forest proximity and the purpose of the other 85 variables is to ensure a reliable estimate.

Along with the growth in relevant and available variables comes, the challenge of achieving parsimonious models with reliable estimates while dealing adequately with the issues of omitted variable bias and multicollinearity inherent to spatial hedonic models (LeSage and Pace 2009).

Because of the often strong correlation between different spatial variables describing urban qualities, omitted variable bias is a major concern in hedonic models, when data sets appear incomplete. However, as the set of explanatory variables grow more complete, multicollinearity becomes a challenge to the practical application and reliable estimation of parsimonious hedonic models for environmental valuation. These problems, if not handled adequately, may reduce at least the efficiency with which we can estimate and draw inference on parameters of interest, but may potentially also imply biased estimates (LeSage and Pace 2009).

In this paper, we use an empirical application to demonstrate that model reduction under these circumstances is not trivial, and we evaluate two common approaches in an empirical case. The first approach applies principal component analysis (PCA), which is used to construct a set of new orthogonal variables capturing a large part of the variation in the available 86 explanatory variables. The second approach is based on stepwise regression model reduction, where we automated variable selection using Variance Inflation Indexes (VIF) and Akaike Information Criteria (AIC), thus reducing the number of variables by removing first those that are highly collinear and then those that have little additional explanatory power. We evaluate the effects of these two approaches on the estimated implicit price, comparing parameter estimates and variances across the resulting hedonic models with the corresponding estimates from a full model containing all available variables.

While PCA is only occasionally used for model reduction in the environmental valuation literature (e.g. Lake et al. 1998), it is more common in the real estate literature e.g. (Thériault et al. 2003; Bitter et al. 2007), just like stepwise regression approaches have been applied on several occasions (Dunse and Jones 1998; Kong et al. 2007; Yoo et al. 2012). Our purpose is to highlight the possible differences between the approaches in terms of their effect on e.g. the implicit prices of environmental variables, which is of interest in applied environmental valuation.

We have chosen to exemplify the effect of the applied variable reduction techniques by focusing on forest proximity. The value of forest proximity, being close to forest lands, has been assessed in numerous hedonic studies (Tyrväinen and Miettinen 2000; Anthon et al. 2005; Cho et al. 2008; Poudyal et al. 2009), and like these we find a positive effect on house prices. However, we demonstrate that this estimate is sensitive to the choice of model reduction approaches.

2 Empirical and econometric methods

2.1 Principle Component Analysis

The PCA is a standard dimensional reduction technique (e.g. Rencher 2002; Jolliffe 2002 and Anderson 2003) that attempts to capture as much as possible of the variance of a dataset, while still reducing the number of dimensions in the dataset (Hastie et al. 2009). The components are orthogonal axes projected onto the dataset, so that the projections are positioned near the largest number of observations. The components' scores describe these orthogonal axes and can be interpreted as new variables.

Following standard notations, the PCA finds the direction of the greatest variance of the vector z based on the $K \times K$ variance–covariance matrix $\mathbb{V}[z] = \Sigma$ where K is the number of variables of the vector z, cf. (1) below. The variables of vector z are standardized to have a mean of zero and a standard deviation of one. The PCA finds a set of principal components weights $a_1, \ldots, a_k$ where the linear function $a'z$ refers to the principal component scores.

$$
\begin{aligned}
a_1 &= \arg \max_{a=\|a\|=1} \nu\left[a'z\right] \\
a_k &= \arg \max_{\substack{a\,=\,\|a\|\,=\,1 \\ a \perp a_1, \ldots, a_{k-1}}} \nu\left[a'z\right]
\end{aligned}
\tag{1}
$$

The component that captures the most amount of variance in the data is the first principal component. The second principal component captures the greatest amount of variance in the subspace orthogonal to the first, etc.

2.2 Stepwise reduction

The stepwise reduction technique automatizes variable selection by reducing the number of available explanatory variables based on an initial set of criteria. In this analysis we apply a stepwise technique using both a backward and a forward stepwise algorithm. In the first stepwise application the potential explanatory variable is subject to a backward selection algorithm removing the variable with the highest VIF value in each step until no variable has a VIF value above 5. The VIF value of variable i is obtained using the R_i^2 value of a regression of all the other explanatory variables on variable i.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

The VIF value will change for all the explanatory variables with each step, as the variable with the highest collinearity is removed.

In the second step the remaining variables are subjected to a forward selection algorithm based on the minimization of AIC. In each step the available explanatory variables are evaluated against the AIC measure. The variable, which provides the largest improvement in AIC is included in the model. The algorithm stops when is not possible to reduce the AIC measure further with the remaining variables. The AIC is calculated as follows:

$$AIC = -2 \log L + 2(edf) \tag{3}$$

where L is the likelihood and the *edf* is the effective degrees of freedom. Essentially, AIC provides a relative measure of goodness of fit, which penalizes the effective degrees of freedom in the model.

2.3 The hedonic model

The hedonic method is well documented in numerous paper and text books, e.g. Palmquist (2005) and Bockstael and McConnell (2007). The hedonic price function is an equilibrium function created by sellers and buyers of properties seeking to maximize their own utility. In equilibrium, the sales price of any house is a function of its characteristics. The model is based on the assumption of weak separability, which means that the marginal rate of substitution between any two characteristics is independent of the level of all other characteristics. Thus, the hedonic model can provide an estimate of the implicit price of the marginal change of a house characteristic (Palmquist 1991, 1992).

The hedonic price function is estimated using a semi-log transformation and Spatial Error Models (SEM) (Anselin 1988), as initial analyses revealed spatial autocorrelation. Spatial lag models were also estimated but provided similar results as the SEM. The SEM can be written as follows:

$$\begin{aligned} y &= X_1\beta_1 + f_2\beta_2 + \varepsilon \\ \varepsilon &= \lambda W\varepsilon + u \end{aligned} \tag{4}$$

where y is an $N\times 1$ vector of logged sales prices, X_1 is a matrix of explanatory variables. The forest proximity variable is f_2. The observation error is the vector ε and β_1 and β_2 are parameters to be estimated. In the SEM, ε is assumed to consist of two terms. The first term capture spatial autocorrelation using the autoregressive parameter, λ, and W which is an $N \times N$ spatial weight matrix. The second term is a vector of noise u which follows the standard assumptions i.i.d.

The spatial weight matrix W defines the extent of the spatial neighborhood effect at each location. The spatial autoregressive error term in the SEM can be understood as a correction term for unobserved omitted variables shared by the local neighborhood, but there is no strict definition of a neighborhood in the literature (Anselin 2006). We defined neighbors by triangulated irregular network polygons around each property, and based our choice of weight matrix, W, on a spatial correlogram analysis based on global Moran's I analyses performed on contiguous neighbors going from the 1^{st} to 8^{th} order neighbors. We found a fairly sharp decline in spatial correlation and based W on 1^{st} order neighbors only.

3 Data sources, research area and variable definitions

3.1 Housing market

For our analysis we chose a market region in the northwestern part of Zealand, in which the development of average house prices across municipalities shared a similar—fairly modest—price trend over the period 1992-2001, when compared with the housing markets in surrounding regions (Fig. 1).

The region covers 1,227 km^2 and has a forest cover of 120 km^2 (9.7 %), which is a bit below the national average of 12–13 %. Forests are a mixture of deciduous, coniferous and mixed species forest stands. The largest city in the survey area is Kalundborg. Households living in the region have a mean distance of 85 km to Copenhagen, which, by Danish standards, is quite far to commute considering that there is no highway and no express trains going in or out of the area.

3.2 Data sources

In Denmark, nationwide data on structural house characteristics are collected and registered in the "Bygnings - og Boligregisteret" (BBR), and sales prices are collected and registered in "Ejendomsstamregisteret" (ESR). "Krydsreferenceregisteret" (KRR) is able to supply ESR and BBR with a common key, which enables these data to be combined. KRR furthermore contains geographic coordinates for every house in Denmark (Hansen 2000).

We constructed location-based variables using ArcGIS 9.2, using data provided by The Danish Geodata Agency (2011) in the kort10 geo-database, by Miljøundersøgelser (2000) in the "Area Information System" (AIS) and by Naturgas Midt-Nord (2000) in the Danish Address and Road Database (DAV). The location-based variables are calculated using Euclidian distance or road network distance. Several different variables representing forest proximity were constructed and evaluated. All performed quite similarly, but for the purpose of this study, we define forest proximity variable simply as the Euclidian distance in steps of 100 m to the nearest forest. The scale of proximity is calculated by $X_{prox} = c_{cutoff} - X_{dist}$ where X_{dist} is Euclidian distance. Furthermore, for homes beyond the cut-off distance the measure of proximity is set to zero, $X_{prox}|X_{prox} < 0 = 0$. The proximity variable is easy to interpret as amenities are associated with positive coefficients. The cutoff value reflects that the service is declining with distance, and beyond some point effectively zero. The cutoff value was initially chosen by mapping out the relationship between the sales price and buffer

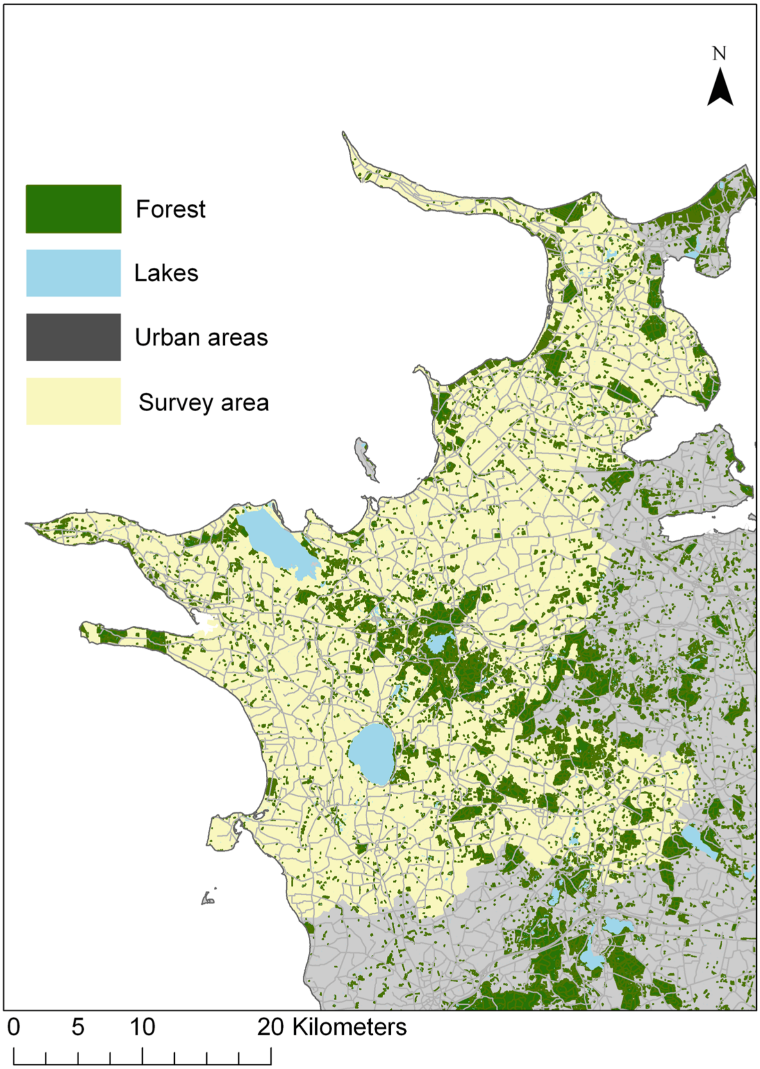

Fig. 1 Land-use map of survey area

distance variables of forest accessibility. We found that the effect of forest proximity was negligible after 600 m.

Data on sales prices for single family houses from 1992 to 2004 are used. To subtract time variation, dummy variables are constructed for each sales year—2004 being the reference year. The data contain 86 explanatory variables that describe structural, neighborhood and environmental variables. After removing 274 incomplete or erroneous observations (missing or implausible technical entries), the remaining 5659 observations formed the basis of our analyses. A thorough description of each variable and descriptive statistics can be found in Supplementary Material (SM).

4 Results

4.1 Model reductions

The correlation matrix of the 86 available variables provided evidence of multicollinearity. We undertook a PCA and a stepwise reduction in order to reduce the problem of multicollinearity while at the same time keeping omitted variable bias to a minimum. Note that 21 of the 86 explanatory variables feed directly into the hedonic models, thus bypassing the model reduction applications. This group of variables covered transaction year dummies and a set of spatial environmental variables. The time dummies are kept in order to ensure the same de-trending across models and the environmental variables are the main focal point of the analysis.

The PCA is calculated using a varimax rotation on the data to create latent variables that describe the underlying structure of the data. The PCA reduced 63 correlated structural and spatial variables to 14 components. Initially, the PCA indicated the presence of 22 components with an eigenvalue above 1, accounting for 70.4 % of the variance in the data. The screeplot of the relationship between the principal components and the eigenvalues is examined in an adjustment step to determine the number of components to extract, based on their combined interpretability. To promote the interpretation of the components, a varimax rotation ensured that the explanatory variables loaded highly on one component and near zero on other components (Hastie et al. 2009). This resulted in the extraction of 14 components accounting for 58.8 % of the total variance of the variables included in the PCA. This is a substantial loss of information, and should be borne in mind in the remaining analysis.

The 14 components (cf. Table 1) represent aspects that are in general intuitively linked. Some examples: Proximity to services and businesses is associated with village and city centers. Institutions like schools, recreational facilities, day care for children are often situated close to each other in Danish urban planning, e.g. to reduce children's need to travel in traffic. District heating, natural gas and similar underground infrastructures are buried under main roads. Solitary farm houses rarely have public sewage but instead forms of mechanical treatment. The older the house, the larger the likelihood that walls are half-timbered and roofs are thatch.

As explained earlier, the stepwise model reduction is conducted in two steps. In the first step the full set of explanatory variables is subjected to a backward selection using a VIF value larger than 5 as a threshold. In total 14 variables are removed in this step. In the second step the remaining variables are subjected to a forward selection, based on the AIC criteria. An additional 19 variables are removed from the model.

4.2 The hedonic house price model

In Table 2 we present the estimates of the forest proximity parameter and model diagnostics for the three versions of the hedonic model. Parameter estimates of the other explanatory variables in the three models can be found in Appendix. The hedonic models include a model using the full set of available explanatory variables, a model which applies the 14 components of the PCA as explanatory variables and a model

Table 1 Selected principal components and their loadings

	Components	Variables	Eigenvalues	Explained % variance	Loadings >0–45
1	Accessibility/ substitutability–infrastructure/retail	Retail	7.6687	11.6192	0.9799
		Supply of retail			0.9765
		Copenhagen city center			0.9248
		Highway exit			0.8860
		Harbor			−0.8630
		Supply of services			0.7543
		Supply of cinemas and theatres			0.6524
		Hospital			−0.5794
		Station			0.5600
2	Substitutability—Public institution	Supply of sports facilities	4.0438	6.1270	0.9654
		Supply of cultural institutions			0.9422
		Supply of healthcare centers			0.7627
		Public cultural institutions			0.6701
3	Accessibility—Service Institutions	Day nursery	3.9082	5.9216	0.7866
		healthcare center			0.7781
		School			0.7435
		Sport facility			0.6966
		Cinemas and theaters			0.6614
		Service store			0.5825
4	The size of the house	Living space	2.7685	4.1947	0.8467
		Toilets			0.7404
		Bathrooms			
		Rooms			0.7354
		Bathrooms			0.6700
5	Farm houses	Public sewage	2.7515	4.1689	−0.8293
		Mechanical treatment			0.7818
		Property size			0.4919
6	Heating with electricity	Electric heating	2.7144	4.1128	0.9245
		Electric stove			0.9198
		Central heating			−0.6427
		Heated by oil			−0.5793
7	Private water supply	Private water supply	2.1566	3.2676	0.8537
		Public water supply			−0.8535
8	Energy & road access	District heating	2.1205	3.2129	−0.6682
		Major road			0.6644
		Natural gas			0.4885
9	Tile roof	Asbestos roof	2.0696	3.1357	−0.8629
		Tile roof			0.7973

Table 1 continued

	Components	Variables	Eigenvalues	Explained % variance	Loadings >0–45
10	Small buildings	Small buildings	1.9386	2.9372	0.8315
		Size of small buildings			0.8090
11	Brick construction	Brick	1.8970	2.8742	−0.8616
		Concrete			0.7108
		Timber			0.5060
12	Age of the house	Half-timbered	1.7362	2.6307	0.7549
		Thatched roof			0.6801
		Age			0.5037
13	Heating—stove and	Heated by coal	1.7044	2.5825	0.8141
	coal	Stove			0.7946
14	Carport and basement	Car port	1.3694	2.0749	0.5919
		Basement			0.4993
		Outhouse			−0.4532
Total variance explained				58.8 %	

See text for intuitive explanation of the grouping
Variables unaccounted for (less than 0.45 loading): covered terrace, garage, patio, top story, waste water tank, electric stove complimentary, buildings, floors, wood—complimentary heating, low basement corrugated iron roof, felt roof, flat roof, private sewage, concrete roof

Table 2 Comparing the hedonic model estimates of the forest proximity parameter

	GLM full model	PCA model reduction	Stepwise model reduction
Forest proximity variable	0.00609 (0.00287)*	0.01164 (0.00277)***	0.00571 (0.00260)*
Lambda		0.08492 (0.02173)***	0.05273 (0.02171)*
R-squared	0.56237	0.50510	0.56083
AIC	3926.564	4583.81	3916.64
Correct signs %	0.72	0.81	0.80
Likelihood Ratio	1875.28	−2253.90	−1902.32
Moran's I	0.01899*	−0.00024	−0.00009
df	5572	5622	5604

N = 5659: () standard error
* significant at 5 %, ** significant at 1%, *** significant at 0.1%

which use the selected variables from the stepwise reduction as explanatory variables. Note that all three models contain transaction year dummies and have a set of selected environmental variables in common. Furthermore, standard errors and significance levels for all hedonic models are based on heteroscedasticity and autocorrelation consistent covariance matrices. The model containing the full set of available variables is estimated by a Generalized Linear Model (GLM). SEM is sensitive to multicollinearity due to issues of singularity. It was therefore not possible to estimate the hedonic model with the full set of available variables using a SEM.

The full GLM model explains 56 % of the variance according to the R^2 using 86 variables, which is only marginally higher than the R^2 of the model based on the stepwise reduction which uses 54 variables. The model with principal components variables has an R^2 around 50 %, but uses only 36 variables. The model based on stepwise reduction had the lowest AIC value, while the PCA based model notably has a much higher AIC value. The two models based on PCA and stepwise reduction have a relatively high number of significant parameter estimates with the expected sign compared with the full model. Note, that the global Moran's I index indicates that spatial autocorrelation is low for all three models. This is likely a result of a having a lot of spatial variables in the models, suggesting that little is left out of the full model.The global Moran's Index is significantly different from zero in the full model, while it is insignificant in both SEM applications.

The stepwise and the PCA based model reduction approaches effectively reduce the multicollinearity problems in the models. However, we find that while the standard error of the forest parameter is 2.87×10^{-3} in the full model, it is only improved marginally to 2.6×10^{-3} in the reduced models, as in this case the correlation between this variable and others is modest. While efficiency gains seem modest, we find a clear difference in the mean estimates of the forest proximity parameter between the PCA-based and the stepwise reduced models. The parameter estimate of forest proximity variables in the PCA models are almost double the size of the corresponding estimate in the full and the stepwise reduced models. This indicates that some of the information lost using the PCA approach may correlate with the forest variable perhaps implying that an omitted variable bias has been introduced. This observation stresses the caution needed when pursuing the estimation of parsimonious models from large data sets.

5 Concluding discussion

In hedonic valuation studies, there is usually a focus on one or a few environmental variables of interest, whereas the rest of the hedonic price function must be designed to obtain the most efficient and unbiased estimates as available information allows. Earlier hedonic studies have often worked on fairly small house price dataset with relatively small spatial extent and a limited number of relevant spatially distributed covariates. However, data availability has grown in recent years and large-scale hedonic models now present both a challenge to and an opportunity for applied environmental valuation. It remains a challenge to achieve parsimonious reliable models and estimates, while dealing adequately with the issues of omitted variable bias and multicollinearity inherent to spatial hedonic models (LeSage and Pace 2009).

In this paper we, evaluate two common model reduction techniques in an empirical application using a very large set of relevant variables, and demonstrate that model reduction under these circumstances is not trivial, and may easily affect the estimate of the environmental valuation parameters of interest, here a forest proximity variable. The first approach applied PCA, to construct a set of new orthogonal variables capturing a large part of the variation in the available 86 explanatory variables. The second approach is based on stepwise regression model reduction, where we automated variable selection using VIF and AIC. Comparing the results of the reduced models with a full model, we find that neither of the model reduction approaches reduce the standard error of the forest proximity estimate much, compared with the inefficient full model. However, the estimate of the forest proximity variable is almost double the size in the PCA-based reduced model compared with the full model and the stepwise reduced model, which are very similar. The finding is likely to be case specific, but it stresses the need for caution when building hedonic models from large scale data sets.

We have focused here on two applied approaches to model reduction in hedonic models used for applied environmental valuation research. The performance of the model reduction techniques could be improved. One option for improving the performance of a PCA-type of approach could be to undertake a simultaneous estimation of the hedonic models and the PCA components, latent house or neighborhood qualities or similar. Such an estimation procedure should at least improve efficiency, but may also reduce the loss of information and hence the risk of omitted variable bias, as this affect the overall likelihood of the model. Another approach could be further development of structural models, which may also handle issues like measurement error due to some variables being poorly observed or proxies (Suparman et al. 2013).

However, while the two-stage PCA approach may not be optimal from an efficiency point of view, it is important to stress that it is used in that way. Similar reservations about e.g. path dependent outcomes exist for the stepwise reduction approach. The point of our paper is exactly to illustrate possible caveats for applied environmental valuation studies in the non-trivial choice between these two currently applied model reduction techniques.

Acknowledgments The authors acknowledge funding from the Municipality of Aarhus, Grontmij and the Danish National Science Foundation.

Appendix

Here we present a table, which provides the parameter estimates of all variables included in the three hedonic house price models, as well as the relevant model diagnostics. The first model is the 'Full model' including all available control variables, the second is the model based on a PCA reduction of the variables and the third model is based on the stepwise reduction approach. The first model is based on a simple GLM estimate while the two later models are based on the spatial error model which correct for spatial autocorrelation in the error term. The estimates of the three models are presented together with relevant model performance tests.

Variables	GLM full model		PCA model reduction		Stepwise model reduction	
	Estimates	t-value	Estimates	z-value	Estimates	z-value
(Intercept) (+)	13.1939 (0.4219)***	31.26972	13.6491 (0.0244)***	559.6299	13.3824 (0.0705)***	189.7223
Component 1 infrastructure retail (+)			−0.0446 (0.0075)***	−5.9499		
Component 2 public institution (+)			−0.0299 (0.0055)***	−5.396		
Component 3 services (+)			−0.0784 (0.0054)***	−14.575		
Component 4 size (+)			0.1935 (0.0056)***	34.2386		
Component 5 farm house (+)			−0.0505 (0.0061)***	−8.2265		
Component 6 electric heating (−)			0.0096 (0.0049)*	1.9675		
Component 7 private water supply (−)			−0.0089 (0.0056)	−1.5865		
Component 8 energy and road (−)			−0.0131 (0.0048)**	−2.7532		
Component 9 tile roof (+)			−0.0472 (0.0043)***	−10.9545		
Component 10 small buildings (+)			0.0127 (0.0053)*	2.3856		
Component 11 brick (−)			−0.0386 (0.0055)***	−7.0553		
Component 12 age (−)			−0.0455 (0.0061)***	−7.4459		
Component 13 coal and stove (+)			−0.0505 (0.0054)***	−9.3332		
Component 14 carport and basement(+)			−0.0421 (0.0076)***	−5.5394		
Living space(+)	0.0039 (2e−04)***	17.64799			0.0039 (2e−04)***	22.0944
Age	−0.0031 (2e−04)***	−13.9864			−0.0031 (2e−04)***	−14.8214
Station (−)	0.0000 (0)***	−5.02412			−0.00002 (0)***	−9.6357

Variables	GLM full model		PCA model reduction		Stepwise model reduction	
	Estimates	t-value	Estimates	z-value	Estimates	z-value
Basement (+)	0.0016 (1e−04)***	11.09494			0.0016 (1e−04)***	12.072
Size of small buildings (+)	0.0010 (2e−04)***	4.39324			0.0010 (2e−04)***	4.2807
Thatched roof (−)	0.2194 (0.0619)***	3.54649			0.1683 (0.0391)***	4.3049
Timber (−)	−0.1112 (0.0548)*	−2.03045			−0.1956 (0.0411)***	−4.7586
Toilets (+)	0.0596 (0.0107)***	5.57694			0.0577 (0.0107)***	5.4057
Stove (+)	0.0079 (0.105)	0.0753			−0.1184 (0.0289)***	−4.1026
Property size (+)	0.00003 (0)***	6.00728			0.00002 (0)***	5.9966
Healthcare center (−)	−0.00001 (0)***	−3.60777			−0.00002 (0)***	−4.9603
Car port (+)	−0.1957 (0.0861)	−2.2719			−0.2079 (0.0835)*	−2.4903
Concrete (−)	−0.0374 (0.0389)	−0.96018			−0.1141 (0.0152)***	−7.5311
Concrete roof (−)	−0.0608 (0.0532)	−1.14347			−0.1094 (0.0252)***	−4.3429
Tile roof (+)	−0.0028 (0.0482)	−0.05886			−0.0462 (0.0124)***	−3.7331
Patio (+)	0.0716 (0.015)***	4.78582			0.0720 (0.0147)***	4.8974
Heated by oil (−)	−0.1107 (0.09)	−1.22979			−0.0487 (0.011)***	−4.4364
Top story (−)	−0.0006 (2e−04)**	−2.59738			−0.0006 (2e−04)**	−2.7503
Roof felt (−)	−0.0424 (0.0598)	−0.70909			−0.0866 (0.0366)*	−2.3664
Mechanical treatment (−)	0.0412 (0.0778)	0.53035			−0.0353 (0.0167)*	−2.1113
Low basement (−)	−0.0452 (0.0199)*	−2.27266			−0.0496 (0.0189)**	−2.6233
Corrugated iron roof (−)	−0.0425 (0.0628)	−0.67753			−0.0828 (0.0407)*	−2.0326
Covered terrace (+)	0.0275 (0.0156)	1.76084			0.0330 (0.0152)*	2.164
Harbor (−)	0.0000 (0)	−0.25012			0.0000 (0)**	2.8433
Service store (−)	−0.0001 (0)**	−2.90066			−0.0001 (0)	−1.5786
Small buildings (+)	0.0180 (0.0088)*	2.04407			0.0210 (0.0086)*	2.4603

Variables	GLM full model		PCA model reduction		Stepwise model reduction	
	Estimates	t-value	Estimates	z-value	Estimates	z-value
Floors (−)	−0.0788 (0.0626)	−1.25798			−0.0779 (0.06)	−1.2977
Heated by coal (+)	−0.1287 (0.096)	−1.34004			−0.0632 (0.0354)	−1.7836
Heated by natural gas (+)	−0.0940 (0.0904)	−1.04052			−0.0256 (0.0153)	−1.6716
Cinema and theatre (+)	0.000001 (0)	1.04746			0.00001 (0)	1.3762
Garage (+)	0.0406 (0.0293)	1.38907			0.0397 (0.0286)	1.3901
Outhouse (+)	−0.0373 (0.0288)	−1.29452			−0.0344 (0.0278)	−1.2388
1992 (−)	−0.7868 (0.0262)***	−30.0832	−0.7742 (0.0268)***	−28.9009	−0.7864 (0.0256)***	−30.668
1993 (−)	−0.8076 (0.0262)***	−30.8132	−0.7947 (0.0269)***	−29.509	−0.8054 (0.0257)***	−31.3142
1994 (−)	−0.7501 (0.0258)***	−29.0748	−0.7363 (0.0263)***	−28.0054	−0.7509 (0.0254)***	−29.6211
1995 (−)	−0.7225 (0.025)***	−28.8749	−0.7059 (0.0257)***	−27.4975	−0.7217 (0.0246)***	−29.2756
1996 (−)	−0.6317 (0.0274)***	−23.0224	−0.6165 (0.0281)***	−21.9282	−0.6330 (0.0269)***	−23.5576
1997	−0.5056 (0.0276)***	−18.3383	−0.4998 (0.0281)***	−17.7757	−0.5060 (0.027)***	−18.7231
1998 (−)	−0.4369 (0.0278)***	−15.694	−0.4189 (0.0283)***	−14.7994	−0.4378 (0.0274)***	−15.9665
1999 (−)	−0.3396 (0.0278)***	−12.2235	−0.3372 (0.0284)***	−11.8603	−0.3389 (0.0273)***	−12.4197
2000 (−)	−0.2628 (0.0289)***	−9.08181	−0.2690 (0.0292)***	−9.2118	−0.2599 (0.0283)***	−9.1896
2001 (−)	−0.1721 (0.0284)***	−6.05146	−0.1641 (0.0292)***	−5.6141	−0.1700 (0.0281)***	−6.0532
2002 (−)	−0.1473 (0.0323)***	−4.56621	−0.1611 (0.0327)***	−4.9234	−0.1522 (0.0318)***	−4.7849
2003 (−)	−0.1011 (0.0334)**	−3.02656	−0.0989 (0.0344)**	−2.8752	−0.1022 (0.0331)**	−3.089
Renovated in 1970s (+)	0.0749 (0.0138)***	5.42159	0.0634 (0.0144)***	4.4018	0.0780 (0.0134)***	5.808
Renovated in 1980s (+)	0.1153 (0.015)***	7.65973	0.1109 (0.0159)***	6.9812	0.1120 (0.0148)***	7.5743
Renovated in 1990s (+)	0.1281 (0.0236)***	5.43987	0.1302 (0.0242)***	5.3811	0.1280 (0.0231)***	5.5423
Railway tracks (−)	−0.0173 (0.0048)***	−3.62761	−0.0185 (0.0048)***	−3.8582	−0.0163 (0.0045)***	−3.612
Large road (−)	−0.0374 (0.0266)	−1.40981	−0.0582 (0.0276)*	−2.1106	−0.0356 (0.0256)	−1.3885

Variables	GLM full model		PCA model reduction		Stepwise model reduction	
	Estimates	t-value	Estimates	z-value	Estimates	z-value
Voltage line (−)	0.0000 (0)	−1.28392	0.0000 (0)	−1.6655	0.0000 (0)*	−2.0353
Coast (+)	−0.0078 (0.0032)*	−2.4401	−0.0029 (0.003)	−0.99	−0.0073 (0.0028)**	−2.6427
Coast^2 (+)	0.0005 (1e−04)***	4.66607	0.0004 (1e−04)***	4.1463	0.0006 (1e−04)***	5.281
Forest (+)	0.0061 (0.0029)*	2.11937	0.0116 (0.0028)***	4.201	0.0057 (0.0026)*	2.1969
Brick (+)	0.0797 (0.0366)*	2.17972				
Half timbered (−)	0.1042 (0.0596)	1.74771				
Asbestos roof (+)	0.0487 (0.0474)	1.02824				
Flat roof (−)	0.0072 (0.0576)	0.12439				
District heating (+)	0.0634 (0.1333)	0.47545				
Central heating (+)	0.1233 (0.0998)	1.23561				
Electric stove (−)	0.1488 (0.075)*	1.98235				
Electric heating (−)	−0.0942 (0.125)	−0.754				
Complimentary heating by wood (+)	0.0155 (0.011)	1.41035				
Complimentary heating by electric stove (−)	0.0752 (0.0679)	1.10824				
Public water supply (+)	0.1154 (0.08)	1.44133				
Private water supply (−)	0.1232 (0.0772)	1.59579				
Public sewage (+)	0.0737 (0.0769)	0.95834				
Private sewage (−)	−0.0069 (0.0901)	−0.07663				
Waste water tank (−)	0.0496 (0.0945)	0.52546				
Buildings (+)	0.0312 (0.1002)	0.311				
Rooms (+)	−0.00001 (0.0052)	−0.00107				
Bathrooms (+)	−0.0139 (0.015)	−0.92889				

Variables	GLM full model		PCA model reduction		Stepwise model reduction	
	Estimates	t-value	Estimates	z-value	Estimates	z-value
Day nursery (−)	0.0000 (0)	0.37373				
School (−)	0.0000 (0)	0.37824				
Sport facility (−)	0.0000 (0)	0.04339				
Supply of sports facilities (−)	0.0000 (0)	1.31423				
Supply of healthcare center (−)	0.0000 (0)	1.70155				
Public cultural institutions (−)	0.0000 (0)	−0.77729				
Supply of cultural institutions (−)	0.0000 (0)	−1.60771				
Supply of cinema and theatre (−)	0.0000 (0)	−1.16905				
Supply of services (−)	0.0000 (0)	0.2263				
Retail (−)	0.0000 (0)	−0.36101				
Supply of retail (−)	0.0000 (0)	−0.22807				
Highway exit (−)	0.0000 (0)	−0.05758				
Major road (−)	0.0000 (0)	0.38755				
Copenhagen city center (−)	0.0000 (0)	−0.62288				
Hospital (−)	0.0000 (0)	1.3645				
Lambda			0.08492 (0.0217)***		0.05273 (0.0217)*	
R-square	0.56905		0.5084		0.56515	
Adjusted R-square	0.56237		0.5051		0.56083	
Number of variables	87		36		54	
Relative number of correct signs	0.72		0.81		0.80	
Akaike info criterion (AIC)	3926.56		4583.8176		3916.6394	
Likelihood ratio	−1875.28		−2253.908		−1902.319	
Global Moran's I	0.01899*		−0.00024		−0.00009	

N=5659: (+)/(−) expected sign, () standard error, * significant at 5 %, ** significant at 1%, *** significant at 0,1%

References

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, New York (2003)
Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer, Dordrecht (1988)
Anselin, L.: Spatial Econometrics. In: Mills, Patterson, A. (eds.) Palgrave Handbook of economics, Econometric theory. vol. 1, pp. 901–969. Palgrave Macmillian, Basingstoke (2006)
Anthon, S., Thorsen, B.J., Helles, F.: Urban-fringe afforestation projects and taxable hedonic values. Urban For Urban Green **3**(2), 79–91 (2005)
Bitter, C., Mulligan, G.F., Dall'erba, S.: Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. J. Geogr. Syst. **9**(1), 7–27 (2007)
Bockstael, N.E., McConnell, K.E.: Environmental and Ressource Valuation with Revealed Preference—A theoretical Guide to Empirical Models. Springer, Berlin (2007)
Cavailhès, J., Brossard, T., Foltête, J.-C., Hilal, M., Joly, D., Tourneux, F.-P., Tritz, C., Wavresky, P.: GIS-based hedonic pricing of landscape. Environ. Resour. Econ. **44**(4), 571–590 (2009)
Cho, S.-H., Poudyal, N.C., Roberts, R.K.: Spatial analysis of the amenity value of green open space. Ecol. Econ. **66**(2–3), 403–416 (2008)
Dubin, R.A., Goodman, A.C.: Valuation of education and crime neighborhood characteristics through hedonic housing prices. Popul. Environ. **5**(3), 166–181 (1982)
Dunse, N., Jones, C.: A hedonic price model of office rents. J. Property Valuat. Invest. **16**(3), 297–312 (1998)
Garrod, G., Willis, K.: The environmental economic impact of woodland: a two stage hedonicprice model of the amenity value of forestry in Britain. Appl. Econ. **24**(7), 715–728 (1992)
Gibbons, S., Mourato, S., Resende, G.: The amenity value of English nature: a hedonic price approach. SERC Discussion Papers, Spatial Economics Research Centre (SERC) (2011)
Hansen, H.S.: Digitale kort og administrative registre. Faglig rapport fra DMU. nr. 330 (2000)
Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning–data mining, inference, and prediction. Springer, Berlin (2009)
Jolliffe, I.T.: Principal Component Ananlysis, 2nd edn. Springer, Heidelberg (2002)
Kong, F., Yin, H., Nakagoshi, N.: Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: a case study in Jinan City. China. Landsc. Urban Plann. **79**(3–4), 240–252 (2007)
Kuethe, T.H.: Spatial fragmentation and the value of residential housing. Land Econ **88**(1), 16–27 (2012)
Lake, I.R., Lovett, A.A., Bateman, I.J., Langford, I.H.: Modelling environmental influences on property prices in an urban environment. Comput. Environ. Urban Syst. **22**, 121–136 (1998)
LeSage, J., Pace, K.R.: Introduction to Spatial Econometrics. Taylor & Francis Group LLC., Stockholm (2009)
Miljøundersøgelser, D. (ed.): Areal, Information Systemet–AIS (2000)
Morancho, A.B.: A hedonic valuation of urban green areas. Landsc. Urban Plann. **66**(1), 35–41 (2003)
Mukherjee, S., Caplan, A.: GIS-based estimation of housing amenities: the case of high grounds and stagnant streams. Lett. Spat. Resour. Sci. **4**(1), 49–61 (2011)
Midt-Nord, N. (ed.): Danish Address and Road, Database (2000)
Palmquist, R.: Hedonic methods. In: Braden, J.E., Kolstad, C.D. (eds.) Measuring the Demand for Environmental Quality, pp. 77–120. Elsevier, Amsterdam (1991)
Palmquist, R.B.: Valuing Localized Externalities. J. Urban Econ. **31**, 59–68 (1992)
Palmquist, R.B.: Property Value Models, Chapter 16. In: Karl-Gran, M., Jeffrey, R.V. (eds.) Handbook of Environmental Economics, vol. 2, pp. 763–819. Elsevier, Amsterdam (2005)
Poudyal, N.C., Hodges, D.G., Tonn, B., Cho, S.H.: Valuing diversity and spatial pattern of open space plots in urban neighborhoods. Forest Policy Econ. **11**(3), 194–201 (2009)
Rencher, A.C.: Methods of Multivariate Analysis. Wiley, New York (2002)
Suparman, Y., Folmer, H., Oud, J.H.L.: Hedonic price models with omitted variables and measurement errors: a constrained autoregression-structural equation modeling approach with application to urban Indonesia. J. Geogr. Syst. (2013)
The Danish Geodata Agency: kort10 (2011)
Thériault, M., Des Rosiers, F., Villeneuve, P., Kestens, Y.: Modelling interactions of location with specific value of housing attributes. Property. Manage. **21**(1), 25–62 (2003)

Tyrväinen, L., Miettinen, A.: Property Prices and Urban Forest Amenities. J. Environ. Econ. Manage. **39**(2), 205–223 (2000)

Yoo, S., Im, J., Wagner, J.E.: Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County. NY. Landsc. Urban Plan. **107**(3), 293–306 (2012)