6 Accounting for spatial variation of land prices in hedonic imputation house price indexes: A semi-parametric approach

Submitted for review. Co-author: Jan de Haan.

Abstract: Location is capitalized into the price of the land the structure of a property is built on, and land prices can be expected to vary significantly across space. We account for spatial variation of land prices in hedonic house price models using geospatial data and a semi-parametric method known as mixed geographically weighted regression. To measure the impact on aggregate price change, quality-adjusted (hedonic imputation) house price indexes are constructed for a small city in the Netherlands and compared to price indexes based on more restrictive models, using postcode dummy variables or no location information at all. We find that, although taking spatial variation of land prices into account improves the model performance, the Fisher house price indexes based on the different hedonic models are almost identical. The land and structures price indexes, on the other hand, are sensitive to the treatment of location.

Keywords: Geospatial information, hedonic modeling, land and structure prices, mixed geographically weighted regression, residential property.

JEL: C14, C33, C43, E31, R31.

§ 6.1 Introduction

Housing markets have two distinct features: every house is unique and houses are sold infrequently. This is problematic for the construction of house price indexes because the usual matched-model method, where the prices of goods are tracked over time, breaks down. Hedonic regression methods and repeat sales methods deal with these problems. The uniqueness of properties is mainly due to location. Within a single neighborhood, the value of two properties with similar structures can differ significantly, depending on the exact location.

Repeat sales indexes fully control for location since they track the prices of the 'same' properties over time (in a regression framework). The problem with repeat sales methods is threefold. First, because they only use matched pairs of houses during the sample period, these methods ignore single sales and are therefore inefficient and prone to sample selection bias. Second, standard repeat sales methods do not adjust for quality changes of the individual houses. Third, these methods cannot provide information on the shadow prices of the property characteristics and thus do not allow the estimation of, for example, price indexes of the land the structure sits on. Given these problems with repeat sales methods, we focus on hedonic regression methods.

Traditional hedonic price indexes also have a number of disadvantages. First, data on housing characteristics must be available. Second, location is typically included in hedonic models at some aggregated level, such as postcode areas, rather than at the individual property level, potentially leading to 'location bias', which is a form of omitted variable bias. Third, land is often not included as an independent variable, again potentially giving rise to bias and making it impossible to estimate price indexes for land. Geospatial data, i.e. information on the exact location of the dwellings in terms of geographic coordinates such as longitude and latitude, can help attenuate the latter disadvantages. Our aim is to show how this can be done and how hedonic house price indexes can be constructed accordingly.

A general problem with the estimation of hedonic models for housing is omitted variables bias. Not properly accounting for location can be a major cause of bias and often leads to spatial autocorrelation of the error terms. As mentioned above, the easiest way to deal with the problem is to include dummy variables for postcode areas. Another straightforward approach, which has also been frequently investigated empirically, is to include explanatory variables for all kinds of amenities. While being of interest because it provides information on the shadow prices of the amenities, this method is rather data intensive and, just like the inclusion of dummy variables, cannot fully capture location effects. As a result, some omitted variables bias and spatial autocorrelation will likely remain.

In recent years, more sophisticated methods have been put forward to handle the problem of spatial autocorrelation. Spatial error models attempt to explicitly model the spatial autocorrelation while spatial lag models include the value of neighbor properties in the model. Both methods can be used in a time dummy hedonic framework, where the model is estimated on pooled data for the whole sample period and price indexes are computed from the time dummy coefficients (Hill et al. 2009; Dorsey et al. 2010). Also within the time dummy hedonic framework, Thanos et al. (2016) comprehensively control for both spatial and temporal effects in computing house price index. It is also possible to apply these spatial (and temporal) methods in a hedonic imputation framework (Rambaldi and Rao, 2011; 2013). Another method uses a spatio-temporal filter which eliminates spatial autocorrelation in order to

estimate an index for a dwelling with specific characteristics (Pace et al., 1998; Tu et al., 2004; Sun et al., 2005).

A disadvantage of the above parametric methods is that a spatial weight matrix has to be specified a priori but that its precise structure is unknown. Nonparametric or semi-parametric methods are more suitable to account for spatial dependence. Semi-parametric methods have become increasingly popular. The effect of variables relating to location, for example, can be estimated nonparametrically in 'characteristics space', whereas the effect of variables relating to the structure of the property can be estimated parametrically, as in traditional hedonic models.

In this paper, we assume that location affects the price of land but not the price of structures. That is, we postulate that land prices vary across space whereas the price of structures is 'fixed'. We deal with this type of spatial nonstationarity using a semi-parametric approach known as Mixed Geographically Weighted Regression (MGWR) in which the land prices are estimated by Geographically Weighted Regression (GWR), a nonparametric method proposed by Brunsdon et al. (1996) and Fotheringham et al. (1998b). An additional advantage is that we will be able to plot a continuous surface of land prices.

Apart from the fact that it deals with spatial nonstationarity in a straightforward way, GWR enables us to model the local form of autocorrelation. Moreover, it allows land prices to vary not only across space but also across time by estimating the model for each period separately. The latter is a prerequisite for the construction of hedonic imputation price indexes. In conclusion, (M)GWR is a flexible approach, which can be seen as a generalization of traditional hedonic methods.

We are specifically targeting statistical agencies engaged in the compilation of house price indexes. This has several consequences. The agencies should have access to geocoded data, but this is hardly a problem these days. The methods applied should be relatively easy to explain. Most importantly, the price indexes should be non-revisable. This means that the use of the time dummy method, where previously published index numbers change when the sample period is extended and new data is added, is ruled out. This strengthens the case for constructing hedonic imputation indexes.

Furthermore, our paper tries to fill a gap in the recent *Handbook on Residential Property Price Indices* (Eurostat et al., 2013) in which the use of geospatial data in the estimation of hedonic house price models is not very well covered. The Handbook uses data for detached dwellings sold in the Dutch city of "A" from the first quarter of 2005 to the second quarter of 2008 to illustrate the various methods. We exploit sales data for the city of "A" also but extend the data set in three dimensions. We have data from the first quarter of 1998 to the fourth quarter of 2007, so our data set covers a period of 10 years. Note that we will use annual rather than quarterly data in our empirical work. The range of characteristics for the structures is broader than that in the

Handbook. Finally, we include houses other than detached dwellings.

The paper proceeds as follows. Section 6.2 outlines some basic ideas. Our hedonic model is linear, with non-transformed property price as the dependent variable and size of land and size of structures as explanatory variables. A normalized version, with price per square meter of living space as the dependent variable, is discussed as well. We also address the inclusion of additional characteristics to describe the quality of structures, including age of the structure to adjust for depreciation. Section 6.3 describes how we treat location. As mentioned before, location is capitalized into the price of land, and we would expect land prices to differ at the property level. The GWR and MGWR models and the way in which they are estimated are explained in detail. Section 6.4 shows how we calculate hedonic imputation indexes. Section 6.5 presents empirical evidence for the Dutch city of "A" and discusses the results. Section 6.6 concludes and identifies potential improvements.

§ 6.2 A simplification of the 'builder's model'

§ 6.2.1 Some basic ideas

Our starting point is the 'builder's model' proposed by Diewert et al. (2011; 2015). It is assumed that the value of a property i in period t, p_i^t, can be split into the value v_{iL}^t of the land the structure sits on and the value v_{iS}^t of the structure:

$$p_i^t = v_{iL}^t + v_{iS}^t. \tag{1}$$

The value of land for property i is equal to the plot size in square meters, z_{iL}^t, times the price of land per square meter, α^t, and the value of the structure equals the size of the structure in square meters of living space, z_{iS}^t, times the price of structures per square meter, β^t. After adding an error term u_i^t with zero mean, model (1) becomes

$$p_i^t = \alpha^t z_{iL}^t + \beta^t z_{iS}^t + u_i^t. \tag{2}$$

The (shadow) prices of both land and structures in (2) are the same for all properties, irrespective of their location. In section 6.3 we relax this assumption and allow for spatial variation of, in particular, the price of land. The 'builder's model' takes depreciation of the structures into account, a topic we address in section 6.2.2.

Equation (2) can be estimated on data of a sample S^t of properties sold in period t. This approach, however, suffers from at least three problems. First, the model has no intercept term, which hampers the interpretation of R^2 and the use of standard tests in Ordinary Least Squares (OLS) regression. Second, a high degree of collinearity between land size and structure size can be expected, so that α^t and β^t will be estimated with low precision. Finally, heteroskedasticity is likely to occur since the absolute value of the errors tends to grow with increasing property prices.

Our next step is to divide the left hand side and right hand side of equation (2) by structure size z_{iS}^t, giving

$$p_i^{t*} = \alpha^t r_i^t + \beta^t + \epsilon_i^t, \tag{3}$$

where $p_i^{t*} = p_i^t / z_{iS}^t$ is the *normalized property price*, i.e. the value of the property per square meter of living space, $r_i^t = z_{iL}^t / z_{iS}^t$ denotes the ratio of plot size to structure size, and $\epsilon_i^t = u_i^t / z_{iS}^t$. This resolves the first two problems as the model now has an intercept term and a single explanatory variable.

However, the normalization is unlikely to resolve the issue of unstable parameter estimates. Estimating (3) by OLS regression is equivalent to estimating (2) by Weighted Least Squares (WLS) using weights equal to $1/z_{iS}^t$. That is, dividing by z_{iS}^t adjusts for heteroskedasticity when the error variance in (2) would be proportional to the square of structure size. This kind of error variance seems quite extreme, so this weighting system may not help reduce the heteroskedasticity problem. Also, the ratios r_i^t (as well as the normalized values p_i^{t*}) may exhibit relatively little dispersion.

Some statistical agencies publish changes in normalized rather than unadjusted property prices, often prices per square meter of structures, to adjust for compositional change of the properties sold. We do not recommend this approach because it is changes in unadjusted property prices and price changes most people will be interested in. Yet, given that (3) is a straightforward regression model, including an intercept term, we do favor specification (3) over (2).

§ 6.2.2 Adding structures characteristics

A potential weakness of hedonic modeling for housing is omitted variables, leading to biased (OLS) parameter estimates and predicted prices. Omitted variables in the models (2) and (3) can relate to land or structures. Improving the treatment of land is the topic of section 6.3. In the present section, we discuss the inclusion of additional characteristics for structures. There are two main issues: depreciation and renovation of structures have been ignored so far, and the use of size as the only price-determining feature seems too simplistic.

Following Diewert et al.(2015), we initially assume a straight-line depreciation model. The adjusted value of the structure is $\beta^t \left(1 - \delta^t \alpha_i^t\right) z_{iS}^t$, where δ^t is the depreciation rate and α_i^t is age of the structure. Information on renovations at the level of individual dwellings is unavailable so that $-\delta^t \alpha_i^t$ measures the effect of *net* depreciation, i.e. the combined effect of 'true' depreciation and renovation. Written in linear form, the adjusted structures value is $\beta^t z_{iS}^t - \beta^t \delta^t \alpha_i^t z_{iS}^t$. Adding the second term to the right-hand side of equation (2) yields

$$p_i^t = \alpha^t z_{iL}^t + \beta^t z_{iS}^t - \beta^t \delta^t \alpha_i^t z_{iS}^t + u_i^t. \tag{4}$$

We do not know the exact age of the structures, but we do know the building period in

decades, from which we can calculate approximate age in decades. Thus, age in our data set is an ordinal (categorical) variable. The net depreciation rate is of course ordinal as well. Using multiplicative dummy variables D_{ia}^t that take on the value 1 if in period t property i belongs to age category a ($a = 1, \cdots, A$) and the value 0 otherwise, and after reparameterizing such that $\beta^t z_{iS}^t$ is no longer a separate term, model (4) becomes $p_i^t = \alpha^t z_{iL}^t + \sum_{a=1}^{A} \gamma^t D_{ia}^t z_{iS}^t + u_i^t$. To be able to use standard techniques, we modify this model as follows:

$$p_i^t = \alpha^t z_{iL}^t + \sum_{a=1}^{A} \gamma_a^t D_{ia}^t z_{iS}^t + u_i^t. \tag{5}$$

No restrictions are placed on the parameters γ_a^t, and the new functional form is neither continuous nor smooth. This is somewhat problematic from a theoretical point of view, because it is at odds with the initial straight-line depreciation model. On the other hand, our approach introduces some flexibility. Age of the structures is not only important for modeling depreciation, it can also be seen as an attribute of the dwelling itself in that houses built in a particular decade are more in demand than other houses, perhaps for their architectural style or other age-related attributes.

Diewert et al. (2015) also show how to incorporate the number of rooms. The new value of the structures becomes $\beta^t \left(1 - \delta^t a_i^t\right) \left(1 + \mu^t z_{iR}^t\right) z_{iS}^t$, where μ^t is the parameter for the number of rooms z_{iR}^t. The linear form for this expression is $\beta^t z_{iS}^t + \beta^t \mu^t z_{iR}^t z_{iS}^t - \beta^t \delta^t a_i^t z_{iS}^t - \beta^t \delta^t \mu^t a_i^t z_{iR}^t z_{iS}^t$. Using dummies D_{ir}^t for the number of rooms with the value 1 if in period t the property belongs to category r ($r = 1, \cdots, R$) and the value 0 otherwise, and reparameterizing again, the extended version of (5)becomes

$$p_i^t = \alpha^t z_{iL}^t + \sum_{a=1}^{A} \gamma_a^t D_{ia}^t z_{iS}^t + \sum_{r=1}^{R} \lambda_r^t D_{ir}^t z_{iS}^t + \sum_{a=1}^{A} \sum_{r=1}^{R} \eta_{ar}^t D_{ia}^t D_{ir}^t z_{iS}^t + u_i^t. \tag{6}$$

Next, in order to save degrees of freedom, we ignore the 'second-order' effects due to the interaction terms $D_{ia}^t D_{ir}^t$, yielding

$$p_i^t = \alpha^t z_{iL}^t + \sum_{a=1}^{A} \gamma_a^t D_{ia}^t z_{iS}^t + \sum_{r=1}^{R} \lambda_r^t D_{ir}^t z_{iS}^t + u_i^t = \alpha^t z_{iL}^t + \left[\sum_{a=1}^{A} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R} \lambda_r^t D_{ir}^t \right] z_{iS}^t + u_i^t \tag{7}$$

The second expression shows that the price of structures, i.e. the price per square meter of living space, equals $\gamma_a^t + \lambda_r^t$ for properties in age class a ($a = 1, \cdots, A$) and category r ($r = 1, \cdots, R$) for number of rooms. A high degree of multicollinearity can occur among the various structures components, but we do not worry about this because we are only interested in the combined effect. Multicollinearity between these components and plot size might still be a problem though. Dividing the first expression in (7) by z_{iS}^t gives

$$p_i^{t*} = \theta^t + \alpha^t r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R-1} \lambda_r^t D_{ir}^t + \epsilon_i^t \tag{8}$$

We included an intercept term θ^t and then excluded dummy variables for age class A and category R for the number of rooms to identify the model.

Model (8) is a straightforward estimating equation for the overall property price per square meter of living space. Additional categorical variables for the structures can be incorporated in a similar way as the number of rooms. As a matter of fact, we will use type of house instead of the number of rooms in our empirical work.

§ 6.3 Land and spatial nonstationarity

§ 6.3.1 Location and the price of land

Location is the most important omitted variable in the hedonic models presented so far. In many empirical studies, location is treated as a 'separate characteristic' by including additive locational dummy variables in models for the *overall* property price. This is not the solution we prefer. Location is definitely capitalized into property prices. However, at least within relatively small regions or cities, the price of structures is most likely to be more or less constant across space. It is the price of the land the structure is built on that can vary significantly across different locations, even within a single neighborhood. The question then arises as to how this spatial variation, or spatial nonstationarity as it is sometimes referred to, in the price of land should be modeled.

We could make the simplifying assumption that the price of land varies across postcode areas but is the same within each postcode area k ($k = 1, \cdots, K$) and denoted by α_k^t. This idea is widely used in empirical studies, such as Diewert and Shimizu (2013) who estimated the 'builder's model' for Tokyo. Using *multiplicative* postcode dummy variables D_{ik}, which take on the value of 1 if property i belongs to k and the value 0 otherwise, an improved version of model (7) for the unadjusted property price is

$$p_i^t = \sum_{k=1}^{K} \alpha_k^t D_{ik} z_{iL}^t + \sum_{a=1}^{A} \gamma_a^t D_{ia}^t z_{iS}^t + \sum_{r=1}^{R} \lambda_r^t D_{ir}^t z_{iS}^t + u_i^t, \tag{9}$$

and an improved version of model (8) for the normalized property price is

$$p_i^{t*} = \theta^t + \sum_{k=1}^{K} \alpha_k^t D_{ik} r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R-1} \lambda_r^t D_{ir}^t + \epsilon_i^t \tag{10}$$

The assumption of equal land prices within postcode areas could be too crude, depending of course on the level of detail of the postcode system. Generalized versions of the models (9) and (10) are obtained by assuming that the price of land can differ at the individual property level, i.e. at the micro location. We denote the property-specific

land price by α_i^t, yielding

$$p_i^t = \alpha_i^t z_{iL}^t + \sum_{a=1}^{A} \gamma_a^t D_{ia}^t z_{iS}^t + \sum_{r=1}^{R} \lambda_r^t D_{ir}^t z_{iS}^t + u_i^t \tag{11}$$

and

$$p_i^{t*} = \theta^t + \alpha_i^t r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R-1} \lambda_r^t D_{ir}^t + \epsilon_i^t. \tag{12}$$

Models (11) and (12) obviously cannot be estimated by standard regression techniques. In section 6.3.2 we will discuss a semi-parametric approach that does allow us to estimate these models. Because the method utilizes data on the prices of neighboring properties (in addition to the price of property *i* itself) to estimate α_i^t, it is not necessarily true that the use of models (11) or (12) will lead to aggregate price indexes that are very different from those found by using models (9) or (10).

§ 6.3.2 Mixed Geographically Weighted Regression

One method that deals with spatial nonstationarity of property prices is the 'expansion method' (Casetti, 1972; Jones and Casetti, 1992). The property price, or in our case the price of land, can be viewed as an unknown function of the property's location in terms of latitude x_i and longitude y_i or a similar geographic coordinate system. This function can be approximated using a Taylor-series expansion of some order; typically, second-order approximations are applied. The expansion method makes use of geospatial data but is basically parametric as it calibrates a prespecified parametric model for the trend of land prices across space (Fotheringham et al. 1998a).

The method we will apply, referred to as *Geographically Weighted Regression* (GWR), deals with spatial nonstationarity in a truly nonparametric fashion (Brunsdon et al. 1996; Fotheringham et al. 1998b). Let us remove the structures characteristics from model (11) for a moment and thus consider land as the only independent variable. Using $\alpha_i = \alpha(x_i, y_i)$, the model becomes

$$p_i = \alpha(x_i, y_i) z_{iL} + u_i. \tag{13}$$

Note that we have dropped the superscript *t* for convenience; it should be clear that we estimate all models for each time period separately. Note further that land prices can be estimated for each location in the area under study, not just for the sample observations, enabling us to plot a continuous surface of land prices.

Model (13) can be estimated using a moving kernel window approach, which is essentially a form of WLS regression. In order to obtain an estimate for the price of land $\alpha(x_i, y_i)$ for property *i*, a weighted regression is run where each related observation *j*, i.e. each neighboring property, is given a weight $w_{ij} (i \neq j)$. The weights should follow a monotonic decreasing function of distance d_{ij} between (x_i, y_i) and (x_j, y_j). There is a

range of possible functional forms from which we have chosen the frequently-used *bi-square function*

$$w_{ij} = \begin{cases} \left(1 - d_{ij}^2/h^2\right)^2 & \text{if} \quad d_{ij} < h \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where h denotes the bandwidth. The choice of bandwidth involves a trade-off between bias and variance. A larger bandwidth generates an estimate with larger bias but smaller variance whereas a smaller bandwidth produces an estimate with smaller bias but larger variance. The usual solution is to select the optimal bandwidth by minimizing the *cross-validation* (CV) statistic

$$CV(h) = \sum_{i=1}^{n} \left[p_i - \hat{p}_{\neq i}(h)\right]^2 \tag{15}$$

where $\hat{p}_{\neq i}(h)$ is the predicted price of property i where the observations for i have been omitted from the calibration process.

The nonparametric GWR approach to dealing with spatial nonstationarity of the price of land has to be adjusted for the fact that models (11) and (12) include structures characteristics with spatially fixed parameters. This leads to a specific instance of the semi-parametric Mixed GWR (MGWR) approach discussed by Brunsdon et al. (1999), where some parameters are spatially fixed and the remaining parameters are allowed to vary across space. To outline the estimation procedure, it will be useful to change over to matrix notation. Denoting the number of observations by n, model (11) can be written in matrix form as

$$\mathbf{P} = \mathbf{Z}_L \otimes \boldsymbol{\alpha} + \mathbf{Z}_S\boldsymbol{\beta} + \mathbf{u}, \tag{16}$$

where $\boldsymbol{\alpha} = (\alpha(x_1, y_1), \alpha(x_2, y_2), \cdots, \alpha(x_n, y_n))^T$ is a vector of land prices to be estimated, $\otimes$ is an operator that multiplies each element of $\boldsymbol{\alpha}$ by the corresponding element of $\mathbf{Z}_L$, $\mathbf{Z}_S$ is the matrix of structures characteristics included in model (11), given by

$$\mathbf{Z}_S = \begin{pmatrix} D_{11}z_{1S} & D_{12}z_{1S} & \cdots & D_{1j}z_{1S} \\ D_{21}z_{2S} & D_{22}z_{2S} & \cdots & D_{2j}z_{2S} \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1}z_{nS} & D_{n2}z_{nS} & \cdots & D_{nj}z_{nS} \end{pmatrix},$$

and $\beta = (\beta_1, \beta_2, \cdots, \beta_n)^T$ is the vector of coefficients relating to $\mathbf{Z}_S$ to be estimated.

We follow Fotheringham et al. (2002), who proposed an estimation method that is less computationally intensive than the method described by Brunsdon et al. (1999). We will broadly describe the actual estimation procedure and present the estimators for the parameters, but we do not provide the exact MGWR algorithm. For details, the readers can refer to Fotheringham et al. (2002), Mei et al. (2006), and Geniaux and Napoléone (2008). To economize on notation, we write the GWR projection or hat

matrix as

$$\mathbf{S} = \begin{pmatrix} z_{1L}\left[\mathbf{Z}_L^T\mathbf{W}(x_1,y_1)\mathbf{Z}_L\right]^{-1}\mathbf{Z}_L^T\mathbf{W}(x_1,y_1) \\ z_{2L}\left[\mathbf{Z}_L^T\mathbf{W}(x_2,y_2)\mathbf{Z}_L\right]^{-1}\mathbf{Z}_L^T\mathbf{W}(x_2,y_2) \\ \vdots \\ z_{nL}\left[\mathbf{Z}_L^T\mathbf{W}(x_n,y_n)\mathbf{Z}_L\right]^{-1}\mathbf{Z}_L^T\mathbf{W}(x_n,y_n) \end{pmatrix}$$

where $\mathbf{W}(x_i,y_i) = \text{diag}\left[w_1(x_i,y_i), w_2(x_i,y_i), \cdots, w_n(x_i,y_i)\right]$. The calibration of the model consists of four steps:

1. regressing each column of $\mathbf{Z}_S$ against $\mathbf{Z}_L$ using the GWR calibration method and computing the residuals $\mathbf{Q} = (\mathbf{I} - \mathbf{S})\mathbf{Z}_S$;

2. regressing the dependent variable $\mathbf{P}$ against $\mathbf{Z}_L$ using the GWR approach and then computing the residuals $\mathbf{R} = (\mathbf{I} - \mathbf{S})\mathbf{P}$;

3. regressing the residuals $\mathbf{R}$ against the residuals $\mathbf{Q}$ using OLS in order to obtain the estimates $\hat{\boldsymbol{\beta}} = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{R}$;

4. subtracting $\mathbf{Z}_S\hat{\boldsymbol{\beta}}$ from $\mathbf{P}$ and regressing this part against $\mathbf{Z}_L$ using GWR to obtain estimates $\hat{\alpha}(x_i,y_i) = \left[\mathbf{Z}_L^T\mathbf{W}(x_i,y_i)\mathbf{Z}_L\right]^{-1}\mathbf{Z}_L^T\mathbf{W}(x_i,y_i)(\mathbf{P} - \mathbf{Z}_S\hat{\boldsymbol{\beta}})$.

The predicted values for the property prices can be expressed as

$$\hat{\mathbf{P}} = \mathbf{S}(\mathbf{P} - \mathbf{Z}_S\hat{\boldsymbol{\beta}}) + \mathbf{Z}_S\hat{\boldsymbol{\beta}} = \mathbf{L}\mathbf{P}, \tag{17}$$

with $\mathbf{L} = \mathbf{S} + (\mathbf{I} - \mathbf{S})\mathbf{Z}_S\left[\mathbf{Z}_S^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{Z}_S\right]^{-1}\mathbf{Z}_S^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})$.

The parameter estimates and the predicted property prices obviously depend on the choice of weights, hence on the choice of bandwidth h. The optimal value for h is determined by minimizing the CV statistic given by (15). In the case of MGWR, the CV statistic is equivalent to (Mei et al., 2006)

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_i - \hat{p}_i(h)}{1 - l_{ii}(h)}\right]^2 \tag{18}$$

where $\hat{p}_i(h)$ is the predicted price for property i and $l_{ii}(h)$ is the i_{th} diagonal element of matrix $\mathbf{L}$ in equation (17).

§ 6.4 Hedonic imputation price indexes

This section addresses the issue of estimating quality-adjusted property price indexes. Suppose that sample data is available for periods $t = 0, \cdots, T$, where 0 is the base period (the starting period of the time series we want to construct), and suppose model (12) has been estimated separately for each period. The predicted property prices,

obtained using MGWR, are given by $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \left[\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{r=1}^{R-1} \hat{\lambda}_r^t D_{ir}^t\right] z_{iS}^t$. For short, we write the predicted price of structures, $\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{r=1}^{R-1} \hat{\lambda}_r^t D_{ir}^t$, as $\hat{\beta}_i^t$ and the predicted overall property price as $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t$ $(t = 0, \cdots, T)$.

We denote the sample of properties sold in the base period by S^0. The hedonic imputation Laspeyres property price index going from period 0 to period t is defined by

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \hat{p}_i^{t(0)}}{\sum_{i \in S^0} \hat{p}_i^0} \tag{19}$$

Equation (19) may need some explanation. All quantities are equal to 1, reflecting the fact that each property is considered unique. The index is not affected by compositional change because it is based on a single sample. Most, if not all, of the properties sold in period 0 are not re-sold in period t, and the 'missing prices' have to be imputed by $\hat{p}_i^{t(0)}$. We have also replaced the observed base period prices p_i^0 by the predicted values $\hat{p}_i^0$, a method known as *double imputation*[1].

The $\hat{p}_i^{t(0)}$ are estimated period t constant-quality property prices, i.e. estimates of the prices that would prevail in period t for properties sold in period 0 if the properties' price-determining characteristics were equal to those of the base period, which serves to adjust for quality changes of the individual properties. These constant-quality prices are estimated by $\hat{p}_i^{t(0)} = \hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0$, where $\hat{\beta}_i^{t(0)} = \hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^0 + \sum_{r=1}^{R-1} \hat{\lambda}_r^t D_{ir}^0$ denotes the estimated constant-quality price of structures.

Substitution of $\hat{p}_i^0 = \hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0$ and $\hat{p}_i^{t(0)} = \hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0$ into (19) yields

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \left[\hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0\right]}{\sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]} = \hat{s}_L^0 \frac{\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0}{\sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0} + \hat{s}_S^0 \frac{\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0}{\sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0} \tag{20}$$

where $\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0 / \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0$ is a price index of land and $\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0 / \sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0$ is a price index of structures. Equation (20) decomposes the overall house price index into structures and land components; the weights $\hat{s}_L^0 = \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0 / \sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]$ and $\hat{s}_S^0 = \sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0 / \sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]$ are estimated shares of land and structures in the total value of property sales in period 0. The double imputation method ensures that the weights sum to unity.

The price indexes of land and structures in (20) are Laspeyres-type indexes and can be written as weighted averages of price relatives for the individual properties. For example, the Laspeyres price index of land can be written as $\sum_{i \in S^0} \hat{s}_{iL}^0 \left(\hat{\alpha}_{iL}^t / \hat{\alpha}_{iL}^0\right)$, where the weights $\hat{s}_{iL}^0 = \hat{\alpha}_i^0 z_{iL}^0 / \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0$ for the price relatives $\hat{\alpha}_{iL}^t / \hat{\alpha}_{iL}^0$ reflect the shares of

1 Hill and Melser (2008) discuss different types of hedonic imputation methods in the context of housing. For a general discussion of the difference between hedonic imputation indexes and time dummy indexes, see Diewert et al. (2009) and de Haan (2010).

the properties in the estimated value of land (implicitly) sold in period 0. Properties with relatively large value shares, like properties in wealthy and sought-after neighborhoods with large plot sizes and high land prices, therefore have a big influence on the index.

An alternative to the Laspeyres index is the hedonic double imputation Paasche price index, defined on the sample S^t of properties sold in period $t (t = 1, \cdots, T)$:

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \hat{p}^t_i}{\sum_{i \in S^t} \hat{p}^{0(t)}_i}. \tag{21}$$

The imputed constant-quality prices $\hat{p}^{0(t)}_i$ are estimates of the prices that would prevail in period 0 if the property characteristics were those of period t, which are estimated as $\hat{p}^{0(t)}_i = \hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}$, where $\hat{\beta}^{0(t)}_i = \hat{\theta}^0 + \sum_{a=1}^{A-1} \hat{\gamma}^0_a D^t_{ia} + \sum_{r=1}^{R-1} \hat{\lambda}^0_r D^t_{ir}$ denotes the period 0 constant-quality price of structures. By substituting the constant-quality prices and the predicted prices $\hat{p}^t_i = \hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS}$ into (21), the hedonic imputation Paasche index can be written as

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \left[\hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS}\right]}{\sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]} = \hat{s}^{t(0)}_L \frac{\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL}}{\sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}} + \hat{s}^{t(0)}_S \frac{\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS}}{\sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}} \tag{22}$$

where $\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL} / \sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}$ and $\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS} / \sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}$ are Paasche price indexes of land and structures, which are weighted by $\hat{s}^{t(0)}_L = \sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL} / \sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]$ and $\hat{s}^{t(0)}_S = \sum_{i \in S^t} \hat{\beta}^0_i z^t_{iS} / \sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}\right]$. The weights are now of a hybrid nature and reflect the shares of land and structures in the estimated total value of property sales in period t, evaluated at base period prices.

A drawback of the above indexes is that they are based on the sample of either the base period or the comparison period t, but not on both samples. When constructing an index going from 0 to t, the sales in both periods should ideally be taken into account in a symmetric fashion. The double imputation Fisher price index

$$P^{0t}_{Fisher} = \left[P^{0t}_{Laspeyres} \times P^{0t}_{Paasche}\right]^{\frac{1}{2}} \tag{23}$$

does so by taking the geometric mean of the Laspeyres and Paasche price indexes. Note that, because the Fisher index number formula is not consistent in aggregation, it is not possible to provide an exact decomposition of the Fisher property index into structures and land components.

Double imputation Laspeyres, Paasche and Fisher property price indexes and the land price indexes based on the more restrictive hedonic models (10) or (8) are found by replacing $\hat{\alpha}^0_i$ and $\hat{\alpha}^t_i$ in (20) and (22) by the corresponding postcode-specific estimates

$\hat{\alpha}_k^0$ and $\hat{\alpha}_k^t$ or the city-wide estimates $\hat{\alpha}^0$ and $\hat{\alpha}^t$. In the latter case, the estimated land price index of course equals $\hat{\alpha}^t/\hat{\alpha}^0$, irrespective of the index number formula used.

§ 6.5 Empirical evidence

§ 6.5.1 The data set

The data set we utilize was provided by the Dutch Association of Real Estate Agents. It contains residential property sales for a small city (population is around 60,000) in the northeastern part of the Netherlands, the city of "A", and covers the first quarter of 1998 to the fourth quarter of 2007. Statistics Netherlands has geocoded the data. We decided to exclude sales on condominiums and apartments since the treatment of land deserves special attention in this case. The resulting total number of sales in the data set during the ten-year period is 6,058, representing approximately 75% of all residential property transactions in "A".

The data set contains information on the time of sale, transaction price, a range of structures characteristics, and land characteristics. We included only three structures characteristics in our models, i.e., usable floor space, building period and type of house. For land, we used plot size and postcode or latitude/longitude. Initially, we deleted 43 observations with missing values or prices below €10,000, properties with more than 10 rooms and those with ratios of plot size to structure size (usable floor space) larger than 10 as well as transactions in rural areas. Finally, we removed 32 outliers or influential observations detected by Cook's distance and were left with 5,983 observations during the sample period.

Table A1 in the Appendix reports summary statistics by year for the numerical variables. Both the average transaction price and the price per square meter significantly increased from 1998 to 2007. Average land size and usable floor space were quite stable over time. The urban area of the city of "A" seems to have expanded along the east-west axis; the standard deviation of the *x* coordinate in later years is generally much larger than that in earlier years.

§ 6.5.2 Estimation results for hedonic models

Given the small size of the city of "A" and the resulting low number of observations, we decided to use annual rather than quarterly data. We estimated three normalized hedonic models: model (8), which does not include location (denoted by OLS), model (10) with 9 postcode dummy variables (OLSD), and model (12) with property-specific land prices (MGWR).

When estimating the MGWR model, we used the adaptive bi-square function to construct the weighting scheme, given that the transactions were not evenly distributed across space. In this case, the bandwidth is generally referred to as the window size, and the choice of window size is equivalent to the choice of the number of nearest neighbors. To find the optimal value, we varied the window size from 10% to 95% using a 5% interval and selected the size that yielded the lowest CV score as given by equation (18). Each annual sample has a unique optimal window size. The CV scores indicated that a 10% window size was preferred for most of the years, except for 1999, 2000 and 2002, with an optimal size of 15%, and 2003, with an optimal size of 30%. However, for the construction of price indexes, we would prefer using the same window size for all years, especially since the number of sales is almost evenly spread across the whole period. So we chose a window size of 10% for each year, leading to 60 nearest neighbors that were used in the estimation of the annual MGWR models.

TABLE 6.1 Parameter estimates for structures characteristics, 2007

	OLS	OLSD	MGWR
Intercept	1480.70***	1405.41***	1395.76***
	(46.93)	(53.71)	(57.51)
Building period:1960-1970	-370.48***	-389.50***	-398.40***
	(25.94)	(36.67)	(41.75)
Building period:1971-1980	-311.17***	-261.50***	-323.50***
	(23.36)	(33.96)	(41.75)
Building period:1981-1990	-232.93***	-173.08***	-226.14***
	(23.37)	(32.59)	(42.87)
Building period:1991-2000	-58.64***	-49.34*	-115.13***
	(21.64)	(26.55)	(37.26)
Terrace	-285.65***	-264.34***	-187.28***
	(35.17)	(35.24)	(37.32)
Corner	-281.36***	-274.54***	-192.85***
	(31.77)	(31.18)	(34.07)
Semidetached	-122.89**	-149.50***	-96.93**
	(47.96)	(47.57)	(48.73)
Duplex	-151.08***	-147.24***	-104.56***
	(30.60)	(30.17)	(31.03)

Notes:Standard errors are reported in parentheses; ***, ** and * denote significance at the 1% , 5% and 10% level, respectively.

As an illustration, Table 6.1 shows the 2007 parameter estimates for the structures characteristics. Almost all of the estimates differ significantly from zero at the 1% level. To some extent they vary across the different models. For example, the OLS intercept term is relatively high compared to the OLSD and MGWR intercepts. Note that, since dummy variables for houses built after 2000 and for detached houses were not included, the intercept measures the price in euros of structures per square meter of living space for detached houses built after 2000. In accordance with a priori expectations, detached dwellings are more expensive than other types of houses. For

all models, there is a clear tendency for the structures to become less expensive as they are getting older.

TABLE 6.2 Summary statistics for estimated land prices

	OLS	OLSD		MGWR				
		Mean	S.D.	Min	Max	Median	Mean	S.D.
1998	116.80	131.50	31.14	72.30	231.03	122.66	125.49	28.66
1999	154.64	178.50	34.85	105.95	223.66	174.07	167.77	30.39
2000	239.77	239.41	36.24	138.53	319.32	251.34	241.83	44.27
2001	214.54	235.58	47.59	110.41	295.01	229.52	226.70	48.77
2002	234.77	245.11	38.41	156.15	323.63	255.05	242.23	40.89
2003	166.07	185.11	44.23	82.12	248.23	179.93	172.26	44.55
2004	186.40	197.19	29.75	104.95	254.20	197.70	195.41	33.78
2005	226.13	224.11	36.55	127.53	299.74	214.19	205.89	35.17
2006	202.84	195.77	30.85	125.90	274.24	207.43	201.27	32.05
2007	214.87	236.73	27.96	141.46	286.91	235.07	229.25	30.99

Notes: For OLS, the land price estimates are reported. For OLSD, the columns show the weighted mean and standard deviation of the estimated land prices for 9 postcode areas where the weights are equal to the share of transactions within each postcode area. For MGWR, the columns provide summary statistics for the land price estimates of all transacted properties.

Table 6.2 contains summary statistics for the estimated price per square meter of land from the three models. The three average land price series exhibit a similar pattern over time, which differs substantially from the changes in the average transaction price of the properties (see Table A1 in the Appendix). After a sharp increase in 1999, the estimated average land price fluctuated during a couple of years, experienced a dramatic drop in 2003, and then increased again.

As mentioned earlier, a virtue of MGWR is that it allows us to plot a continuous map with estimated prices of land per square meter. For the year 2007, such a map is depicted in Figure 6.1 for the city of "A", where the land prices have been rescaled to the range [0,1]. The postcode areas are indicated as well. While the spatial pattern in Figure 6.1 is largely consistent with the pattern found using the OLSD model (shown in Figure A1 in the appendix), the MGWR land prices estimates do vary within some of the areas. This suggests that the use of postcode dummies, as in the OLSD model, is a rather crude strategy to incorporate spatial variation of land prices.

To formally compare the performance of the three hedonic models, two statistics were calculated, the Corrected Akaike Information Criterion (AICc) and the Root Mean Square Error (RMSE). The AICc takes into account the trade-off between goodness of fit and degrees of freedom. The AICc expressions for the OLS and OLSD models can be found in Hurvich and Tsai (1989). And for MGWR models, it is defined by

$$AICc = 2nln(\hat{\sigma}) + nln(2\pi) + n\left(\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right)$$

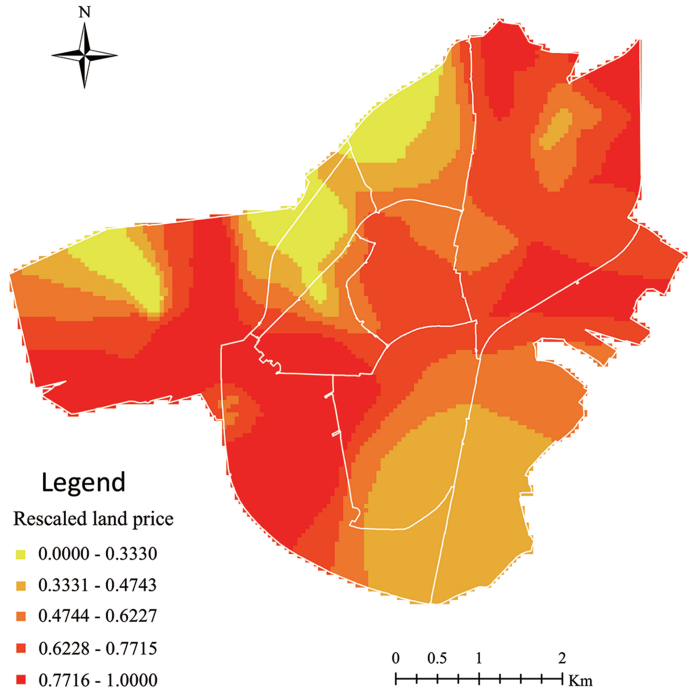

FIGURE 6.1 Price of land per square meter, 2007

where $\hat{\sigma}$ is the estimated standard deviation of the error term and $tr(\mathbf{S})$ the trace of the hat matrix described in section 6.3.2. The RMSE measures the variability of the absolute prediction errors of the models and is given by

$$RMSE = \frac{1}{n}\sqrt{\sum_i (p_i - \hat{p}_i)^2}.$$

Table 6.3 shows the AICc and RMSE and their differences for the three models. A rule of thumb states that if the difference in the AICc for two models is larger than 3, a significant difference exists in their performance (Fotheringham et al. 2002). It can be seen that the OLSD model performs much better than the OLS model in all years, as we would expect, and in turn that the MGWR model outperforms the OLSD model (except for 2003, when the difference is insignificant). The same ranking is found if the RMSE is used to assess the various models. These results confirm the earlier finding that land prices vary across space, both across and within postcode areas.

Although MGWR is obviously better suited to model the variation of land prices and to predict property prices, the OLSD model does a surprisingly good job. In several years,

TABLE 6.3 Model comparison

Dependent variable = House price per square metre of living space												
	OLS		OLSD				MGWR					
	AICc	RMSE	AICc	ΔAIC_{DO}	RMSE	$\Delta RMSE_{DO}$	AICc	ΔAIC_{MD}	ΔAIC_{MO}	RMSE	$\Delta RMSE_{MD}$	$\Delta RMSE_{MO}$
1998	6487.71	91.32	6372.45	-115.26	80.89	-10.43	6366.21	-6.24	-121.50	77.85	-3.04	-13.47
1999	7056.56	146.82	6990.52	-66.04	136.14	-10.68	6982.93	-7.59	-73.63	131.28	-4.86	-15.54
2000	7216.89	151.11	7164.26	-52.63	142.00	-9.11	7127.51	-36.75	-89.38	133.30	-8.70	-17.81
2001	7380.41	147.00	7294.10	-86.31	134.36	-12.64	7279.66	-14.44	-100.75	128.87	-5.49	-18.13
2002	7718.63	152.44	7643.97	-74.66	141.19	-11.25	7632.34	-11.63	-86.29	135.62	-5.57	-16.82
2003	7769.06	159.02	7702.07	-66.99	148.23	-10.79	7701.91	-0.16	-67.15	143.93	-4.30	-15.09
2004	7968.62	159.66	7947.61	-21.01	154.80	-4.86	7927.92	-19.69	-40.70	147.91	-6.89	-11.75
2005	8060.84	161.52	7993.88	-66.96	150.93	-10.59	7984.11	-9.77	-76.73	145.10	-5.83	-16.42
2006	8597.81	175.46	8565.36	-32.45	168.94	-6.52	8517.73	-47.63	-80.08	157.67	-11.27	-17.79
2007	9006.25	177.18	8960.58	-45.67	169.24	-7.94	8929.11	-31.47	-77.14	159.98	-9.26	-17.20

Notes: ΔAIC_{DO} is equal to AICc for OLSD minus AICc for OLS; ΔAIC_{MD} and ΔAIC_{MO} are equal to AICc for MGWR minus AICc for OLSD and OLS, respectively; $\Delta RMSE_{DO}$, $\Delta RMSE_{MD}$ and $\Delta RMSE_{MO}$ have a similar meaning.

for example in 1998, 1999 and 2003, the inclusion of postcode dummy variables accounts for the major part of the variance in overall property prices, almost as much as the MGWR model does.

§ 6.5.3 Hedonic imputation price indexes

Changes in average property prices and their land and structure components are affected by compositional change and quality change of the traded properties. The hedonic house price indexes and the land and structures components that we estimated control for these effects. We estimated chained rather than direct indexes because imputing the 'missing prices' over a very long period of time may not be useful and because the value shares of land and structures will then be updated annually. A drawback of chaining is that the resulting price indexes cannot be exactly decomposed because they are not consistent in aggregation.

In Figures 6.2-6.4, the estimated double imputation hedonic Laspeyres, Paasche and Fisher price indexes for the overall property are plotted, based on the three models (OLS, OLSD, and MGWR). A comparison of Figures 6.2 and 6.3 shows that, for each model, the chained Laspeyres index sits above the Paasche index, as expected. The Laspeyres and Paasche indexes based on OLSD and MGWR are very similar; for the Laspeyres index, the difference can even hardly be noticed. This result is in accordance with our finding that the OLSD model captures the spatial nonstationarity of land prices reasonably well.

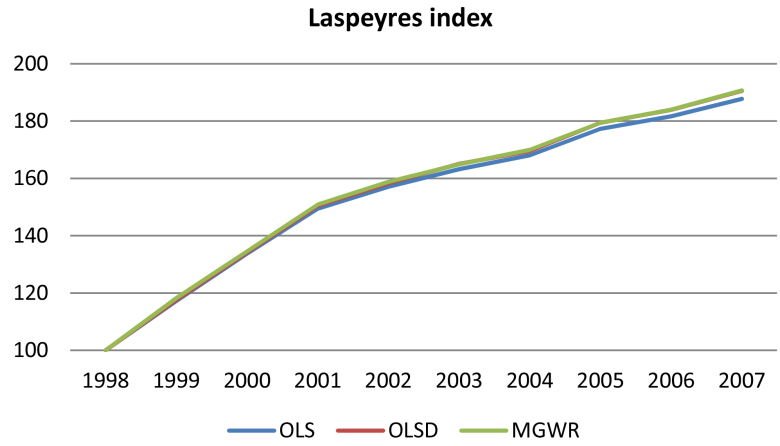

FIGURE 6.2 Hedonic imputation Laspeyres house price index

Not using location information at all does make a difference though, at least for the

Laspeyres and Paasche house price indexes. The OLS-based Laspeyres and Paasche indexes seem to be biased downwards and upwards, respectively. However, the biases almost cancel out in the Fisher index: the OLS-based Fisher index is very similar to the OLSD-based and MGWR-based Fisher indexes. In other words, the hedonic imputation Fisher house price index is insensitive to the treatment of location in the hedonic model, which is a surprising result.

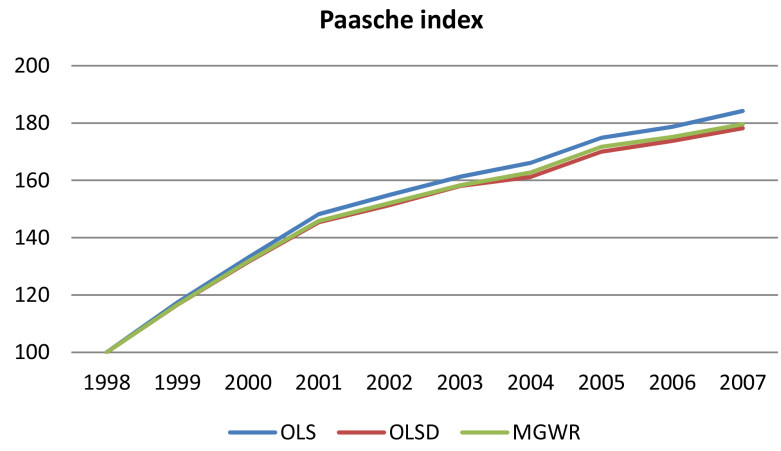

FIGURE 6.3 Hedonic imputation Paasche house price index

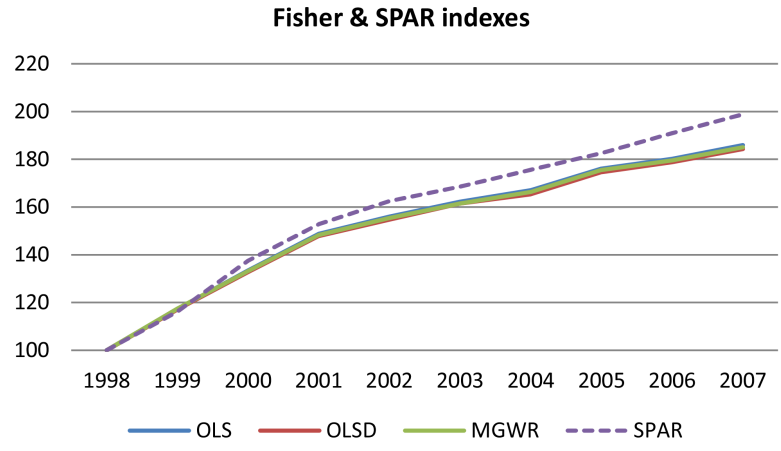

FIGURE 6.4 Hedonic imputation Fisher house price index and SPAR index

The house price index for the Netherlands published by Statistics Netherlands is also plotted in Figure 6.4. This official index is based on the Sale Price Appraisal Ratio (SPAR) method (de Haan et al. 2009; de Vries et al. 2009). Our hedonic indexes show a

much more modest price increase. There may be two reasons for this. First, house prices in the city of "A" appreciated less compared to the rest of the country. Second, our indexes better adjust for quality changes while the SPAR method only adjusts for compositional change of the properties sold. We think that the second reason is more important.

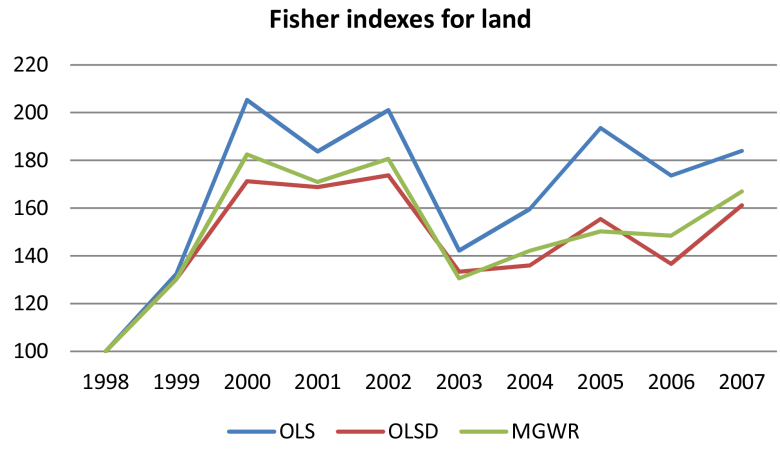

FIGURE 6.5 Hedonic imputation Fisher price indexes for land

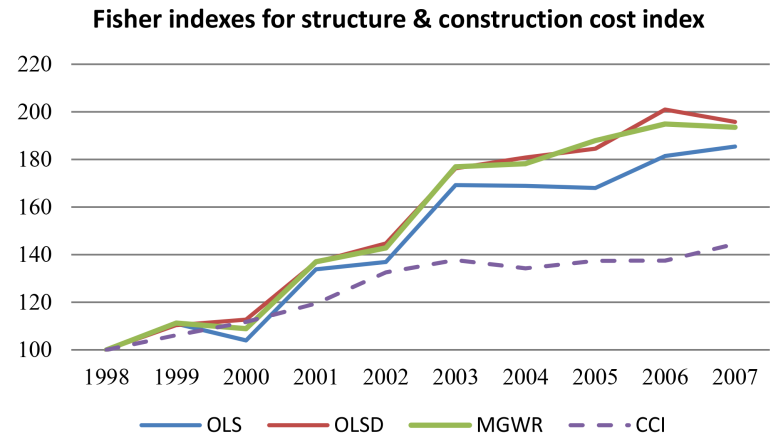

FIGURE 6.6 Hedonic imputation Fisher price indexes for structures and official construction cost index

The picture changes when we look at the Fisher indexes for the price of land in Figure 6.5. The OLSD- and MGWR-based indexes, which explicitly account for location, are

similar, although the MGWR-based index is less volatile, at least during 2003-2007. However, the OLS-based index seems to be significantly upward biased. For example, between 1999 and 2000 as well as between 2003 and 2005, the OLS-based index rises much faster than the other two indexes.

Figure 6.6 shows the Fisher price indexes for structures based on the three models. Again, the OLSD-based and MGWR-based indexes are similar. The OLS-based index sits below the other indexes, but the difference is less pronounced than for land. This is in line with our expectations: location should affect the price of land and is modeled accordingly, but it should leave the price of structures unaffected.

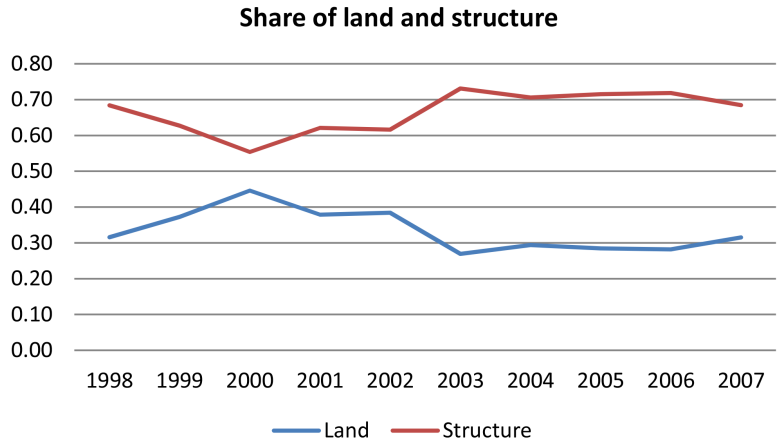

FIGURE 6.7 Estimated value shares of land and structures, MGWR-based

Figure 6.7 shows the MGWR-based value share estimates for both structures and land. Prior to 2003, these shares are quite volatile, but from 2003 on they remain fairly constant. The average estimated shares for structures and land across the entire sample period are 0.67 and 0.33. The OLS- and OLSD-based shares show similar patterns and levels; the average shares for structures are 0.68 and 0.66, respectively, hence for land 0.32 and 0.34. Given that the estimated value share of structures is twice as large as that of land, the overall house price indexes are affected most by changes in structures prices. Yet, combining Figures 6.4, 6.5, 6.6 and 6.6 suggests that the increase in house prices between 1998 and 2001 has been driven mainly by the increase of land prices: both the (average) price of land and its value share show a sharp increase.

§ 6.5.4 Discussion

Figures 6.5, 6.6 and 6.7 raise a number of issues. The first issue is the volatility of the land and structures price indexes. Volatile series can of course be expected with sparse data (and without smoothing). Another cause might be multicollinearity. Diewert et al. (2015) found that multicollinearity (between land and structure size) led to price changes for land and structures which consistently had opposite signs. To deal with this form of multicollinearity, some studies (e.g.,Diewert et al. 2009; Diewert and Shimizu 2013; Francke and van de Minne 2016) included exogenous information in the hedonic models; they all used the officially published construction cost index as the measure of price change for structures. Put differently, their models do not provide an endogenously determined price index of structures. We do not follow their approach because, as we discuss in the next paragraph, multicollinearity does not seem to be the most important issue and because the trend of the endogenous price index of structures might be more consistent with the evolution of the market values of structures.

In Figure 6.8, the MGWR-based Fisher price indexes for land and structures from Figures 6.5 and 6.6 are copied. In some years, for example in 2003 when the land price index suddenly falls and starts to sit below the structures price index, the price changes for land and structures have opposite signs, but in other years the price changes are in the same direction. The variance inflation factor (VIF) for the ratio of plot size to structure size did not point to significant multicollinearity either. Further, there is a considerable amount of variation in these ratios in our data set; see Table A1. We therefore suspect that multicollinearity is not the main issue.

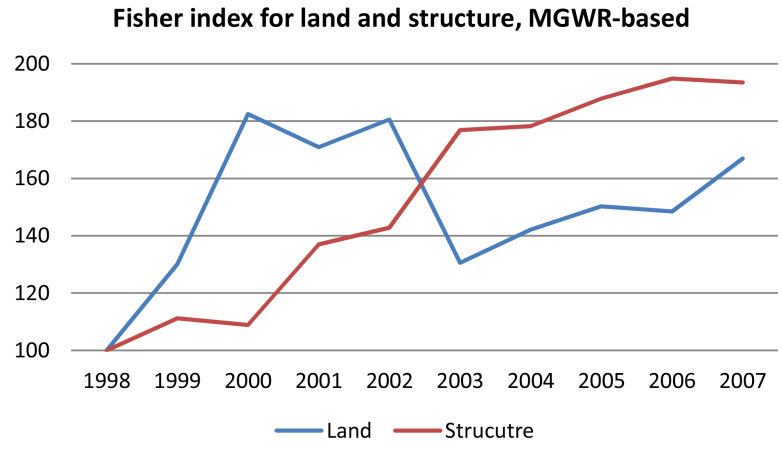

FIGURE 6.8 Chained Fisher price indexes for land and structures, MGWR-based

The second issue is whether the trends of the (Fisher) price indexes for land and structures are plausible. For land, this cannot be checked since information on the

price change of land is not available for the Netherlands [2]. For structures we use the official nationwide construction cost index (CCI) for new dwellings as a benchmark. This price index, rebased to 1998=100, is shown in Figure 6.6 as well. Our structures price indexes rise much faster than the construction cost index, especially during the second half of the sample period when the construction cost index flattens.

At first, a construction cost index does not necessarily have to coincide with an implicit price index for structures derived from a hedonic model. Since structure is producible, it is believed that, in a completely competitive market, the construction cost is equal to the market value of structure (Davis and Heathcote 2007; Davis and Palumbo 2008). This equality might be held in a stable market where the developers can take a sufficiently long time to meet the demand. However, the market of structures in reality tends to be less competitive due to the restriction of new constructions and the high cost of replacing old structures with new ones. In this case, it is very likely that a persistent markup is imposed on structure prices and hence it is no surprising to see the structure price index sitting above the construct cost index. This disparity can be even more striking during a housing boom, which is exactly the case of this study. Kuminoff and Pope (2013), who estimated the land values for US metropolitan areas using a refined hedonic approach that mitigates the omitted variable bias, presented a similar finding that the increase of market value of structures exceeds the growth of replacement cost in the booming period in some places. On the other hand, the flattening of the construction cost index between 2003 and 2007 has been subject of debate in the Netherlands. The discussion arose because the construction cost index increased by only 4.9%, which was even lower than the increase in the CPI of 5.8%, while house prices were still rapidly rising.

Nevertheless, a divergence that large is still a bit worrying. One of the reasons for the strong increase of our structures price indexes could be omitted variables bias – resulting in quality-change bias – because we included only a few structures characteristics in the hedonic models. Unless they are highly collinear with included variables, adding characteristics will lead to better quality adjustment for structures and lower the price indexes for structures if, as can be expected during this period of booming house prices, the quality of structures has improved over time. One obvious omitted variable that is related to depreciation of the structures is the level of maintenance.

The third issue concerns the share of land in the value of properties sold, which was estimated at roughly one third across the sample period. van de Minne and Francke (2012) estimated the share of land for properties (excluding

2 Municipalities do have information on what are sometimes referred to as realizations of the value of land sold to developers of residential construction. These realizations are determined residually, but it is doubtful whether they accurately measure the 'true' value of land.

apartments/condominiums) sold during 2003-2010 in the city of 's Hertogenbosch at 0.39 on average. In a follow-up study (Francke and van de Minne 2016), where they made a distinction between the part of the land plot that the structure sits on and the part used as gardens, the estimate was almost 0.50. It is not unreasonable to find that the value share of land for the city of "A" is lower than that for 's Hertogenbosch. The city of "A" lies in a less prosperous part of the Netherlands with fewer amenities that households appreciate, and we expect this to have a downward effect on the price of land rather than the price of structures, hence on the value share of land.

de Groot et al. (2015), also using hedonic models to decompose property values into land and structures components, estimated the price of land for most Dutch cities, though unfortunately not for "A". They found substantial cross-city differences. For example, the price per square meter of land in 2005 was estimated at 717 euros for the capital city of Amsterdam, 308 euros for 's-Hertogenbosch, and 184 euros for Leeuwarden. Like "A", Leeuwarden is a city in the northeastern part of the Netherlands but bigger. In light of their findings, our MGWR estimates of the average price of land for the city of "A", 206 euros in 2005 (Table 6.2), and the value share of land are not surprisingly low after all.

§ 6.6 Summary and conclusions

Land is often not explicitly included in hedonic models for house prices, which can bias the results. Ignoring spatial nonstationarity of land prices can also generate bias. As far as we know, the present paper is the first attempt to account for nonstationarity of land prices in the construction of hedonic imputation house price indexes. We linearized the 'builder's model' proposed by Diewert et al. (2015), allowed the price of land to vary across individual properties, and estimated the model for the normalized property price (the price of the property per square meter of living space) by MGWR, a semi-parametric method, on annual data for the Dutch city of "A". We then constructed chained imputation Laspeyres, Paasche and Fisher indexes, and compared these indexes with price indexes based on more restrictive models, i.e. a model where land prices vary across postcode areas and a model with no variation in land prices and, both estimated by OLS.

The Fisher house price indexes were quite insensitive to the choice of model, but the Laspeyres and Paasche indexes for the 'fixed' land price model differed from those for the models where location was explicitly included. The use of postcode area dummy variables produced price indexes very similar to indexes obtained by MGWR. Hill and Scholz (2014), who treated location as a 'separate characteristic' in their hedonic models in that they estimated property-specific shift terms for the overall property

price, also concluded that the use of geocoded information did not significantly improve hedonic imputation house price indexes compared to indexes based on models with postcode dummy variables. This result is reassuring for statistical agencies that do not have the expertise or resources to apply more sophisticated methods. It should be noted that the similarity between OLSD-based and MGWR-based house price indexes could also be due to the small size and homogeneity of the city "A" where relatively little variation of land prices can be expected.

Apart from being able to capture spatial variation of land prices at the property level, the MGWR model has two additional advantages. A potential problem with the OLSD model is that if a large number of postcode areas are distinguished, observations in some areas may not be available, leading to difficulties in the construction of hedonic imputation price indexes. The MGWR method deals with this problem by using data of the nearest neighbors which are not necessarily confined to a particular postcode area. Most importantly, the use of nearest-neighbor information in the (semi-parametric) MGWR method makes it possible to properly account for spatial effects in the absence of detailed information on amenities, such as the availability of, and distance to, public transport, green space, schools, shopping centers, and so on.

For some purposes, separate price indexes for land and structures are needed. As was demonstrated already by Diewert et al. (2015), the decomposition into land and structures using hedonic modeling is not straightforward and raises several statistical and functional form issues. First, our MGWR-based price indexes of land and structures for the city of "A" are quite volatile, in spite of the use of annual data, which can be attributed to the sparse data in combination with possibly multicollinearity (though we believe this is less important). Second, the structures price index increases much faster than expected, perhaps due to omitted variables or quality-change bias, i.e. a failure to fully control for changes in structures characteristics. Third, the estimated value share of land seems rather low. The above-mentioned problems may have played a role here, but the low land share could also be a real phenomenon: households do not value a square meter of land in the city of "A" as much as they would do in more prosperous cities with more and better amenities. Anyhow, in future work it would be useful to re-examine our models and compare the results for the city of "A" with those for bigger and more densely populated cities in the western part of the country, like Amsterdam, Rotterdam or The Hague. Having more observations might also enable us to estimate biannual or even quarterly price indexes.

Functional form problems might be even more important. The original 'builder's model' is nonlinear, in particular due to the treatment of net depreciation. We linearized the model, which basically means we ignored interaction terms. Another potential type of misspecification arises from the linear relationship between land price and plot size in our models. As Diewert et al. (2015), Francke and van de Minne (2016) and others have argued that the marginal price of land tends to decrease with plot size.

Diewert et al. (2015) accounted for this form of nonlinearity by using linear splines. In future work we may modify our 'normalized' models by using linear splines as well and estimating different parameters for the plot size to structure size ratio for different categories of lot size or by explicitly specifying some nonlinear function of this ratio.

References

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281-298.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1999). Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science, 39(3), 497-524.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographical research. Geographical Analysis, 4(1), 81-91.

Davis, M. A., & Heathcote, J. (2007). The price and quantity of residential land in the United States. Journal of Monetary Economics, 54(8), 2595-2620.

Davis, M. A., & Palumbo, M. G. (2008). The price of residential land in large US cities. Journal of Urban Economics, 63(1), 352-384.

de Groot, H. L. F., Marlet, G., Teulings, C., & Vermeulen, W. (2015). Cities and the Urban Land Premium. Cheltenham: Eaward Elgar.

de Haan, J. (2010). Hedonic price indexes: A comparison of imputation, time dummy and 're-pricing' methods. Jahrbücher für Nationalökonomie und Statistik, 230(6), 772-791.

de Haan, J., van de Wal, E., & de Vries, P. (2009). The measurement of house prices: A review of the sale price appraisal ratio method. Journal of Economic and Social Measurement, 34(2,3), 51-86.

de Vries, P., de Haan, J., van de Wal, E., & Mariën, G. (2009). A house price index based on the SPAR method. Journal of Housing Economics, 18(3), 214-223.

Diewert, W. E., de Haan, J., & Hendriks, R. (2011). The decomposition of a house price index into land and structures components: A hedonic regression approach. The Valuation Journal, 6(1), 58-105.

Diewert, W. E., de Haan, J., & Hendriks, R. (2015). Hedonic regressions and the decomposition of a house price index into land and structure components. Econometric Reviews, 34(1-2), 106-126.

Diewert, W. E., Heravi, S., & Silver, M. (2009). Hedonic imputation versus time dummy hedonic indexes. In W. E. Diewert, J. S. Greenlees, & C. R. Hulten (Eds.), Price Index Concepts and Measurement (pp. 161-196). Chicago: University of Chicago Press.

Diewert, W. E., & Shimizu, C. (2013). Residential property price indexes for Tokyo. UBC Vancouver School of Economics Discussion Papers 2013-16, The University of

British Columbia.
Dorsey, R. E., Hu, H., Mayer, W. J., & Wang, H. (2010). Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle. Journal of Housing Economics, 19(2), 75-93.
Eurostat, ILO, IMF, OECD, UNECE, & World Bank (2013). Handbook on Residential Property Price Indices. Luxemburg: Publications Office of the European Union.
Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (1998a). Scale issues and geographically weighted regression. In N. J. Tate, & P. M. Atkinson (Eds.), Modelling Scale in Geographical Information Science (pp. 123-140). Chichester: Wiley.
Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: John Wiley & Sons.
Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (1998b). Geographically weighted regression: a natrual evolution of the expansion method for spatial data analysis. Environment and Planning A, 30(11), 1905-1927.
Francke, M. K., & van de Minne, A. M. (2016). Land, structure and depreciation. Real Estate Economics, Forthcoming.
Geniaux, G., & Napoléone, C. (2008). Semi-parametric tools for spatial hedonic models: An introduction to mixed geographically weighted regression and geoadditve models. In A. Baranzini, J. Ramirez, C. Schaerer, & P. Thalmann (Eds.), Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation (pp. 101-127). New York: Springer.
Hill, R. J., & Melser, D. (2008). Hedonic imputation and the price index problem: An application to housing. Economic Inquiry, 46(4), 593-609.
Hill, R. J., Melser, D., & Syed, I. (2009). Measuring a boom and bust: The Sydney housing market 2001 - 2006. Journal of Housing Economics, 18(3), 193-205.
Hill, R. J., & Scholz, M. (2014). Incorporating geospatial data in house price indexes: A hedonic imputation approach with splines. Graz Economics Papers 2014-05, University of Graz.
Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297-307.
Jones, J. P., & Casetti, E. (1992). Applications of the Expansion Method. London: Routledge.
Kuminoff, N. V., & Pope, J. C. (2013). The value of residential land and structures during the great housig boom and bust. Land Economics, 89(1), 1-29.
Mei, C., Wang, N., & Zhang, W. (2006). Testing the importance of the explanatory variables in a mixed geographically weighted regression model. Environment and Planning A, 38(3), 587-598.
Pace, R. K., Barry, R., Clapp, J. M., & Rodriquez, M. (1998). Spatiotemporal Autoregressive Models of Neighborhood Effects. Journal of Real Estate Finance and Economics, 17(1), 15-33.

Rambaldi, A. N., & Rao, D. S. P. (2011). Hedonic predicted house price indices using time-varying hedonic models with spatial autocorrelation. School of Economics Discussion Paper Series 432, University of Queensland.

Rambaldi, A. N., & Rao, D. S. P. (2013). Econometric modeling and estimation of theoretically consistent housing price indexes. Centre for Efficiency and Productivity Analysis Working Paper Series WP04/2013, University of Queesland.

Sun, H., Tu, Y., & Yu, S.-M. (2005). A spatio-temporal autoregressive model for multi-unit residential market analysis. The Journal of Real Estate Finance and Economics, 31(2), 155-187.

Thanos, S., Dubé, J., & Legros, D. (2016). Putting time into space: the temporal coherence of spatial applications in the housing market. Regional Science and Urban Economics, 58, 78-88.

Tu, Y., Yu, S.-M., & Sun, H. (2004). Transaction-Based Office Price Indexes: A Spatiotemporal Modeling Approach. Real Estate Economics, 32(2), 297-328.

van de Minne, A. M., & Francke, M. K. (2012). De waardebepaling van grond en opstal [The determination of the value of land and structures]. Real Estate Research Quarterly, 11, 14-24.

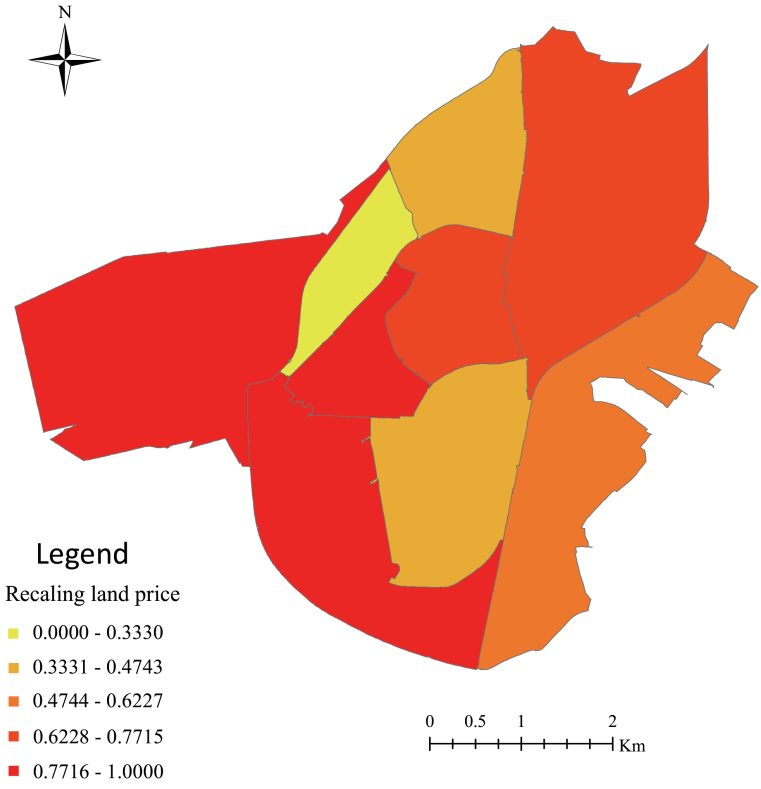

FIGURE A1 Price of land per square meter, 2007, OLSD model

TABLE A1 Summary statistics by year

	Total	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
# of obs.	5983	545	549	559	574	597	597	612	618	651	681
Transaction price (Euro)											
Mean	157073.87	95124.15	117936.77	131907.96	144672.16	151363.75	162956.98	174998.71	180882.00	191491.09	198546.51
S.D.	72782.29	40240.34	53569.32	54793.53	58064.72	53220.31	63278.10	82975.61	68777.60	76120.61	83639.92
Standardized price (Euro)											
Mean	1232.38	742.30	930.70	1039.71	1168.13	1240.63	1287.24	1353.89	1420.07	1469.62	1518.50
S.D.	374.83	206.31	273.33	279.98	293.14	285.56	285.87	296.73	294.31	321.20	348.89
Lot size (m^2)											
Mean	251.57	234.08	259.73	242.23	239.68	239.20	250.46	261.38	248.93	263.15	270.98
S.D.	148.16	135.05	169.59	132.98	120.00	115.39	145.76	163.19	136.00	149.26	187.52
Floor space (m^2)											
Mean	125.87	126.00	125.42	126.48	123.34	122.05	125.29	126.57	125.89	128.52	128.39
S.D.	30.61	23.59	31.99	31.97	29.59	28.16	29.87	36.90	30.29	31.14	30.09
Ratio of lot size to floor space											
Mean	1.96	1.81	2.04	1.89	1.93	1.97	1.96	2.01	1.93	2.01	2.04
S.D.	0.82	0.77	0.99	0.72	0.72	0.80	0.84	0.80	0.72	0.78	0.95
x-coordinate											
Mean	233733.81	233972.85	234200.97	234180.34	233948.97	234007.39	233624.00	233480.63	233519.69	233222.34	233385.19
S.D.	1796.35	1453.72	1427.35	1551.87	1716.67	1713.60	1794.99	1984.82	1927.09	1918.80	1948.29
y-coordinate											
Mean	558597.10	558739.46	558805.54	558830.14	558660.23	558721.99	558522.02	558397.61	558549.11	558429.21	558410.25
S.D.	1414.88	1436.14	1463.14	1428.62	1424.92	1410.80	1451.63	1413.94	1354.34	1322.63	1381.24