Accounting for Spatial Variation of Land Prices in Hedonic Imputation House Price Indexes

Jan de Haan[a] and Yunlong Gong[b]

19 November 2014

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 nonparametric method known as geographically weighted regression. To illustrate the impact on aggregate price change, quality-adjusted house price indexes and the land and structures components are constructed for a city in the Netherlands and compared to indexes based on more restrictive models.

Keywords: geocoded data, hedonic modeling, land and structure prices, non-parametric estimation, residential property.

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

[a] Corresponding author; Division of Process Development, IT and Methodology, Statistics Netherlands, and OTB, Faculty of Architecture and the Built Environment, Delft University of Technology; email: j.dehaan@cbs.nl.

[b] OTB, Faculty of Architecture and the Built Environment, Delft University of Technology; email: y.gong-1@tudelft.nl.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of Statistics Netherlands.

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 locality.

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 throw away information on single sales and are therefore inefficient. 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 various property characteristics and thus do not allow the estimation of, for example, price indexes of the land the structure sits on. Given the problems with repeat sales methods, we focus on hedonic indexes.

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 aggregate level, such as postcode areas, rather than at the individual property level, potentially leading to 'location bias'. Third, land is usually 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 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 since it provides information on the effect of those amenities on the prices or price changes of properties, this approach is very data intensive. Importantly, both methods cannot fully adjust for location, and so some omitted variables bias and spatial autocorrelation will most 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 (Dorsey et al., 2010; Hill et al., 2009). It is also possible to apply these 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. (1998a). An additional advantage is that we will be able to plot a detailed map 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, GWR is a rather flexible method, 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.[1] 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 second quarter of 2008, so our data set covers a period of more than 10 years. Note though that we will use annual rather than quarterly data in our empirical work. The range of structural characteristics is much broader than that in the Handbook. Finally, we include types of houses other than detached dwellings.

The paper proceeds as follows. Section 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 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 4 shows how we calculate hedonic imputation indexes. Section 5 presents empirical evidence for the Dutch city of "A". Section 6 discusses the results, identifies potential improvements and concludes.

[1] For an excellent introduction, see Hill (2013).

2. A simplification of the 'builder's model'

2.1 Some basic ideas

Our starting point is the 'builder's model' proposed by Diewert, de Haan and Hendriks (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.[2] 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 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 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

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

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 and structure

[2] We follow Diewert, de Haan and Hendriks (2015) who used living space (usable floor space) in square meters as a measure of size of the structures. Alternative measures are also possible, for instance the volume of the structure in cubic meters.

size, and $\varepsilon_i^t = u_i^t / z_{iS}^t$. 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. Dividing by z_{iS}^t is a means of adjusting for heteroskedasticity when the error variance in (2) is proportional to the square of structure size; estimating equation (3) by OLS is equivalent to estimating (2) by Weighted Least Squares (WLS) using weights equal to $1/(z_{iS}^t)^2$. This kind of error variance seems quite extreme, and so this weighting system may not be helpful to reduce the heteroskedasticity problem. Also, the ratios r_i^t (and the normalized values p_i^{t*}) will exhibit relatively little dispersion.

Some statistical agencies measure and publish changes in normalized property prices, often the price per square meter of structures in order 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 favor specification (3) over (2).

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. Omitted factors relating to land are addressed in section 3. Here we describe our approach to including additional characteristics for structures. There are two issues: depreciation and renovation of the structures has not been taken into account so far, and the use of size as the only measure of quality of the structures seems too simplistic.

Following Diewert, de Haan and Hendriks (2015), we initially assume a straight-line depreciation model. The adjusted value of the structure is $\beta^t(1-\delta^t a_i^t)z_{iS}^t$, where δ^t is the depreciation rate and a_i^t is age of the structure. Information on renovations at the level of individual dwellings is unavailable so that $-\delta^t a_i^t$ 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 $\beta^t z_{iS}^t - \beta^t \delta^t a_i^t z_{iS}^t$. Adding the second term to the right-hand side of equation (2) yields

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

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 a categorical variable. The net depreciation rate is of course categorical as well.[3] Using multiplicative dummy variables D_{ia}^{t} that take on the value 1 if in period *t* property *i* belongs to age category a $(a = 1,...,A)$ and the value 0 otherwise, and after reparameterizing such that $\beta^{t} z_{iS}^{t}$ is no longer a separate term, model (4) is equivalent to $p_{i}^{t} = \alpha^{t} z_{iL}^{t} + \sum_{a=1}^{A} \gamma^{t} D_{ia}^{t} z_{iS}^{t} + u_{i}^{t}$. To be able to use standard estimation techniques, we modify this model as follows:

$$p_{i}^{t} = \alpha^{t} z_{iL}^{t} + \sum_{a=1}^{A} \gamma_{a}^{t} D_{ia}^{t} z_{iS}^{t} + u_{i}^{t} . \qquad (5)$$

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 for other reasons.

Diewert, de Haan and Hendriks (2015) also show how to incorporate the number of rooms. The new value of the structures becomes $\beta^{t}(1-\delta^{t} a_{i}^{t})(1+\mu^{t} z_{iR}^{t}) z_{iS}^{t}$, where μ^{t} is the parameter for the number of rooms z_{iR}^{t}.[4] 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 $(r = 1,...,R)$ and the value 0 otherwise, and reparameterizing again, the extension of (5) becomes

$$p_{i}^{t} = \alpha^{t} z_{iL}^{t} + \sum_{a=1}^{A} \gamma_{a}^{t} D_{ia}^{t} z_{iS}^{t} + \sum_{r=1}^{R} \lambda_{r}^{t} D_{ir}^{t} z_{iS}^{t} + \sum_{a=1}^{A} \sum_{r=1}^{R} \eta_{ar}^{t} D_{ia}^{t} D_{ir}^{t} z_{iS}^{t} + u_{i}^{t} . \qquad (6)$$

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

[3] Diewert, de Haan and Hendriks (2015) treated approximate age as a continuous variable, despite the fact that it is in fact categorical. They found that the estimated net depreciation rate was quite volatile, which was not consistent with their a priori expectation of a stable depreciation rate, and subsequently estimated models where the depreciation rate was kept constant over time. However, we are not interested in the depreciation rate itself and accept any volatility.

[4] Note that Diewert, de Haan and Hendriks (2015) did not allow the parameter to change over time.

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

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 $(a = 1,..., A)$ and category r $(r = 1,..., R)$ 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

$$p_i^{t*} = \theta^t + \alpha^t r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R-1} \lambda_r^t D_{ir}^t + \varepsilon_i^t \, . \qquad (8)$$

We included an intercept term θ^t and 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 structures can be included in a similar way as was done for the number of rooms. As a matter of fact, in our empirical work we will use type of house instead of the number of rooms.

3. Land and spatial nonstationarity

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, the price of structures is most likely to be approximately constant across space, at least within relatively small regions or cities. 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 $(k = 1,..., K)$ and denoted by

α_k^t. 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

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

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

$$p_i^{t*} = \theta^t + \sum_{k=1}^{K} \alpha_{k(K)}^t D_{ik} r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{r=1}^{R-1} \lambda_r^t D_{ir}^t + \varepsilon_i^t. \qquad (10)$$

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 found by assuming that the price of land can in principle differ at the individual property level, i.e. at the micro location. We denote the property-specific land price by α_i^t, yielding

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

and

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

Models (11) and (12) obviously cannot be estimated by standard regression techniques. In section 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 obtained by using models (9) or (10).

3.2 Accounting for spatial variation of land prices

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 seen 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., 1998b).

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., 1998a).[5] Let us remove the structural 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, but it should be clear that we estimate all models for each time period separately. Note also that the prices of land can be estimated for all points in space, not just for the sample observations, enabling us to depict a surface of land prices for the entire study area.

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 weight w_{ij} should be 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. In this paper we have chosen the frequently-used *bi-square function* given by:

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

where *h* denotes the bandwidth defining the rate of decrease in terms of distance. 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. This bias-variance trade-off motived us to choose the bandwidth by minimizing the *cross-validation* (CV) statistic

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

[5] For a comparison of geographically weighted regression and the spatial expansion method, see Bitter et al. (2007).

where $\hat{y}_{\neq i}(h)$ is the fitted value of y_i with the observations for point i 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 structural 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) in which some parameters are spatially fixed and the remaining parameters are allowed to vary across space. To describe the estimation procedure, it is 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 $\boldsymbol{\alpha} = (\alpha(x_1, y_1), \alpha(x_2, y_2), ..., \alpha(x_n, y_n))^T$ 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 $\mathbf{Z}_L$, and $\mathbf{Z}_S$ is the matrix of structural characteristics included in model (11), given by

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

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

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).[6] To economize on notation, we write the GWR hat matrix as

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

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

[6] 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, see Fotheringham et al. (2002), Mei et al. (2006) and Geniaux and Napoléone (2008).

(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 **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 **R** against the residuals **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 **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

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

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 depend on the choice of weights, hence on the choice of bandwidth *h*. The optimal value for *h* is determined by minimizing the CV score, as mentioned above.

4. Hedonic imputation price indexes

This section addresses the issue of estimating quality-adjusted property price indexes.[7] Suppose that sample data is available for periods $t = 0,...,T$, where 0 is the base period (the starting period of the time series we want to construct), and suppose model (11) has been estimated separately for each period. The predicted property prices, obtained using MGWR, are given by $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + [\hat{\theta}^t + \sum_{a=1}^{A-1}\hat{\gamma}_a^t D_{ia}^t + \sum_{r=1}^{R-1}\hat{\lambda}_r^t D_{ir}^t]z_{iS}^t$. For short, we write the predicted price of structures, $\hat{\theta}^t + \sum_{a=1}^{A}\hat{\gamma}_a^t D_{ia}^t + \sum_{r=1}^{R}\hat{\lambda}_r^t D_{ir}^t$, as $\hat{\beta}_i^t$ and the predicted overall property price as $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t$ $(t = 0,...,T)$.

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

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \hat{p}_i^{t(0)}}{\sum_{i \in S^0} \hat{p}_i^0}, \qquad (18)$$

[7] In this paper we only discuss *sales-based* property price indexes. For an explanation of the difference between sales-based and stock-based indexes, see Eurostat et al. (2013).

Equation (18) may need some explanation. All quantities are set equal to 1 because each property is unique. Because the index is based on a single sample, it will not be affected by compositional change. Most, if not all, of the properties traded in period 0 are not re-sold in period *t*, and the 'missing prices' therefore need to be imputed by $\hat{p}_i^{t(0)}$. We have also replaced the observable base period prices p_i^0 by predicted prices $\hat{p}_i^0$, a method known as *double imputation*.[8]

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 $\hat{p}_i^{t(0)} = \hat{\alpha}_i^t z_{iL}^0 + [\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^0 + \sum_{r=1}^{R-1} \hat{\lambda}_r^t D_{ir}^0] z_{iS}^0$. For brevity, we use $\hat{\beta}_i^{t(0)}$ for the estimated constant-quality price of structures, $\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^0 + \sum_{r=1}^{R-1} \hat{\lambda}_r^t D_{ir}^0$.

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 (18) yields

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} [\hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0]}{\sum_{i \in S^0} [\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0]} = \hat{s}_L^0 \frac{\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0}{\sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0} + \hat{s}_S^0 \frac{\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0}{\sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0}, \tag{19}$$

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 (19) decomposes the overall house price index into structures and land components; the weights $\hat{s}_L^0 = \sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0 / \sum_{i \in S^0} [\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0]$ and $\hat{s}_S^0 = \sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0 / \sum_{i \in S^0} [\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0]$ 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 (19) 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 $\sum_{i \in S^0} \hat{s}_{iL}^0 (\hat{\alpha}_{iL}^t / \hat{\alpha}_{iL}^0)$, 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 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.

[8] Hill and Melser (2008) discuss different types of hedonic imputation indexes 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).

An alternative to the Laspeyres price index given by (19) is the hedonic double imputation Paasche price index, defined on the sample S^t of properties sold in period *t* $(t = 1,...,T)$:

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \hat{p}^t_i}{\sum_{i \in S^t} \hat{p}^{0(t)}_i}. \tag{20}$$

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 equation (20), the imputation Paasche index can be written as

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} [\hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS}]}{\sum_{i \in S^t} [\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}]} = \hat{s}^{t(0)}_L \frac{\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL}}{\sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}} + \hat{s}^{t(0)}_S \frac{\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS}}{\sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}}, \tag{21}$$

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 $\hat{s}^{t(0)}_L = \sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL} / \sum_{i \in S^t} [\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}]$ and $\hat{s}^{t(0)}_S = \sum_{i \in S^t} \hat{\beta}^0_i z^t_{iS} / \sum_{i \in S^t} [\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS}]$. 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

$$P^{0t}_{Fisher} = \left[P^{0t}_{Laspeyres} \times P^{0t}_{Paasche} \right]^{\frac{1}{2}} \tag{22}$$

does so by taking the geometric mean of the Laspeyres and Paasche price indexes. In the empirical section of the paper, we will estimate all three types of indexes. An exact decomposition of the Fisher index into structures and land components is not possible. Due to the fixed weights, the Laspeyres index and its decomposition are relatively easy to explain. So, even though we prefer the Fisher index, we are inclined to implement the Laspeyres index in statistical practice when the numerical differences are small.

5. Empirical evidence

5.1 The data set

The data set we will use 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 second quarter of 2008. 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 our data set during the ten-year period is 6,397, 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 characteristics for the structure, and characteristics for land. We included only three structural 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. After removing 44 observations with missing values, transaction prices below €10,000, more than 10 rooms, or ratios of plot size to structure size (usable floor space) larger than 10, we were left with 6,353 observations during the sample period.

Table A1 in the Appendix reports summary statistics by year for the numerical variables. The average transaction price significantly increased from 1998 to 2007 and then slightly decreased during the first half of 2008 (when the financial crisis started). 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.

5.2 Estimation results for hedonic models

Given the small size of the city of “A” and the relatively low number of observations, we decided to use annual data; in future work we will probably be using bi-annual data. Three normalized hedonic equations were estimated: model (8), which has no location characteristics at all (denoted as OLS in the tables and figures below), model (10) with 8 postcode dummy variables (OLSD), and model (12) with property-specific land prices

(MGWR). The last model was estimated by mixed geographically weighted regression using the software package GWR4.0.[9]

Considering that the property transactions are not evenly distributed across space, we used the adaptive bi-square function to construct the weighting scheme. In this case, the bandwidth is generally referred to as the window size, and its selection procedure is equivalent to the choice of the number of nearest neighbors. We derived the optimal bandwidth using the 'Golden Section Search' approach based on minimizing CV scores in a window-size range of 10% to 90%. There is a unique optimal window size for each annual sample in terms of prediction power; the CV scores indicated that it was around 10% for most of the years, except for 1998 (51%), 2001 (36%), and 2003 (29%). Yet, for the construction of price indexes, we would prefer a fixed window size for all years, especially since the number of sales is almost evenly spread across the whole period. So we have chosen a window size of 10% for every year, leading to 60 nearest neighbors that were used in the estimation of the MGWR models.

To compare the performance of the three property price 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 and is defined for MGWR models by[10]

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

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 3.2. The RMSE measures the variability of the absolute prediction errors of the models and is given by

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

The AICc and RMSE for each type of model are shown in Table 1. According to a rule of thumb mentioned by Fotheringham et al. (2002), if the difference in the AICc for two models is larger than 3, a significant difference exists in terms of performance. It can be seen that the OLSD model performs much better than the OLS model in all of the periods, which is not so surprising, and in turn that the MGWR model outperforms

[9] The software can be downloaded free of charge from https://geodacenter.asu.edu.

[10] The AICc expressions for the OLS and OLSD models may be found in e.g. Hurvich and Tsai (1989).

the OLSD model. The same ranking is found if the RMSE is used to assess the models. These results suggest that land prices indeed vary across space and that MGWR does a good job in estimating such nonstationarity.

Table 1: Model estimation and comparison

	OLS		OLSD				MGWR			
	AICc	RMSE	AICc	$dAIC_{10}$	RMSE	$dRMSE_{10}$	AICc	$dAIC_{21}$	RMSE	$dRMSE_{21}$
1998	6666.26	101.77	6629.82	-36.44	96.96	-4.81	6599.71	-30.11	91.18	-5.78
1999	7145.61	155.52	7110.61	-35.00	148.37	-7.15	7054.04	-56.57	136.98	-11.39
2000	7380.38	166.91	7342.49	-37.89	158.99	-7.92	7260.25	-82.24	143.44	-15.55
2001	7499.41	154.14	7450.69	-48.72	145.64	-8.50	7414.58	-36.11	137.21	-8.43
2002	7781.70	157.26	7714.56	-67.14	146.61	-10.65	7703.18	-11.38	141.05	-5.56
2003	7927.15	173.70	7877.13	-50.02	164.30	-9.40	7866.85	-10.28	158.42	-5.88
2004	7992.75	161.12	7970.64	-22.11	156.08	-5.04	7950.46	-20.18	149.29	-6.79
2005	8191.19	172.02	8124.69	-66.50	160.87	-11.15	8071.11	-53.58	149.51	-11.36
2006	8736.88	191.27	8717.40	-19.48	186.03	-5.24	8688.51	-28.89	176.35	-9.68
2007	9094.15	183.55	9068.06	-26.09	177.89	-5.66	9037.90	-30.16	168.44	-9.45

Note: $dAIC_{10}$ indicates the difference of AICc between OLS and OLSD, while $dAIC_{21}$ indicates the difference of AICc between OLSD model and MGWR; $dRMSE_{10}$ and $dRMSE_{21}$ have similar meanings.

Table 2: Summary statistics for estimated land prices from the MGWR model

	Min	Max	Median	Mean	Std deviation
1998	0.44	97.18	49.21	44.97	21.74
1999	76.13	195.23	146.48	141.70	30.23
2000	81.62	260.38	196.86	187.82	39.88
2001	89.05	227.31	182.73	173.00	34.00
2002	158.41	305.99	242.15	234.85	33.52
2003	55.03	196.84	142.73	133.84	37.21
2004	79.74	236.14	174.35	171.70	35.43
2005	109.47	276.13	187.90	181.74	36.08
2006	55.37	208.72	142.75	131.99	33.61
2007	58.54	223.41	166.35	164.82	31.86

Table 2 contains summary statistics for the price per square meter of land for the transacted properties, estimated using MGWR. The average estimated land price is quite volatile; the change over time differs greatly from that of the average transaction price of the properties (see Table A.1 in the Appendix). Following a sharp increase in 1999, the estimated average land price peaked in 2002, experienced a dramatic drop in 2003, and then increased again. The value in the starting year 1998 of approximately 45 euros per square meter of land is extremely low. This has a big impact on the corresponding land price indexes, as we will see in section 5.3.

As an illustration of the estimated hedonic models, the 2007 parameter estimates for the structure characteristics are given in Table 3. Note that almost all estimates differ significantly from zero at the 1% level. Dummy variables for dwellings built after 2000 and for detached houses were not included, and so the intercept term measures the price of structures per square meter of living space (in euros) for detached houses built after 2000. The estimated intercept for MGWR is rather high in comparison with OLSD. For each model, there is a clear tendency for structures to become less expensive as they are getting older. Also, detached dwellings are more expensive than other types of houses, which accords with a priori expectations.

Table 3: Parameter estimates for structural characteristics, 2007

	OLS	OLSD	MGWR
Intercept	1561.00**	1472.04**	1633.70**
	(46.93)	(55.59)	(75.35)
Building period:1960-1970	-367.23**	-310.09**	-411.55**
	(26.85)	(36.97)	(45.21)
Building period:1971-1980	-308.01**	-255.16**	-378.17**
	(24.19)	(35.68)	(44.86)
Building period:1981-1990	-230.45**	-178.98**	-259.74**
	(24.21)	(34.25)	(45.46)
Building period:1991-2000	-54.42*	-58.87*	-124.04**
	(22.41)	(27.87)	(38.28)
Terrace	-326.68**	-286.66**	-309.05**
	(35.80)	(36.78)	(42.04)
Corner	-303.89**	-280.98**	-278.44**
	(32.67)	(32.67)	(35.04)
Semidetached	-156.63**	-165.54**	-195.84**
	(49.37)	(49.85)	(52.39)
Duplex	-171.43**	-149.10**	-170.19**
	(31.49)	(31.63)	(33.94)

Note: Standard errors in brackets;** and * denote significance at the 1% and 5% level, respectively.

5.3 A comparison of different hedonic 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, and it will be interesting to see how they are affected by the choice of hedonic model (OLS, OLSD, or MGWR). We have estimated chained rather than direct indexes because imputing the 'missing prices' over a long period of time may not be useful and because the land and structures will be updated annually. A disadvantage of chaining is that the resulting indexes cannot be exactly decomposed since they are not consistent in aggregation.

In Figures 1-3, the estimated double imputation hedonic Laspeyres, Paasche and Fisher price indexes for the property as a whole are plotted, based on the three models. For each model, the (chained) Laspeyres index sits above the Paasche, as expected. The indexes based on OLSD and MGWR are almost the same; the differences can hardly be noticed in the graphs. So, for the house price index, the inclusion of a limited number of location dummy variables produces satisfactory results, despite the fact that the OLSD model performs not as good as MGWR. Not using location information at all makes a difference though: the Laspeyres and Paasche indexes from the OLS model seem to be biased downwards and upwards, respectively. The biases almost cancel out in the Fisher index, which is very similar to the Fisher indexes produced with the other two models.

Figure 1: Chained hedonic imputation Laspeyres house price index

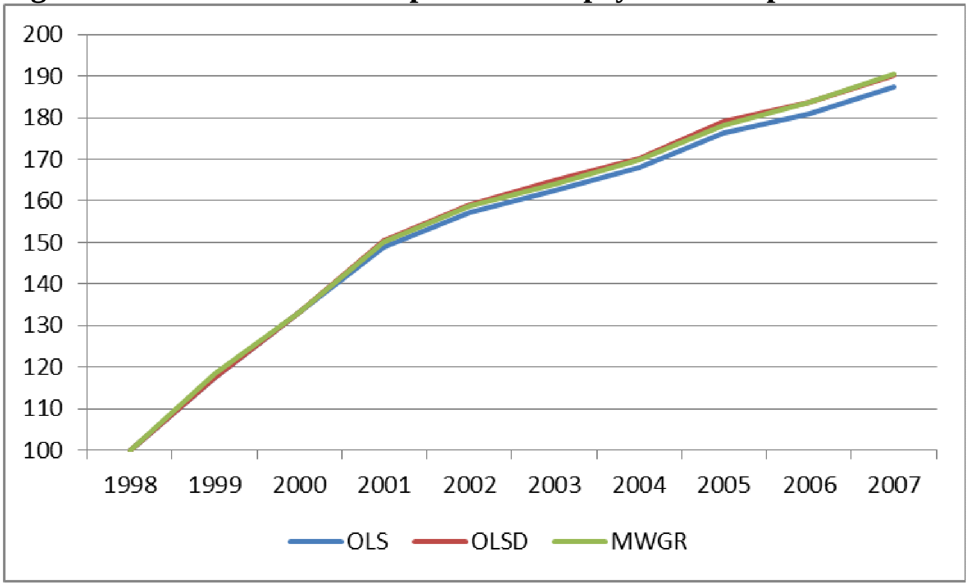


Figure 2: Chained hedonic imputation Paasche house price index

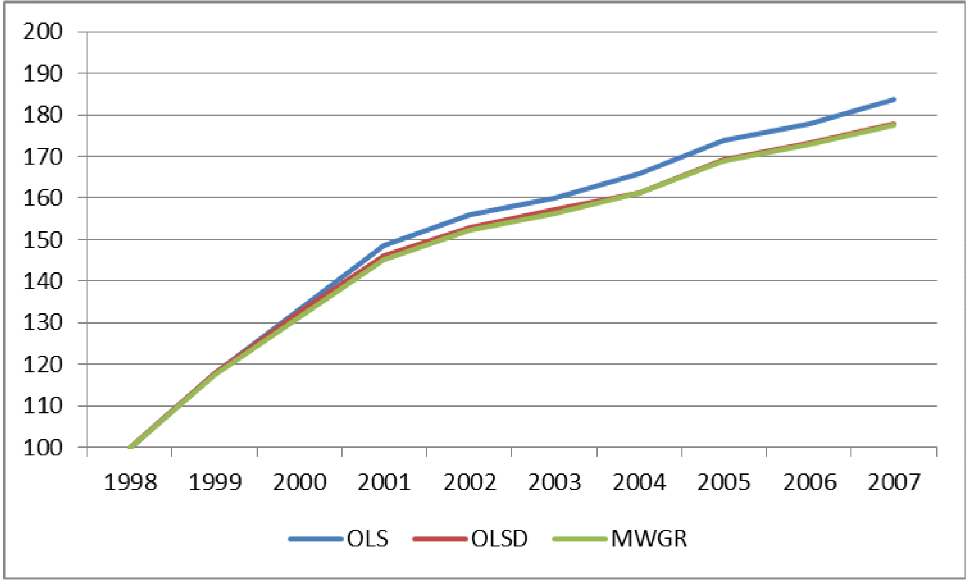
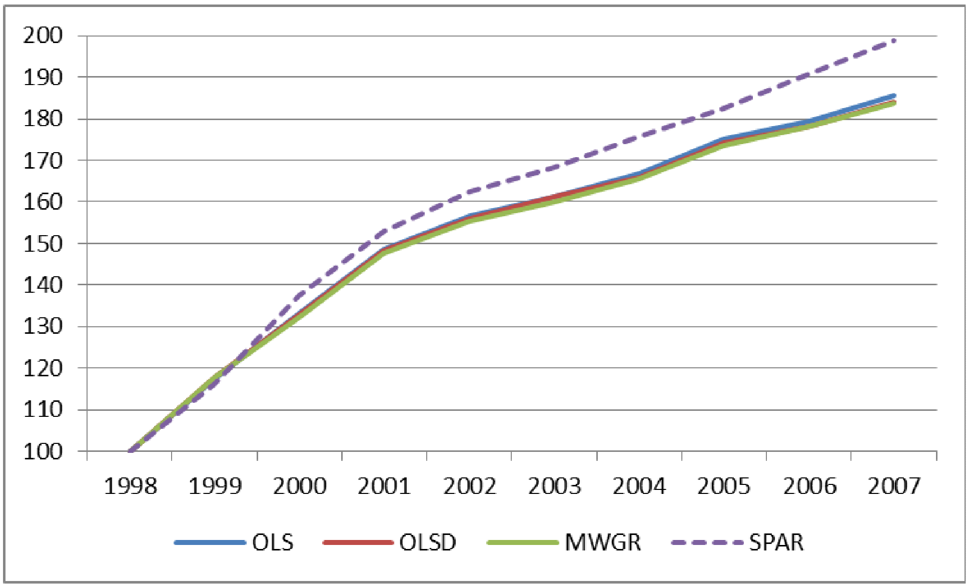


Figure 3: Chained hedonic imputation Fisher house price index and official SPAR index



We tentatively conclude that the double imputation Fisher house price index is insensitive to the treatment of location in the hedonic model. The official house price index for the Netherlands is also plotted in Figure 3.[11] Our hedonic indexes show a more modest price increase. There may be at least two reasons for this: house prices in the

[11] The official index is based on the Sale Price Appraisal Ratio (SPAR) method. For more information on this method, see de Haan et al. (2009) and de Vries et al. (2009).

city of “A” appreciated less compared to the rest of the country, or our indexes better adjust for quality changes. We think that the second reason is more important.

The picture changes when we look at the Fisher indexes for the price of land in Figure 4. The OLS- and OLSD-based indexes are similar, but the MWGR-based index behaves differently. For example, between 1998 and 1999 the MWGR-based index rises much faster than the other two indexes, and between 2005 and 2006 the MWGR-based index rises whereas the other two indexes fall. These results are surprising; for land in particular, we would expect the OLSD-based index to be similar to the MWGR-based index since both indexes explicitly account for location.

Figure 4: Chained hedonic imputation Fisher price indexes for land

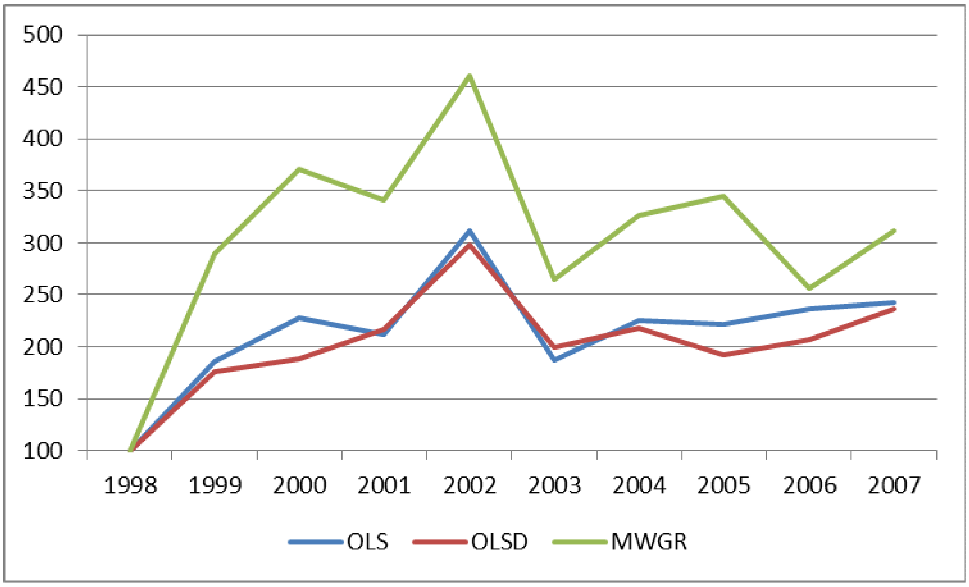
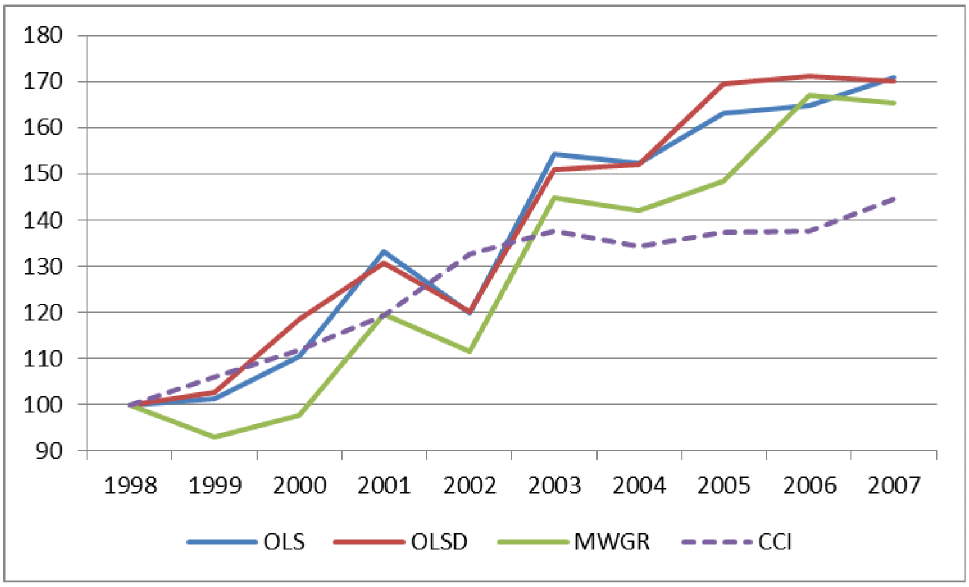


Figure 5 shows the hedonic imputation Fisher price indexes for structures based on the three models. While the differences cannot be ignored, they are less pronounced than the differences obtained for land. This is in accordance with a priori expectations: location should affect the price of land, and is modeled as such, but should leave the price of structures unaffected.

Figures 4 and 5 raise a number of issues. The first issue is whether the trends of the (Fisher) indexes for land and structures are plausible. For land, this will be difficult to check because information on the price change of land is currently unavailable for the Netherlands. For structures we use the nationwide official construction cost index (CCI) for dwellings, published by Statistics Netherlands, as a benchmark. This index, rebased to 1998=100, is also plotted in Figure 5. During the first half of the sample period, our price indexes for structures exhibit roughly the same trend as the construction cost index. During the second half of the sample period, the construction cost index flattens, but the structures price indexes keep rising. A construction cost index does not necessarily have to be identical to an implicitly derived price index for structures, and it may suffer from some measurement problems,[12] but this divergence is nevertheless puzzling.

Figure 5: Chained hedonic imputation Fisher price indexes for structures and official construction cost index



Importantly, the overall property price indexes are affected most by the changes in structures prices; the average estimated value share for structures across the sample period is 0.73 for the OLS and OLSD models, and 0.74 for MWGR. Figure 6 shows the OLSD-based estimates of the value shares for land and structures. The volatility of the shares in Figure 6, and also the volatility of the price indexes for land and structures in Figures 4 and 5, is striking. We would not expect the 'true' shares and price indexes to be very volatile. The volatility can be caused by problems such as the small number of observations, multicollinearity, heteroskedasticity, or outliers in the data. Of course, the small-number problem can only be circumvented by using data for a bigger city, which could also enable us to estimate bi-annual instead of annual indexes.

[12] The flattening of the construction cost index prior 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.

Figure 6: Estimates of value shares of land and structures, OLSD-based

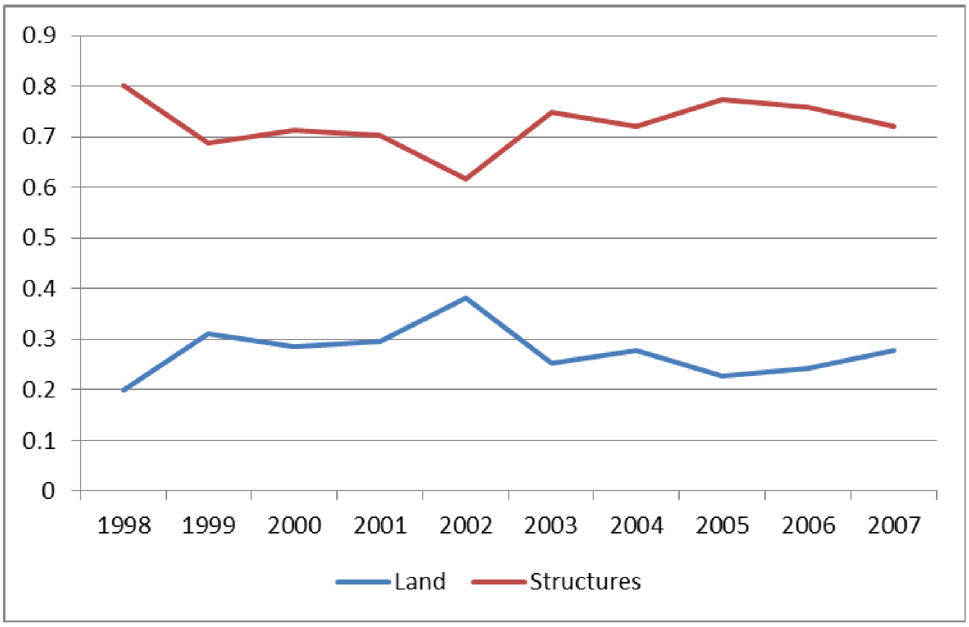
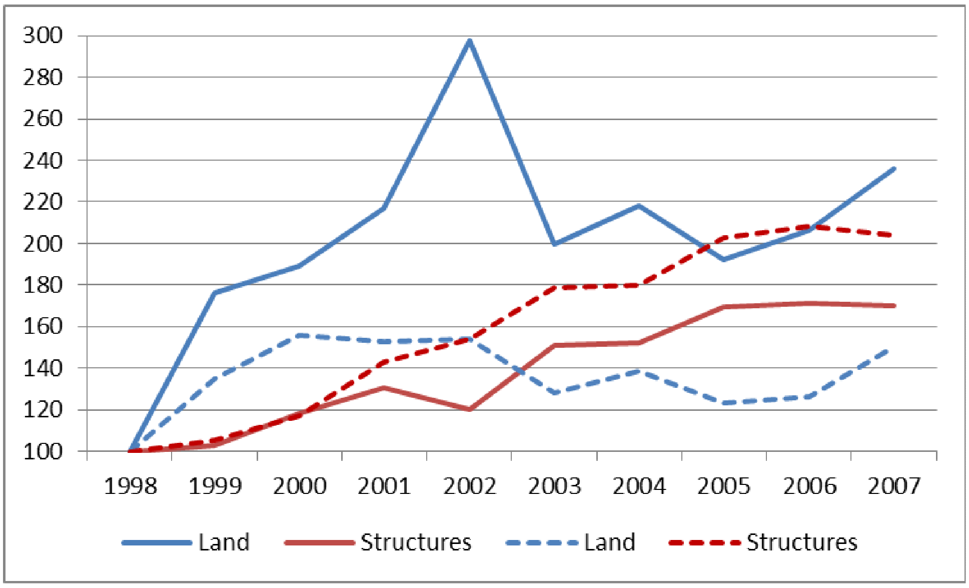


Figure 7: Chained Fisher price indexes for land and structures, OLSD-based



Multicollinearity was a big problem faced by Diewert et al. (2015) in estimating the builder’s model. It resulted in price changes for land and structures that consistently had opposite signs. In Figure 7, the OLSD-based Fisher indexes for land and structures from Figures 4 and 5 are copied. In some years, like in 2002, the price changes for land and structures have opposite signs, but in other years the price changes are in the same direction. We therefore suspect that multicollinearity is not the main issue involved. The variance inflation factor (VIF) for the estimated parameters for the ratio of plot size and structure size did not point to significant multicollinearity either.

The use of the property price per square meter of living space as the dependent variable in the models (i.e. the normalization) likely reduced multicollinearity, but it can have led to instability of the parameter estimates for land and structures if it resulted in 'classical' heteroskedasticity where the regression residuals grow with increasing ratios of plot size to structure size. For the OLS and OLSD models, the Breusch-Pagan test did indeed point to heteroskedasticity.[13] A related problem is the relatively small variation in the plot size to structure size ratios.

Scatterplots of the normalized prices against the plot size to structure size ratios showed some extreme outliers; most of them are in the higher ranges of the normalized prices and ratios. To check if deleting outliers would stabilize the indexes, we removed all observations with ratios of plot size to structure size larger than 5 (instead of 10), re-ran OLSD regressions and calculated chained double imputation price indexes again. The new OLSD-based Fisher indexes for land and structures are depicted by the dashed lines in Figure 7. Compared with the initial indexes the volatility is slightly reduced, but the trends have changed dramatically: the new structure price index sits above the old index and the new land price index sits far below the old one. This troubling result is touched upon in section 6 below.

6. Discussion and conclusions

Land is typically 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 using spatial econometrics. We linearized the 'builder's model' proposed by Diewert, de Haan and Hendriks (2015), allowed the price of land to vary at the individual property level, and estimated the model for the normalized property price (i.e., the price of the property per square meter of living space) by MGWR, a semi-parametric method, on annual data for

[13] Actually, we should have used a heteroskedasticity-consistent estimator for the standard errors in the OLS and OLSD models. Note that there is no formal heteroskedasticity test for the MWGR model.

the Dutch city of “A”. We then constructed chained imputation Laspeyres, Paasche and Fisher indexes and compared them with price indexes based on more restrictive models: a model with no variation in land prices and a model where land prices can vary across postcode areas, 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) also concluded that the use of geocoded data and spatial econometrics did not improve hedonic imputation house price indexes over models with postcode dummy variables.[14] This result is reassuring for statistical agencies that do not have the expertise or resources to use more sophisticated methods.

For some purposes, separate price indexes for land and structures are needed. As was demonstrated by Diewert, de Haan and Hendriks (2015), this can be a difficult task. A potential problem is multicollinearity, which arises because (in the ‘builder’s model’) the value of the property is split into the value of land and the value of structures: if the estimated price of land is too high, then the estimated price of structures will be too low, given plot and structure size. Probably due to the normalization of the property price, our estimates did not appear to suffer from severe multicollinearity.

Yet, our estimated price indexes for land and structures were very volatile. We can think of at least two reasons. First, the normalization of the property price resulted in heteroskedastic errors (and relatively little variation in the plot size to structure size ratios), leading to unstable coefficients and volatile indexes. Thus, although we reduced multicollinearity, at the same time we introduced heteroskedasticity.

The second reason for the volatility of the estimated land and structures indexes might be the linear relation postulated in our models between normalized property price and plot size to structure size ratio. Most likely, the ‘true’ relationship is nonlinear, and the linear specification produced outliers in the higher ranges. The misspecification was confirmed when we deleted all observations with plot size to structures size ratios larger than 5; the volatility of the land and structure price indexes from the OLSD model (with postcode dummy variables) was reduced somewhat but the trends changed significantly.

[14] Hill and Scholz (2014) treated location as a separate characteristic in their hedonic models in that they estimated property-specific shift terms for the overall property price rather than the price of land.

The probable cause is that the price of land is dependent on the size of the land plot: the price per square meter of land tends to fall with increasing plot size. Diewert, de Haan and Hendriks (2015) adjusted for this type of nonlinearity using linear splines to model the price of land. In future work we want to modify our models in the same spirit, either by using splines as well or by explicitly specifying some nonlinear function.

What worries us most is the extreme volatility of the MWGR-based indexes for land and structures. The MWGR method makes use of prices of neighboring properties, and since neighboring properties may be expected to have similar plot sizes, our results are unexpected and counterintuitive. We lack an explanation of this finding, but it does suggest that the semi-parametric MGWR approach produces inherently unstable results. Thus, while the MWGR model outperforms the other two models in terms of statistical criteria (AICc and RMSE) and produces a house price index that is very similar to the OLSD model, it aggravates instability and does not seem appropriate for estimating the land and structures components.

References

Bitter, C., G.F. Mulligan and S. Dall'erba (2007), "Incorporating Spatial Variation in Housing Attribute Prices: A Comparison of Geographically Weighted Regression and the Spatial Expansion Method", *Journal of Geographical Systems* 9, 7-27.

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

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton (1999), "Some Notes on Parametric Significance Tests for Geographically Weighted Regression", *Journal of Regional Science* 39, 497-524.

Casetti, E. (1972), "Generating Models by the Expansion Method": Applications to Geographic Research", *Geographic Analysis* 4, 81-91.

Diewert, W.E., S. Heravi and M. Silver (2009), "Hedonic Imputation versus Time Dummy Hedonic Indexes", pp. 87-116 in W.E. Diewert, J. Greenlees and C. Hulten (eds.), *Price Index Concepts and Measurement*, Studies in Income and Wealth, Vol. 70. Chicago: University of Chicago Press.

Diewert, W.E., J. de Haan and R. Hendriks (2011), "The Decomposition of a House Price Index into Land and Structures Components: A Hedonic Regression Approach", *The Valuation Journal* 6, 58-106.

Diewert, W.E., J. de Haan and R. Hendriks (2015), “Hedonic Regressions and the Decomposition of a House Price index into Land and Structure Components”, *Econometric Reviews* 34, 106-126. DOI: 10.1080/07474938.2014.944791.

Dorsey, R.E., H. Hu, W.J. Mayer, and H.C. Wang (2010), “Hedonic versus Repeat-Sales Housing Price Indexes for Measuring the Recent Boom-Bust Cycle”, *Journal of Housing Economics* 19, 75-93.

Eurostat, ILO, IMF, OECD, UNECE and World Bank (2013), *Handbook on Residential Property Price Indices*. Luxemburg: Publications Office of the European Union.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton (1998a), “Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis”, *Environment and Planning A* 30, 1905-1927.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton (1998b), “Scale Issues and Geographically Weighted Regression”, in N. Tate (ed.), *Scale Issues and GIS*. Chichester: Wiley.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton (2002), *Geographically Weighted Regression: the Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons.

de Haan, J. (2010), “Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and ‘Re-pricing’ Methods”, *Jahrbücher fur Nationalökonomie und Statistik* 230, 772-791.

de Haan, J., P. de Vries and E. van der Wal (2009), “The Measurement of House Prices: A Review of the Sale Price Appraisal Ratio Method”, *Journal of Economic and Social Measurement* 34, 51-86.

Geniaux, G. and C. Napoléone (2008), “Semi-parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditive Models”, pp. 101-127 in A. Baranzini jr., C. Schaerer and P. Thalmann (eds.), *Hedonic Methods in Housing Markets – Pricing Environmental Amenities and Segregation*. New York: Springer.

Hill, R.J. (2013), “Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy”, *Journal of Economic Surveys* 27, 879-914.

Hill, R.J. and D. Melser (2008), “Hedonic Imputation and the Price Index Problem: An Application to Housing”, *Economic Inquiry* 46, 593-609.

Hill, R.J., D. Melser, and I. Syed (2009), “Measuring a Boom and Bust: The Sydney Housing Market 2001-2006”, *Journal of Housing Economics* 18, 193-205.

Hill, R.J and M. Scholz (2014), “Incorporating Geospatial Data into House Price Indexes: A Hedonic Imputation Approach with Splines”, Graz Economics Paper 2014-05, Department of Economics, University of Graz.

Hurvich, C.M. and C.L. Tsai (1989), "Regression and Time Series Model Selection in Small samples", *Biometrika* 76, 297-307.

Jones, J.P. and E. Casetti (1992), *Applications of the Expansion Method*. London: Routledge.

Mei, C.L., N. Wang and W. X. Zhang (2006), "Testing the Importance of the Explanatory Variables in a Mixed Geographically Weighted Regression Model", *Environment and Planning A* 38, 587-598.

Pace, R.K., R. Barry, J.M. Clapp, and M. Rodriquez (1998), "Spatiotemporal Autoregressive Models of Neighborhood Effects", *Journal of Real Estate Finance and Economics* 17, 15-33.

Rambaldi, A.N. and D.S.P. Rao (2011), "Hedonic Predicted House Price Indices Using Time-Varying Hedonic Models with Spatial Autocorrelation", Discussion paper 432, School of Economics, University of Queensland.

Rambaldi, A.N. and D.S.P. Rao (2013), "Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes", Working paper WP04/2013, Centre for Efficiency and Productivity Analysis, School of Economics, University of Queensland.

Sun, H., Y. Tu, and S. Yu (2005), "A Spatio-Temporal Autoregressive Model for Multi-Unit Residential Market Analysis", *Journal of Real Estate Finance and Economics* 31, 155-187.

Tu, Y., S. Yu, and H. Sun (2004), "Transaction-Based Office Price Indexes: A Spatiotemporal Modelling Approach", *Real Estate Economics* 32, 297-328.

de Vries, P., J. de Haan, E. van der Wal and G. Mariën (2009), "A House Price Index Based on the SPAR Method", *Journal of Housing Economics* 18, 21-223.

Appendix

Table A 1: Summary statistics by year

	Total	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Obs.	6353	550	551	563	579	599	601	613	622	653	684	338
Transaction price (Euro)												
Mean	159654.54	96461.46	118282.01	132876.58	145231.68	151706.44	163434.80	175121.07	182076.49	191212.41	199038.99	196620.71
S.D.	74337.38	42692.67	54082.15	56393.94	58437.87	53463.20	64048.12	82963.12	70840.62	76253.88	83984.94	82259.77
Unit price (Euro/m^2)												
Mean	1249.28	748.42	934.22	1044.42	1172.44	1244.20	1286.49	1355.08	1423.38	1467.40	1520.72	1510.61
S.D.	380.79	217.29	282.93	288.05	298.26	291.71	288.65	297.94	298.21	324.48	350.24	341.84
Parcel size (m^2)												
Mean	255.99	243.78	262.41	249.28	244.72	239.37	255.78	262.63	255.93	264.42	275.15	258.76
S.D.	163.27	175.75	175.04	162.38	137.39	115.26	164.37	165.98	164.14	150.79	198.42	168.94
Floor space (m^2)												
Mean	126.15	126.43	125.36	126.72	123.41	122.01	125.83	126.56	126.29	128.57	128.49	128.14
S.D.	30.91	24.08	31.94	32.31	29.71	28.12	30.67	36.87	30.75	31.12	30.10	32.69
Ratio of parcel to floor space												
Mean	1.99	1.86	2.07	1.93	1.97	1.97	1.99	2.02	1.97	2.01	2.06	1.97
S.D.	0.90	0.95	1.08	0.86	0.87	0.80	0.90	0.85	0.84	0.79	1.03	0.92
XCoordinate												
Mean	233714.07	233973.13	234204.27	234185.17	233930.83	234003.15	233633.95	233484.33	233535.37	233224.28	233385.63	233324.00
S.D.	1810.91	1458.76	1426.00	1549.06	1728.41	1712.31	1794.27	1985.32	1930.81	1916.71	1946.71	2011.40
YCoordinate												
Mean	558584.51	558727.16	558799.54	558822.96	558648.40	558716.46	558521.05	558394.83	558548.00	558430.55	558404.22	558447.70
S.D.	1406.93	1441.34	1464.41	1430.78	1426.64	1411.69	1450.61	1414.46	1352.07	1322.21	1381.35	1240.58