DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 695–720, http://dx.doi.org/10.2478/JOS-2018-0033

Accounting for Spatial Variation of Land Prices in Hedonic Imputation House Price Indices: a Semi-Parametric Approach

Yunlong Gong[1,2] *and Jan de Haan*[2,3]

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 indices are constructed for a small city in the Netherlands and compared to price indices based on more restrictive models, using postcode dummy variables, or no location information at all. We find that, while taking spatial variation of land prices into account improves the model performance, the Fisher house price indices based on the different hedonic models are almost identical. The land and structures price indices, on the other hand, are sensitive to the treatment of location.

Key words: Geospatial information; hedonic modeling; land and structure prices; mixed geographically weighted regression; residential property

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

1. Introduction

The construction of house price indices is difficult because houses are traded infrequently and because properties are unique in terms of their location and structural characteristics. Hedonic regression and repeat sales methods both deal with these problems. The repeat sales method controls for location and unchanged structural characteristics as the prices of the 'same' properties are tracked over time (in a regression framework). However, this method suffers from several problems. For example, since only houses that are sold at least twice in the data set are used, it ignores single sales and is prone to sample selection bias. Also, the repeat sales method cannot provide information on the shadow prices of the property characteristics and thus does not allow the estimation of, for example, price

[1] Department of Land Resource Management, China University of Mining and Technology, Daxue Road 1, 221116 Xuzhou, China. Email: ylgong@cumt.edu.cn
[2] OTB-Research for the Built Environment, Delft University of Technology, Juliannalaan 134, 2628 BL Delft, The Netherlands.
[3] IT and Methodology Division, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands. Email: j.dehaan@cbs.nl
Acknowledgments: The authors would like to thank participants at the Economic Measurement Group Workshop, December 2014, Sydney, and at the Second International Conference of the Society for Economic Measurement, July 2015, Paris, for helpful comments on an earlier version of the article. Thanks are also given to two anonymous referees and the editor for critical review and constructive comments.

indices of the land the structure sits on. Given these problems with repeat sales methods, we focus on hedonic regression methods.

The hedonic regression method has its limitations as well. A general problem in the context of housing is omitted variables bias; it is not possible to include all the structural characteristics into the model, even if data on these characteristics were available (which is usually not the case). In addition, the true relationship between housing characteristics and house prices is unknown. The treatment of location is an important issue. One may for instance include locational variables in the model, such as distance to the city center and amenities. However, this is a rather data intensive method, and listing all the nodes of interest within the area is virtually impossible. Instead, researchers often include dummy variables at some aggregate level, such as postcode areas, to approximate the location effects. This is obviously a crude approach and could potentially lead to "location biases". In this article, we focus on the use of geospatial data, that is information on the exact location of properties in terms of geographic coordinates, to measure the effect of location.

Not properly accounting for location is likely to result in spatial correlation of house prices, which will impact on the precision of parameter estimates in hedonic models. Spatial correlation can be modeled in various ways, for instance via spatial lags or spatial errors, where a spatial weight matrix is designed to relate the feature of a point in space to the features of neighboring points. Such spatial econometric methods have been applied in time dummy hedonic models to estimate house price indices (Hill et al. 2009; Dorsey et al. 2010). Spatial error modelling has also been combined with state-space house price models which allow the parameters to follow a stochastic process along the time dimension; the price index can then be constructed through imputations (Rambaldi and Rao 2011, 2013). Others have directly extended the spatial filter by including time so that both spatial and temporal correlations are accounted for; these spatiotemporal autoregressive (STAR) models can generate a price index surface (Pace et al. 1998; Tu et al. 2004; Sun et al. 2005).

A disadvantage of the above methods is that the value of location and land is not explicitly modeled. For some purposes, like taxation and national accounting, being able to decompose the property value into land and structures values would be quite useful (Diewert et al. 2015; Rambaldi et al. 2015). In the present article, we attain this decomposition using a simplified version of the so-called builder's model (Diewert et al. 2011, 2015). We further assume that the value of location is capitalized into the price of land but not into the price of structures so that land prices are expected to vary across space whereas the price of structures is 'fixed'. The spatial variation of land prices is estimated by Geographically Weighted Regression (GWR), a nonparametric method proposed by Brunsdon et al. (1996) and Fotheringham et al. (1998b). Combining the land and structures components, we form a semi-parametric house price model and estimate it by Mixed Geographically Weighted Regression (MGWR). The (annual) house price index and its land and structures components are subsequently constructed in an imputation framework.

Our article tries to fill a gap in the *Handbook on Residential Property Price Indices* (Eurostat et al. 2013) in which the use of geospatial data to estimate hedonic house price models is not well covered. Geospatial data has been used before to estimate house price indices using a semi-parametric method. Clapp (2004), for example, estimated the value of location and overall property price change by Local Polynomial Regression (LPR). Our

work differs from Clapp's approach in a number of ways. The most important difference is that we incorporate the value of location into land prices and hence are able to construct a land price index, whereas Clapp treats it as an additive term to house value and thus cannot distinguish between land and structures values.

The article proceeds as follows. Section 2 outlines some basic ideas about the hedonic house price model that decomposes the property value into land and structures values and about the inclusion of additional structural characteristics into the model. Section 3 explains how we treat location; the GWR and MGWR approaches will be discussed in detail. Section 4 describes how we calculate the hedonic imputation indices. Section 5 presents empirical evidence for the Dutch city of "A" and discusses the results. Section 6 concludes and identifies some potential improvements.

2. A Simplification of the 'Builder's Model'

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 of the land $\left(\alpha^t z_{iL}^t\right)$, the value of the structure $\left(\beta^t z_{iS}^t\right)$ and a random error term u_i^t with zero mean:

$$p_i^t = \alpha^t z_{iL}^t + \beta^t z_{iS}^t + u_i^t. \tag{1}$$

The land and structure values are assumed to be proportional to the plot size z_{iL}^t and the size of living space z_{iS}^t, respectively. The shadow prices of land and structures in (1), α^t and β^t, are the same for all properties, irrespective of their location. In Section 3 we relax this assumption and allow for spatial variation in the price of land.

When applying Model (1) to the data of a sample S^t of properties sold in period t, a few problems arise. 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. To resolve these drawbacks, Equation (1) is divided by structure size z_{iS}^t, giving

$$p_i^{t^*} = \alpha^t r_i^t + \beta^t + \varepsilon_i^t, \tag{2}$$

where $p_i^{t^*} = p_i^t / z_{iS}^t$ is the *normalized property price*, that is, the value of the property per square meter of living space, $r_i^t = z_{iL}^t / z_{iS}^t$ is the ratio of plot size to structure size, and $\varepsilon_i^t = u_i^t / z_{iS}^t$. The model now has an intercept term and a single explanatory variable. In what follows, we focus on this normalized model.

2.2. Adding Structures Characteristics

Models (1) and (2) only incorporate structure size and plot size, which may lead to omitted variable bias. Here we discuss the inclusion of additional characteristics for the structures by linearizing the method proposed by Diewert et al. (2015).

We first consider the age effect and assume a straight-line depreciation model. The adjusted value of the structure is $\beta^t\left(1-\delta^t a_i^t\right)z_{iS}^t$, where δ^t is the depreciation rate and a_i^t is age of the structure. It is assumed that structure age is available in the data set as an ordinal (categorical) rather than continuous variable. Using multiplicative dummy variables D_{ia}^t that take on the value 1 if in period t property i belongs to age category a ($a = 1, \ldots, A$) and 0 otherwise, and after reparameterizing to eliminate the term $\beta^t z_{iS}^t$, the adjusted value of structure can be expressed as $\sum_{a=1}^{A}\gamma_a^t D_{ia}^t z_{iS}^t$, where γ_a^t represents the unit price of a structure belonging to age category a. While using discrete age may be somewhat problematic, it introduces some flexibility in that age dummies will not only reflect depreciation of structure but also capture vintage effect.

When incorporating another attribute, such as the number of rooms, the new value of the structures becomes $\beta^t\left(1-\delta^t a_i^t\right)\left(1+\mu^t z_{iM}^t\right)z_{iS}^t$, where μ^t is the parameter for the number of rooms z_{iM}^t. Using dummies D_{iM}^t for the number of rooms ($m = 1, \ldots, M$), and reparameterizing again, the new adjusted value of structure becomes $\sum_{a=1}^{A}\gamma_a^t D_{ia}^t z_{iS}^t + \sum_{m=1}^{M}\lambda_m^t D_{im}^t z_{iS}^t + \sum_{a=1}^{A}\sum_{m=1}^{M}\eta_{am}^t D_{ia}^t D_{im}^t z_{iS}^t$. To save degrees of freedom, we ignore the second-order interaction terms $D_{ia}^t D_{im}^t$ and obtain the *normalized* model

$$p_i^{t^*} = \theta^t + \alpha^t r_i^t + \sum_{a=1}^{A-1}\gamma_a^t D_{ia}^t + \sum_{m=1}^{M-1}\lambda_m^t D_{im}^t + \varepsilon_i^t. \tag{3}$$

In this model, an intercept term θ^t is included by excluding dummy variables for age class A and category M. For a property belonging to age class a ($a = 1, \ldots, A-1$) and category m ($m = 1, \ldots, M-1$) for number of rooms, the unit price of structures equals $\theta^t + \gamma_a^t + \lambda_m^t$. Additional categorical variables for the structures can be incorporated in a similar way.

3. Land and Spatial Heterogeneity

3.1. Location and the Price of Land

It is widely accepted that the value of location is capitalized into the price of land. In most empirical studies it is assumed that the price of land varies across postcode areas but is the same within each postcode area. An example is Diewert and Shimizu (2013) who estimated the 'builder's model' for Tokyo. Applying the same strategy to postcode dummy variables D_{ik} as was used in Subsection 2.2 for adding structure characteristics, an improved version of Model (3) for the *normalized* property price is

$$p_i^{t^*} = \theta^t + \sum_{k=1}^{K}\alpha_k^t D_{ik} r_i^t + \sum_{a=1}^{A-1}\gamma_a^t D_{ia}^t + \sum_{m=1}^{M-1}\lambda_m^t D_{im}^t + \varepsilon_i^t. \tag{4}$$

Each postcode area now has its own land price α_k^t. This might be still too crude, however, depending of course on the level of detail of the postcode system. A more general version of Model (4) is found by assuming that the price of land can differ at the individual property level, that is, at the micro location. We denote the property-specific land price

by α_i^t, yielding

$$p_i^{t^*} = \theta^t + \alpha_i^t r_i^t + \sum_{a=1}^{A-1} \gamma_a^t D_{ia}^t + \sum_{m=1}^{M-1} \lambda_m^t D_{im}^t + \varepsilon_i^t. \tag{5}$$

This model obviously cannot be estimated by standard regression techniques. In Subsection 3.2 below, we discuss a semi-parametric approach that enables us to estimate Model (5).

3.2. *Mixed Geographically Weighted Regression*

The parameters for the structures characteristics (θ^t, γ_a^t, and λ_m^t) in Model (5) are constant across space, whereas the land price $\left(\alpha_i^t\right)$ differs between properties. In other words, we account for spatial heterogeneity, or spatial nonstationarity as it is often referred to Brunsdon et al. (1996), of the price of land. One method that deals with spatial heterogeneity of parameters is the 'expansion method' (Casetti 1972; Jones and Casetti 1992). In our case, the price of land would 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 then be approximated using a Taylor-series expansion of some order; typically, second-order approximations are applied. Although the expansion method makes use of geospatial data, it is basically parametric because it calibrates a prespecified parametric model for the trend of land prices across space (Fotheringham et al. 1998a).

In this article, we adopt a truly nonparametric approach, namely *Geographically Weighted Regression* (GWR), to dealing with spatial heterogeneity of parameters (Brunsdon et al. 1996; Fotheringham et al. 1998b). Let us for a moment ignore the structures characteristics to explain how the property-based land prices can be obtained. Defining $\alpha_i = \alpha(x_i, y_i)$ and using matrix notation, Model (5) without structures characteristics can be written as

$$\mathbf{p}^* = \mathbf{r} \circ \boldsymbol{\alpha} + \boldsymbol{\eta}, \tag{6}$$

where $\mathbf{p}^* = (p_1^*, p_2^*, \cdots, p_n^*)^T$, $\mathbf{r} = (r_1, r_2, \cdots, r_n)^T$, $\boldsymbol{\alpha} = (\alpha(x_1, y_1), \alpha(x_2, y_2), \ldots, \alpha(x_n, y_n))^T$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)^T$, and $\circ$ is an operator that multiplies each element of $\boldsymbol{\alpha}$ by the corresponding element of $\mathbf{r}$. We have dropped the superscript t for convenience; it should be clear that we estimate models for each time period separately. In Model (6), the land price at point i is a realization of the continuous function $\alpha(x,y)$ at that point.

Model (6) can be estimated using a moving kernel window approach, which is essentially a form of Weighted Least Squares (WLS) regression. To obtain an estimate for the price of land $\alpha(x_i, y_i)$ for property i, a WLS regression is run on a subset of properties close to i on the premise that a property j which is closer to property i has a bigger influence in the estimation of $\alpha(x_i, y_i)$. That is

$$\alpha(x_i, y_i) = (\mathbf{r}^T \mathbf{w}(x_i, y_i)\mathbf{r})^{-1} \mathbf{r}^T \mathbf{w}(x_i, y_i)\mathbf{p}^*, \tag{7}$$

where $\mathbf{w}(x_i, y_i) = \text{diag}[w_1(x_i, y_i), w_2(x_i, y_i), \ldots, w_n(x_i, y_i)]$ is an n by n spatial weighting matrix. In this way, we are able to estimate land prices not only for observed properties,

but also for any imaginary location within the study area, enabling us to plot a continuous surface of land prices. The predicted values of the house prices are

$$\hat{\mathbf{p}}^* = \mathbf{r} \circ \boldsymbol{\alpha} = \mathbf{s}\mathbf{p}^*, \tag{8}$$

where the so-called hat matrix $\mathbf{s}$ is given by

$$\mathbf{s} = \begin{bmatrix} r_1\left(\mathbf{r}^T\mathbf{w}(x_1,y_1)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_1,y_1) \\ r_2\left(\mathbf{r}^T\mathbf{w}(x_2,y_2)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_2,y_2) \\ \vdots \\ r_n\left(\mathbf{r}^T\mathbf{w}(x_n,y_n)\mathbf{r}\right)^{-1}\mathbf{r}^T\mathbf{w}(x_n,y_n) \end{bmatrix}.$$

The weights w_{ij} $(i \neq j)$ should follow a monotonic decreasing function of distance 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

$$w_{ij} = \begin{cases} \left(1 - d_{ij}^2/h^2\right)^2 & \text{if } d_{ij} < h \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

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

$$CV(h) = \sum_{i=1}^{n} \left[p_i^* - \hat{p}_{\neq i}^*(h)\right]^2, \tag{10}$$

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

The above nonparametric GWR approach to dealing with spatial heterogeneity of land prices has to be extended by including structures characteristics with spatially fixed parameters, as shown in Model (5). 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. The estimation of the MGWR model is more complicated than that of the GWR model. To outline the estimation procedure, we write Model (5) in matrix notation as

$$\mathbf{p}^* = \mathbf{r} \circ \boldsymbol{\alpha} + \mathbf{D}_S\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{p}^*$, $\mathbf{r}$ and $\boldsymbol{\alpha}$ have the same meaning as in Equation (6), $\mathbf{D}_S$ is the matrix of structures characteristics included in Model (5), given by

$$\mathbf{D}_S = \begin{bmatrix} 1 & D^a_{11} & \cdots & D^a_{1,A-1} & D^m_{11} & \cdots & D^m_{1,M-1} \\ 1 & D^a_{21} & \cdots & D^a_{2,A-1} & D^m_{21} & \cdots & D^m_{2,M-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & D^a_{n1} & \cdots & D^a_{n,A-1} & D^m_{n1} & \cdots & D^m_{n,M-1} \end{bmatrix},$$

and $\boldsymbol{\beta} = (\theta, \gamma_1, \cdots, \gamma_{A-1}, \lambda_1, \cdots, \lambda_{M-1})^T$ is the vector of coefficients relating to $\mathbf{D}_S$ to be estimated.

We use the estimation method proposed by Fotheringham et al. (2002), which is less computationally intensive than the method described in Brunsdon et al. (1999). If the parameters $\boldsymbol{\beta}$ were known, the GWR approach (7) could be used to estimate $\boldsymbol{\alpha}$ by regressing $\mathbf{r}$ on $\mathbf{p}^* - \mathbf{D}_S\boldsymbol{\beta}$. Similarly, OLS estimates of $\boldsymbol{\beta}$ could be obtained by regressing $\mathbf{D}_S$ on $\mathbf{p}^* - \mathbf{r} \circ \boldsymbol{\alpha}$ if the property-based parameters $\boldsymbol{\alpha}$ were known. In practice, a four-step estimation procedure is followed; for details, see Fotheringham et al. (2002), Mei et al. (2006) and Geniaux and Napoléone (2008). This four-step procedure involves:

(1) regressing each column of $\mathbf{D}_S$ against $\mathbf{r}$ using the GWR approach described by (7) and then computing the residuals $\mathbf{Q} = (\mathbf{I} - \mathbf{s})\mathbf{D}_S$,
(2) regressing the dependent variable $\mathbf{p}^*$ against $\mathbf{r}$ using the GWR approach (7) 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 $\mathbf{D}_S\hat{\boldsymbol{\beta}}$ from $\mathbf{p}^*$ and regressing this part against $\mathbf{r}$ using GWR approach in (7) to obtain estimates $\hat{\alpha}(x_i, y_i) = [\mathbf{r}^T\mathbf{w}(x_i, y_i)\mathbf{r}]^{-1}\mathbf{r}^T\mathbf{w}(x_i, y_i)(\mathbf{p}^* - \mathbf{D}_S\hat{\boldsymbol{\beta}})$.

The predicted values for the property prices in Equation (11) can be expressed as

$$\hat{\mathbf{p}}^* = \mathbf{s}(\mathbf{p}^* - \mathbf{D}_S\hat{\boldsymbol{\beta}}) + \mathbf{D}_S\hat{\boldsymbol{\beta}} = \mathbf{L}\mathbf{p}^*, \tag{12}$$

with $\mathbf{L} = \mathbf{s} + (\mathbf{I} - \mathbf{s})\mathbf{D}_S\left[\mathbf{D}_S^T(\mathbf{I} - \mathbf{s})^T(\mathbf{I} - \mathbf{s})\mathbf{D}_S\right]^{-1}\mathbf{D}_S^T(\mathbf{I} - \mathbf{s})^T(\mathbf{I} - \mathbf{s})$, which is the hat matrix for Equation (11).

As discussed above, 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 statistic given by (10). In the case of MGWR, the CV statistic is equivalent to (Mei et al. 2006)

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_i^* - \hat{p}_i^*(h)}{1 - l_{ii}(h)}\right]^2, \tag{13}$$

where $\hat{p}_i^*(h)$ is the predicted price for property i and $l_{ii}(h)$ is the ith diagonal element of matrix $\mathbf{L}$ in Equation (12).

4. Hedonic Imputation Price Indices

This section addresses the issue of estimating quality-adjusted property price indices. Suppose sample data is available for periods $t = 0, \ldots, T$, where 0 is the base period (the starting period of the time series we want to construct), and suppose Model (5) has been estimated separately for each period. The predicted property prices are given by $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \left[\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^t\right] z_{iS}^t$. For short, we write the predicted unit price of structures, $\hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^t + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^t$, as $\hat{\beta}_i^t$ and the predicted overall property price as $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t$ $(t = 0, \ldots, T)$.

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

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \hat{p}_i^{t(0)}}{\sum_{i \in S^0} \hat{p}_i^0}, \tag{14}$$

Equation (14) 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 resold in period t, and the 'missing prices' have to be imputed by $\hat{p}_i^{t(0)}$. We also replaced the observed base period prices p_i^0 by the predicted values $\hat{p}_i^0$, a method known as *double imputation*. Hill and Melser (2008) discussed different types of hedonic imputation methods in the context of housing. For a general discussion of the difference between hedonic imputation indices and time dummy indices, see Diewert et al. (2009) and de Haan (2010).

The $\hat{p}_i^{t(0)}$ are estimated period t constant-quality property prices, that is, 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 $\hat{\beta}_i^{t(0)} = \hat{\theta}^t + \sum_{a=1}^{A-1} \hat{\gamma}_a^t D_{ia}^0 + \sum_{m=1}^{M-1} \hat{\lambda}_m^t D_{im}^0$ 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 (14) yields

$$P_{Laspeyres}^{0t} = \frac{\sum_{i \in S^0} \left[\hat{\alpha}_i^t z_{iL}^0 + \hat{\beta}_i^{t(0)} z_{iS}^0\right]}{\sum_{i \in S^0} \left[\hat{\alpha}_i^0 z_{iL}^0 + \hat{\beta}_i^0 z_{iS}^0\right]} = \hat{s}_L^0 \frac{\sum_{i \in S^0} \hat{\alpha}_i^t z_{iL}^0}{\sum_{i \in S^0} \hat{\alpha}_i^0 z_{iL}^0} + \hat{s}_S^0 \frac{\sum_{i \in S^0} \hat{\beta}_i^{t(0)} z_{iS}^0}{\sum_{i \in S^0} \hat{\beta}_i^0 z_{iS}^0}, \tag{15}$$

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 (15) 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.

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 ($t = 1, \ldots, T$):

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \hat{p}^t_i}{\sum_{i \in S^t} \hat{p}^{0(t)}_i}. \tag{16}$$

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 $\hat{\beta}^{0(t)}_i = \hat{\theta}^0 + \sum_{a=1}^{A-1} \hat{\gamma}^0_a D^t_{ia} + \sum_{m=1}^{M-1} \hat{\lambda}^0_m D^t_{im}$ 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 (16), the hedonic imputation Paasche index can be written as

$$P^{0t}_{Paasche} = \frac{\sum_{i \in S^t} \left[\hat{\alpha}^t_i z^t_{iL} + \hat{\beta}^t_i z^t_{iS} \right]}{\sum_{i \in S^t} \left[\hat{\alpha}^0_i z^t_{iL} + \hat{\beta}^{0(t)}_i z^t_{iS} \right]} = \hat{s}^{t(0)}_L \frac{\sum_{i \in S^t} \hat{\alpha}^t_i z^t_{iL}}{\sum_{i \in S^t} \hat{\alpha}^0_i z^t_{iL}} + \hat{s}^{t(0)}_S \frac{\sum_{i \in S^t} \hat{\beta}^t_i z^t_{iS}}{\sum_{i \in S^t} \hat{\beta}^{0(t)}_i z^t_{iS}}, \tag{17}$$

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 indices 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 indices 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

$$P^{0t}_{Fisher} = \left[P^{0t}_{Laspeyres} \times P^{0t}_{Paasche} \right]^{\frac{1}{2}} \tag{18}$$

does so by taking the geometric mean of the Laspeyres and Paasche price indices. The Fisher index formula is not consistent in aggregation, which means that decomposing the Fisher property index into structures and land components like Equation (15) and (17) is not possible. In other words, the Fisher property index can only be derived directly from house price relatives, but not from aggregating the Fisher structures index and land index, whereas the Laspeyres and Paasche indices can be obtained in both ways.

Double imputation Laspeyres, Paasche, and Fisher property price indices and the land price indices based on the more restrictive hedonic Models (4) or (3) are found by replacing $\hat{\alpha}^0_i$ and $\hat{\alpha}^t_i$ in (15) and (17) by the corresponding postcode-specific estimates $\hat{\alpha}^0_k$ and $\hat{\alpha}^t_k$ 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.

5. Empirical Evidence

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 excluded sales of condominiums and apartments as 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 per cent of all residential property transactions in "A".

Our data set contains information on time of sale, transaction price, a range of structures characteristics, and land characteristics. We included only three structures characteristics in our models, that is, usable floor space, type of house and building period. Note that, because a sample period of ten years is relatively short and building period is available in decades only, we decided to use building period in the models rather than approximate age of the structures (from building period in decades and time of sale). For land, we used plot size and postcode or latitude/longitude. We deleted 43 observations with missing values or prices below EUR 10,000, properties with more than ten rooms and those with ratios of plot size to structure size (usable floor space) larger than ten 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. The average transaction price and the price per square meter of floor space increased significantly 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.

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 biannual or quarterly data. We estimated three normalized hedonic models: Model (3), which does not include location and has a fixed land price across the city (denoted by FLP), Model (4) with nine postcode dummy variables, hence with postcode-varying land prices (PCLP), and Model (5) with location-varying land prices (LLP).

The FLP and PCLP models were estimated by OLS, while the MGWR approach described in Subsection 3.2 was used to estimate the LLP model. When applying the MGWR approach, a key point is the selection of the bandwidth in Equation (9) to decide which neighboring transactions will be used in the estimation of the land price for a specific property. Given that the transactions in our data set are not evenly distributed across space, using transactions within a certain distance may not be good practice: properties located in the densely-populated area will have many neighbors while other

properties will have only few. We therefore constructed the weighting scheme using the *adaptive bi-square* function where the bandwidth relates to a fixed number N of nearest neighbors that are used in the estimation process. When computing the weights given by Equation (9), h equals the distance to the Nth nearest neighbor and changes with the target properties. In practice, the choice of N nearest neighbors is equivalent to the choice of window size, that is the fraction of the sample used. To find the optimal value, we varied the window size from ten per cent to 95 per cent using a five per cent interval and selected the size that yielded the lowest CV score as given by Equation (13). Each annual sample then has a unique optimal window size. The CV scores indicated that a ten per cent window size was preferred for most of the years, except for 1999, 2000, and 2002, with an optimal size of 15 per cent, and for 2003, with an optimal size of 30 per cent. However, for the construction of price indices we prefer the same window size for all years, in particular because the number of sales is almost evenly spread across the whole period. So we chose a window size of ten per cent for each year, leading to 60 nearest neighbors that were used in the estimation of the LLP models.

As an illustration, Table 1 contains the 2007 parameter estimates for the structures characteristics. Almost all of the estimates differ significantly from zero at the one per cent level. To some extent they vary across the different models. For example, the FLP intercept term is relatively high compared to the PCLP and LLP intercepts. 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 older.

Table 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 approach is that it allows us to plot a continuous map with estimated prices of land per square meter. To produce the map, we first divided the city of "A" into 50 (meters) $\times$ 50 (meters) grids and retrieved the coordinates of each cell, and then estimated the unit land price of each grid based on their coordinates. For the year 2007, such a map is depicted in Figure 1, where the land prices were rescaled to the range [0, 1]. The postcode areas are indicated as well. While the spatial pattern in Figure 1 is largely consistent with the pattern found using the PCLP model (shown in Figure A1 in the appendix), the land prices estimates from the LLP model do vary within some of the postcode areas. This suggests that the use of postcode dummies, as in the PCLP 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 FLP and PCLP models can be found in

Table 1. Parameter estimates for structures characteristics, 2007.

	FLP	PCLP	LLP
Intercept	1480.70*** (46.93)	1405.41*** (53.71)	1395.76*** (57.51)
Building period: 1960–1970	−370.48*** (25.94)	−389.50*** (36.67)	−398.40*** (41.75)
Building period: 1971–1980	−311.17*** (23.36)	−261.50*** (33.96)	−323.50*** (41.69)
Building period: 1981–1990	−232.93*** (23.37)	−173.08*** (32.59)	−226.14*** (42.87)
Building period: 1991–2000	−58.64*** (21.64)	−49.34* (26.55)	−115.13*** (37.26)
Terrace	−285.65*** (35.17)	−264.34*** (35.24)	−187.28*** (37.32)
Corner	−281.36*** (31.77)	−274.54*** (31.18)	−192.85*** (34.07)
Semidetached	−122.89** (47.96)	−149.50*** (47.57)	−96.93** (48.73)
Duplex	−151.08*** (30.60)	−147.24*** (30.17)	−104.56*** (31.03)

Notes: Model FLP and PCLP are estimated by OLS, while model LLP is estimated using the MGWR approach. Standard errors are reported in parentheses; ***, ** and * denote significance at the 1%, 5% and 10% level, respectively.

Table 2. Summary statistics for estimated land prices.

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

Notes: For FLP, the land price estimates are reported. For PCLP, 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 LLP, the columns provide summary statistics for the land price estimates of all transacted properties.

Hurvich and Tsai (1989); for the LLP model, it is defined by

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

where $\hat{\sigma}$ is the estimated standard deviation of the error term and $tr(\mathbf{L})$ the trace of the hat matrix described in Subsection 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 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 three, a significant difference exists in their performance (Fotheringham et al. 2002). It can be seen that the PCLP model performs much better than the FLP model in all years, as we would expect, and in turn that the LLP model outperforms the PCLP 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 between and within postcode areas.

Although LLP is obviously better suited to model the variation of land prices and to predict property prices, the PCLP model does a good job too. In several years, 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 LLP model does. This does not come as a surprise though, given that the MGWR approach used for estimating the land price of a particular property in the LLP model utilizes the information of neighboring properties, most of which are likely to be located in the same postcode area.

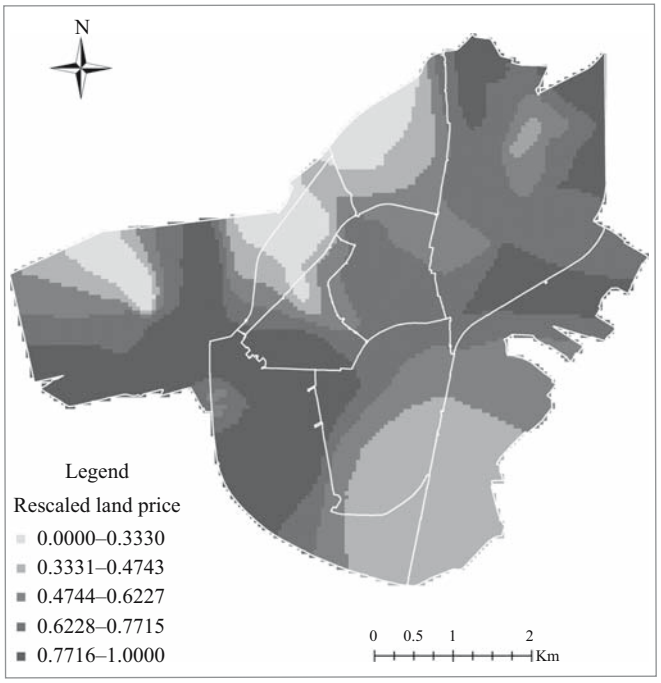

Fig. 1. Price of land per square meter, 2007.

5.3. *Hedonic Imputation Price Indices*

Changes in average property prices, and changes in their land and structure components, are affected by compositional change in the traded properties. Our hedonic house price indices and the land and structures components control for this. We estimated chained rather than direct price indices because the value shares of the land and structures will then be updated at an annual frequency. A drawback of chaining is that the resulting indices cannot be exactly decomposed because they are not consistent in aggregation.

In Figures 2–4, the estimated double imputation hedonic Laspeyres, Paasche, and Fisher price indices for the overall property are plotted, based on the three models (FLP, PCLP, and LLP). A comparison of Figures 2 and 3 shows that, for each model, the chained Laspeyres index sits above the Paasche index, as expected. The Laspeyres and Paasche indices based on PCLP and LLP are very similar; for the Laspeyres index, the difference can even hardly be noticed. This result is in accordance with our previous finding that the PCLP model captures the spatial variation of land prices reasonably well.

Not using location information at all does make a difference, at least for the Laspeyres and Paasche house price indices. The FLP-based Laspeyres and Paasche indices seem to be biased downwards and upwards, respectively. However, the biases almost cancel out in the Fisher indices, as can be seen in Figure 4: the FLP-based Fisher index is very similar to the PCLP-based and LLP-based Fisher indices. 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.

Table 3. Model comparison.

	FLP		PCLP				LLP					
	AICc	RMSE	AICc	ΔAIC_{PF}	RMSE	$\Delta RMSE_{PF}$	AICc	ΔAIC_{LP}	ΔAIC_{LF}	RMSE	$\Delta RMSE_{LP}$	$\Delta RMSE_{LF}$
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

Note: ΔAIC_{PF} is equal to AICc for PCLP minus AICc for FLP; ΔAIC_{LP} and ΔAIC_{LF} are equal to AICc for LLP minus AICc for PCLP and FLP, respectively; $\Delta RMSE_{PF}$, $\Delta RMSE_{LP}$ and $\Delta RMSE_{LF}$ have a similar meaning.

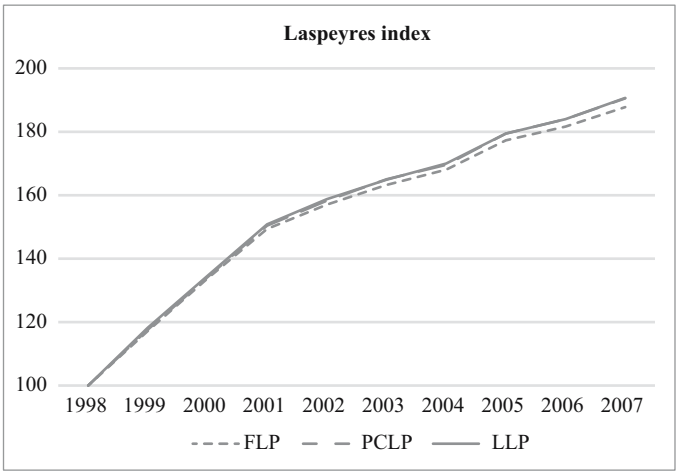

Fig. 2. Hedonic imputation Laspeyres house price index.

Figure 5 plots the Fisher price indices for land. The PCLP- and LLP-based indices, which explicitly account for location, are similar, though the LLP-based index is less volatile, at least during 2003–2007. The FLP-based index seems to be significantly upward biased. For example, between 1999 and 2000 as well as between 2003 and 2005, the FLP-based index rises much faster than the other two indices. A possible explanation is the following. Suppose specific locational attributes improved over time or that consumers' preferences changed towards locations with specific characteristics. This will have caused land prices in some areas to appreciate significantly relative to other areas. If, as in the FLP model, such heterogeneity is not accounted for, bias in the average estimated land price is likely to occur. The treatment of location in the FLP model may not only have produced biased levels of land prices, it might easily have led to a biased trend as well.

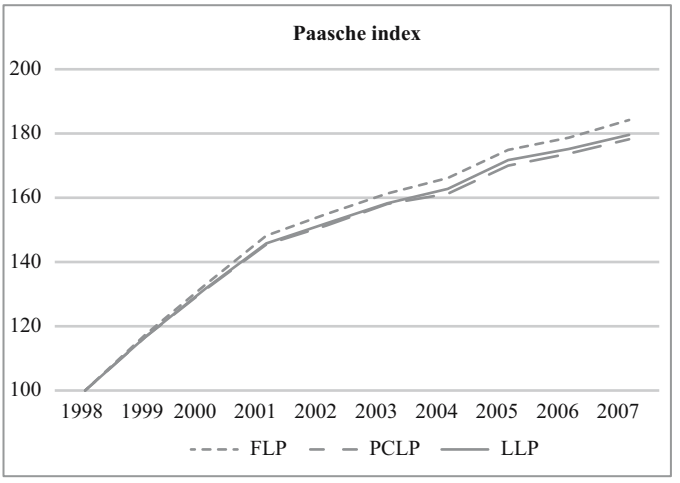

Fig. 3. Hedonic imputation Paasche house price index.

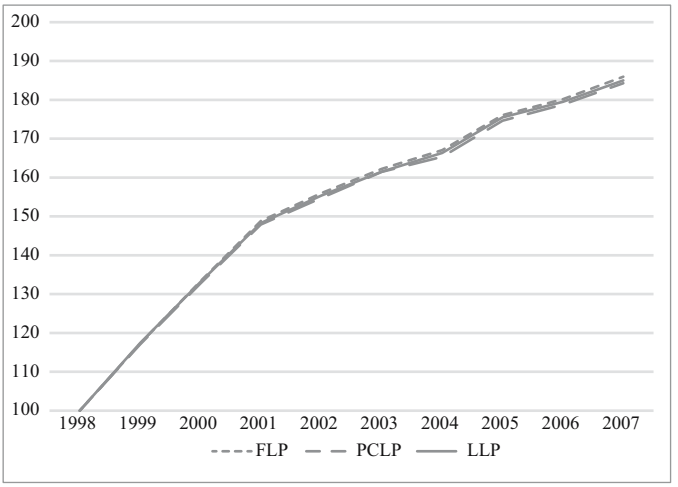

Fig. 4. Hedonic imputation Fisher house price index.

Figure 6 shows the Fisher price indices for structures based on the three models. Again, the PCLP-based and LLP-based indices are similar. The FLP-based index sits below these two indices, which is not surprising given the above results for land. Since the hedonic model in this paper leaves out many structural characteristics, which may be correlated with location, the decomposition of house price index is not strictly orthogonal. In this sense, upward bias in estimated land prices using the FLP model is therefore likely to result in downward bias in structures prices.

Figure 7 shows the LLP-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 FLP- and PCLP-based shares exhibit similar patterns and levels; the value

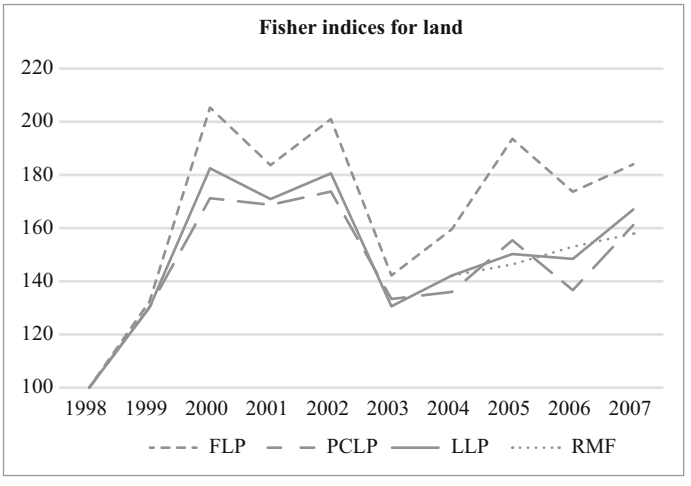

Fig. 5. Hedonic imputation Fisher price indices for land.

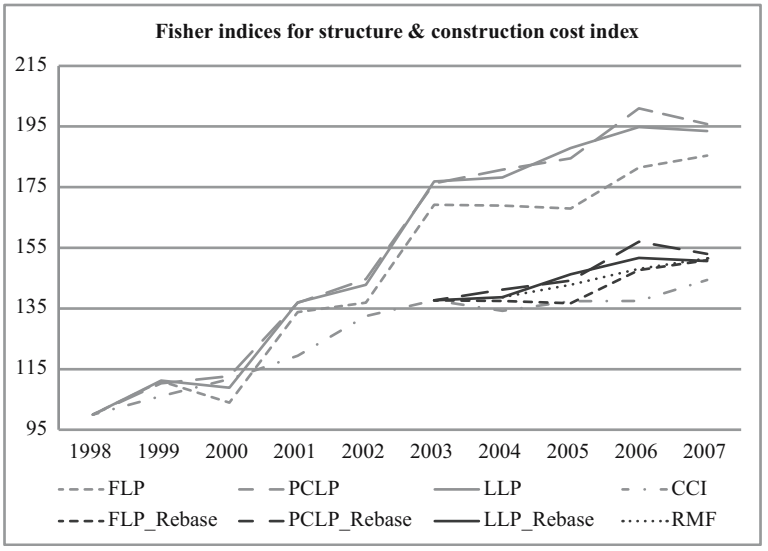

Fig. 6. Hedonic imputation Fisher price indices for structures and official construction cost index.

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, overall house price indices are affected most by changes in structures prices. Yet, combining Figures 4, 5, 6, and 7 suggests that the increase in house prices between 1998 and 2001 was driven mainly by the increase of land prices: both the (average) price of land and its value share show a sharp increase.

5.4. Discussion

Figures 5, 6, and 7 raise a number of issues. The first issue is the volatility of the land and structures price indices. Volatile series can be expected with sparse data (without

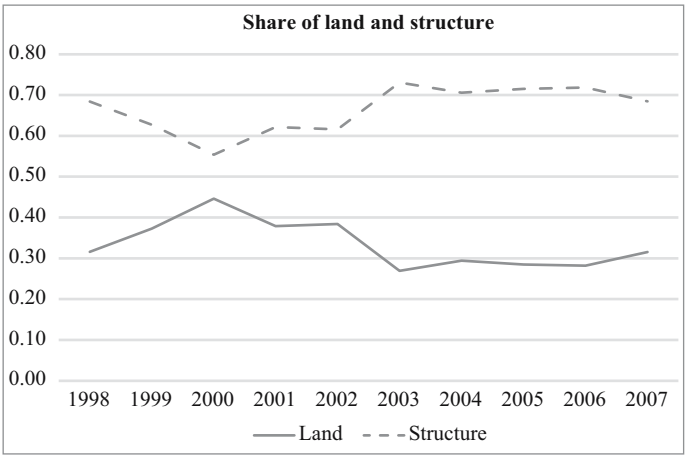

Fig. 7. Estimated value shares of land and structures, LLP model.

smoothing). Another potential cause is 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 type of multicollinearity, they incorporated exogenous information in the hedonic models; see also (e.g., Diewert et al. 2009; Diewert and Shimizu 2013; Francke and van de Minne 2017). More specifically, their (final) models did not endogenously determine a price index of structures but used the published construction cost index as the measure of structures price change. We did not follow their approach for two reasons: an endogenously estimated trend in the price of structures does not necessarily have to be equal to that of construction costs, and multicollinearity does not seem to be the most important issue.

In Figure 8, the LLP-based Fisher price indices for land and structures from Figures 5 and 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.

The second issue is whether the trends of the (Fisher) price indices for land and structures are plausible. For land, this can hardly be checked since information on the price change of land covering our sample period is not available for the Netherlands. Rambaldi et al. (2015), using an unobserved component approach, estimated an endogenous monthly land price index for the city of "A" from August 2003 to June 2008, denoted by RMF index. We converted their series into an annual series by averaging the monthly indices, rebased the resulting index to 2004, and then spliced it on to the LLP land price index for 2004 (see Figure 5). Our LLP hedonic land price index in 2005, 2006, and 2007 is very similar to the RMF index, which is reassuring, except that the latter index is smoother.

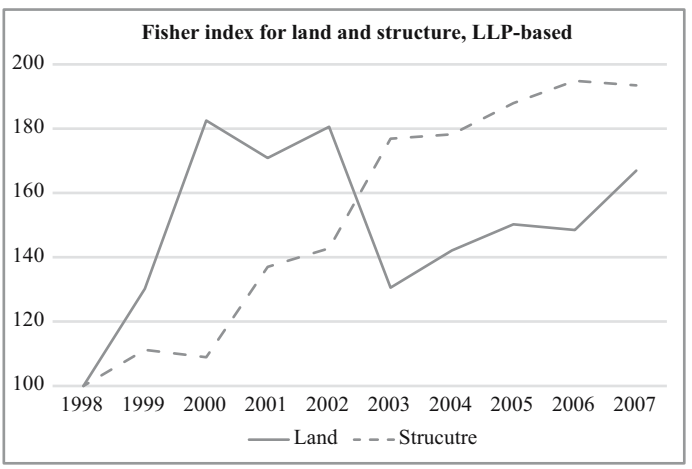

Fig. 8. Chained Fisher price indices for land and structures, LLP model.

For structures we use the nationwide construction cost index (CCI) for new dwellings published by Statistics Netherlands as a benchmark. This price index, rebased to 1998=100, is shown in Figure 6 as well. Our hedonic structures price indices appear to rise much faster than the construction cost index. As mentioned above, a construction cost index does not necessarily have to coincide with an implicit price index for structures derived from a hedonic model. In a competitive market, where developers also have sufficient time to meet demand, construction cost is believed to be equal to the market value of the structure (Davis and Heathcote 2007; Davis and Palumbo 2008). However, in reality the market is characterized by restrictions on new construction and high costs of replacing old structures by new ones. In this case, a markup on construction costs can be expected. During a housing boom, like our study period, the mark up may well be growing over time. Kuminoff and Pope (2013), who estimated land values for US metropolitan areas using a hedonic approach, indeed found that in some (though not all) areas the increase in the market value of the structures exceeded the increase in replacement costs in the booming period.

Omitted variables bias, resulting in quality-change bias, may have played a role as well. We included only a few structures characteristics in the hedonic models. Unless they would be highly collinear with included variables, adding characteristics will lead to better quality adjustment for structures and lower the hedonic price indices for structures if the average quality of structures improved over time.

Importantly, the major part of the differences between our hedonic indices and the construction cost index stems from a big increase in our indices in 2003; as of 2003, the deviation is relatively small. We reproduced the RMF price index for structures estimated by Rambaldi et al. (2015) in Figure 6 and, as was the case for land, their index is very similar to our LLP structures price index in 2005, 2006, and 2007.

The sudden increase in estimated structures prices and drop in estimated land prices in 2003 are worth examining in more detail. At first glance, sample selection bias might matter, for example if the spatial distribution of transacted properties in 2003 was very different from that in other years, or if unique properties, like properties with a very large of plot size to structure size ratios, were transacted in 2003. However, after a careful check of the data, we exclude this possibility. It could be that the 2003 results are "real" in the sense that a shock affected households' decision-making in the Dutch housing market or perhaps in the local market of "A". This is quite plausible given that the house price appreciation rate suddenly dropped from above ten per cent to around four per cent at the time around 2002 or 2003. But it is not clear to us what that shock might have been.

The third issue concerns the low share of land in the value of properties sold, which was estimated at roughly one third across the sample period. Rambaldi et al. (2015) estimated the land value share for the city of "A" during the period 2003–2008 between 0.30 and 0.40. van de Minne and Francke (2012) estimated the share of land for properties (excluding apartments/condominiums) sold during 2003–2010 in the Dutch city of 's Hertogenbosch at 0.39 on average. In a follow-up study (Francke and van de Minne 2017), 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, and we

expect this to have a downward effect on the price of land but not on the price of structures, hence on the value share of land.

De Groot et al. (2015), who also used hedonic modeling 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 EUR 717 for the capital city of Amsterdam, EUR 308 for 's-Hertogenbosch, and EUR 184 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", EUR 206 in 2005 (Table 2), and the value share of land are not surprisingly low after all.

6. Summary and Conclusions

Hedonic house price models used for constructing house price indices usually do not explicitly model the value of land. In the present article, we assumed that the value of location is capitalized into land and attempted to account for spatial variation of land prices in the construction of hedonic imputation house price indices. 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 property price 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 indices, and compared these indices with price indices based on more restrictive models, that is a model where land prices vary across postcode areas and a model with no variation in land prices, both estimated by OLS.

The Fisher house price indices were quite insensitive to the choice of model, but the Laspeyres and Paasche indices 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 indices very similar to indices obtained by MGWR. Hill and Scholz (2017), 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 indices compared to indices 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 PCLP-based and LLP-based house price indices 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-based LLP model has two additional advantages. A potential problem with the PCLP 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 indices. The LLP model deals with this problem by using data of the nearest neighbors which are not necessarily confined to a particular postcode area. Most importantly, The LLP model can generate a continuous map of land prices for a city,

which will be more informative than a discrete map that only shows differences between postcode areas.

For some purposes, separate price indices 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 LLP-based price indices of land and structures for the city of "A" are a bit volatile, compared to indices produced by smoothing methods such as the unobserved component approach (Rambaldi et al. 2015). The volatility may be due to sparse data and also to multicollinearity (though we believe the latter is less important). Second, the structures price index increases faster than the official construction cost index, perhaps due a failure to fully control for changes in structures characteristics. Third, the estimated large drop in land prices and increase in structures prices in 2003 seems a bit unusual. While these results could be caused by methodological issues, they could also reflect the impact of a housing market shock which affected households' preferences. Finally, at first glance, the estimated value share of land seems to be rather low. The above-mentioned issues may have played a role here, but the low land share could also be a real phenomenon: households may not value a square meter of land in the city of "A" as much as they do in more prosperous cities with more and better amenities. 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 cities in the western part of the Netherlands, like Amsterdam, Rotterdam or The Hague. Having more observations might also enable us to estimate biannual or even quarterly price indices.

We did not address functional form problems. 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, and replaced age by building period in the empirical estimation. Another potential type of misspecification arises from the linear relation between land price and plot size in our models. As Diewert et al. (2015), Francke and van de Minne (2017) and others have argued, 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 estimate 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. Furthermore, it would be useful to explicitly allow for net depreciation, as in the original models.

Appendix

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 (EUR)											
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 (EUR/m^2)											
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

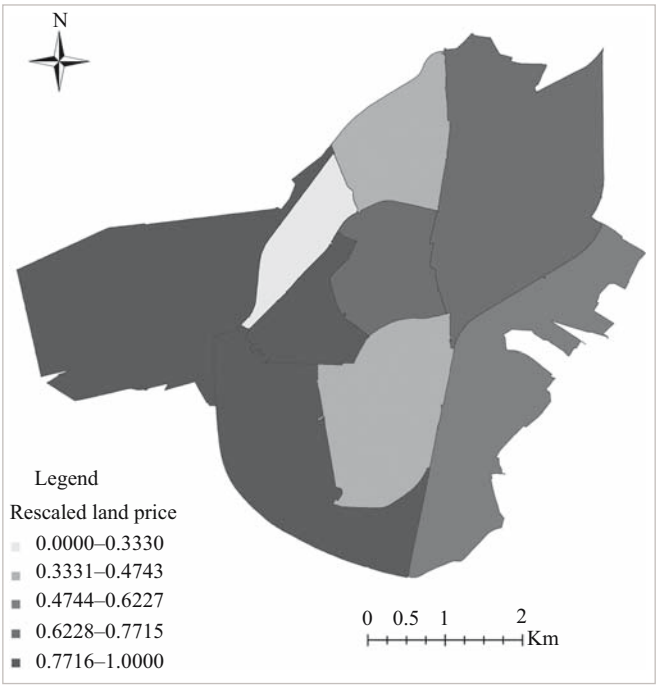

Fig. A1. Price of land per square meter, 2007, PCLP model.

7. References

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28: 281–298. DOI: http://dx.doi.org/10.1111/j.1538-4632.1996.tb00936.x.

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. DOI: http://dx.doi.org/10.1111/0022-4146.00146.

Casetti, E. 1972. "Generating Models by the Expansion Method: Applications to Geographical Research." *Geographical Analysis* 4: 81–91. DOI: http://dx.doi.org/10.1111/j.1538-4632.1972.tb00458.x.

Clapp, J.M. 2004. "A Semiparametric Method for Estimating Local House Price Indices." *Real Estate Economics* 32: 127–160. DOI: http://dx.doi.org/10.1111/j.1080-8620.2004.00086.x.

Davis, M.A. and J. Heathcote. 2007. "The Price and Quantity of Residential Land in the United States." *Journal of Monetary Economics* 54: 2595–2620. DOI: https://doi.org/10.1016/j.jmoneco.2007.06.023.

Davis, M.A. and M.G. Palumbo. 2008. "The Price of Residential Land in Large US Cities." *Journal of Urban Economics* 63: 352–384. DOI: https://doi.org/10.1016/j.jue.2007.02.003.

de Groot, H.L.F., G. Marlet, C. Teulings, and W. Vermeulen. 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: 772–791.

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: 58–105.

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: http://dx.doi.org/10.1080/07474938.2014.944791.

Diewert, W.E., S. Heravi, and M. Silver. 2009. "Hedonic Imputation Versus Time Dummy Hedonic Indexes." In *Price Index Concepts and Measurement*, edited by W.E. Diewert, J.S. Greenlees, and C.R. Hulten, 161–196. Chicago: University of Chicago Press.

Diewert, W.E. and C. Shimizu. 2013. *Residential Property Price Indexes for Tokyo*. Vancouver: The University of British Columbia (UBC Discussion Paper Series No. 13-07).

Dorsey, R.E., H. Hu, W.J. Mayer, and H. Wang. 2010. "Hedonic Versus Repeat-Sales Housing Price Indexes for Measuring the Recent Boom-Bust Cycle." *Journal of Housing Economics* 19: 75–93. DOI: https://doi.org/10.1016/j.jhe.2010.04.001.

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. "Scale Issues and Geographically Weighted Regression." In *Modelling Scale in Geographical Information Science*, edited by N.J. Tate and P.M. Atkinson, 123–140. 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.

Fotheringham, A.S., M.E. Charlton, and C. Brunsdon. 1998b. "Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis." *Environment and Planning A* 30: 1905–1927. DOI: https://doi.org/10.1068/a301905.

Francke, M.K. and A.M. van de Minne. 2017. "Land, Structure and Depreciation." *Real Estate Economics* 45: 415–451. DOI: http://dx.doi.org/10.1111/1540-6229.12146.

Geniaux, G. and C. Napoléone. 2008. "Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditve Models." In *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*, edited by A. Baranzini, J. Ramirez, C. Schaerer, and P. Thalmann, 101–127. New York: Springer.

Hill, R.J. and D. Melser. 2008. "Hedonic Imputation and the Price Index Problem: An Application to Housing." *Economic Inquiry* 46: 593–609. DOI: http://dx.doi.org/10.1111/j.1465-7295.2007.00110.x.

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. DOI: https://doi.org/10.1016/j.jhe.2009.07.010.

Hill, R.J. and M. Scholz. 2017. "Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines." *Review of Income and Wealth* Forthcoming. DOI: http://dx.doi.org/10.1111/roiw.12303.

Hurvich, C.M. and C.L. Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76: 297–307. DOI: https://doi.org/10.1093/biomet/76.2.297.

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

Kuminoff, N.V. and J.C. Pope. 2013. "The Value of Residential Land and Structures During the Great Housig Boom and Bust." *Land Economics* 89: 1–29. DOI: https://doi.org/10.3368/le.89.1.1.

Mei, C., N. Wang, and W. Zhang. 2006. "Testing the Importance of the Explanatory Variables in a Mixed Geographically Weighted Regression Model." *Environment and Planning A* 38: 587–598. DOI: https://doi.org/10.1068/a3768.

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. DOI: https://doi.org/10.1023/A:1007799028599.

Rambaldi, A.N., R.R.J. McAllister, and C.S. Fletcher. 2015. *Decoupling Land Values in Residential Property Prices: Smoothing Methods for Hedonic Imputed Price Indices*. Queensland: University of Queensland (School of Economics Discussion Paper Series No: 549).

Rambaldi, A.N. and D.S.P. Rao. 2011. *Hedonic Predicted House Price Indices Using Time-Varying Hedonic Models with Spatial Autocorrelation*. Queensland: University of Queensland (School of Economics Discussion Paper Series No: 432).

Rambaldi, A.N. and D.S.P. Rao. 2013. *Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes*. Queensland: University of Queensland (CEPA Working Paper Series No: WP03/2013).

Sun, H., Y. Tu, and S.-M. Yu. 2005. "A Spatio-Temporal Autoregressive Model for Multi-Unit Residential Market Analysis." *The Journal of Real Estate Finance and Economics* 31: 155–187. DOI: https://doi.org/10.1007/s11146-005-1370-0.

Tu, Y., S.-M. Yu, and H. Sun. 2004. "Transaction-Based Office Price Indexes: A Spatiotemporal Modeling Approach." *Real Estate Economics* 32: 297–328. DOI: http://dx.doi.org/10.1111/j.1080-8620.2004.00093.x.

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

Received December 2016
Revised July 2017
Accepted November 2017