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 pr

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 construct

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 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, a^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, b^t.[2] After adding an error term u_i^t with zero mean, model (1) becomes

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

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 a^t and b^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

$$p_i^{t*} = a^t r_i^t + b^t + e_i^t , \tag{3}$$

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 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

a_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

$$ {}_i^t \quad {}_{k}^{K} \quad {}_k^t \; {}_{ik} z_{iL}^t \quad {}_{a}^{A} \quad {}_a^t D_{ia}^t z_{iS}^t \quad {}_{r}^{R} \quad {}_r^t D_{ir}^t z_{iS}^t \quad u_i^t $$

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 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 regressions 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

$$P = Z_L \ddot{A} \quad + Z_S \quad + u \tag{16}$$

where $= (a(x_1, y_1), a(x_2, y_2), \ldots, a(x_n, y_n))^T$ is a vector of land prices to be estimated, $\ddot{A}$ is an operator that multiplies each element of by the corresponding element of Z_L, and Z_S is the matrix of structural characteristics included in model (11), given by

$$_S = \begin{matrix} D_{11} z_{1S} & D_{12} z_{1S} & D_{1j} z_{1S} \\ D_{21} z_{2S} & D_{22} z_{2S} & D_{2j} z_{2S} \\ \\ D_{n1} z_{nS} & D_{n2} z_{nS} & D_{nj} z_{nS} \end{matrix}$$

(1) regressing each column of Z_S against Z_L using the GWR calibration method and computing the residuals $Q = (I - S)Z_S$;

(2) regressing the dependent variable P against Z_L using the GWR approach and then computing the residuals $R = (I - S)P$;

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

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

The predicted values for the property prices can be expressed as

$$\hat{P} = S(P - Z_S \hat{\gamma}) + Z_S \hat{\gamma} = LP, \tag{17}$$

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

Equation (18) may need some explanation. All quanti

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 $(t = 1,...,T)$:

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

The imputed constant-quality prices $\hat{p}_i^{0(t)}$ are estimates of the prices that would prevail in period 0 if the property characteristics were those of period t, which are estimated as $\hat{p}_i^{0(t)} = \hat{\alpha}_i^0 z_{iL}^t + \hat{\beta}_i^{0(t)} z_{iS}^t$, where $\hat{\beta}_i^{0(t)} = \hat{\theta}^0 + \sum_{a=1}^{A-1} \hat{\gamma}_a^0 D_{ia}^t + \sum_{r=1}^{R-1} \hat{\lambda}_r^0 D_{ir}^t$ denotes the period 0 constant-quality price of structures. By substituting the constant-quality prices and the predicted prices $\hat{p}_i^t = \hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t$ into equation (20), the imputation Paasche index can be written as

$$P_{Paasche}^{0t} = \frac{\sum_{i \in S^t} [\hat{\alpha}_i^t z_{iL}^t + \hat{\beta}_i^t z_{iS}^t]}{\sum_{i \in S^t} [\hat{\alpha}_i^0 z_{iL}^t + \hat{\beta}_i^{0(t)} z_{iS}^t]} = \hat{\sigma}_L^{t(0)} \frac{\sum_{i \in S^t} \hat{\alpha}_i^t z_{iL}^t}{\sum_{i \in S^t} \hat{\alpha}_i^0 z_{iL}^t} + \hat{\sigma}_S^{t(0)} \frac{\sum_{i \in S^t} \hat{\beta}_i^t z_{iS}^t}{\sum_{i \in S^t} \hat{\beta}_i^{0(t)} z_{iS}^t}, \tag{21}$$

where $\sum_{i \in S^t} \hat{\alpha}_i^t z_{iL}^t / \sum_{i \in S^t} \hat{\alpha}_i^0 z_{iL}^t$ and $\sum_{i \in S^t} \hat{\beta}_i^t z_{iS}^t / \sum_{i \in S^t} \hat{\beta}_i^{0(t)} z_{iS}^t$ are Paasche price indexes of land and structures, which are weighted by = $\in$ $\in$ +

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 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 20

(MGWR). The last model was estimated by mixed geographically weighted regression using the software package GWR 4.0.[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]

$$= 2\ \ln(\hat{\ }) + \ \ln(2\ ) + \frac{+\ (S)}{-\ 2 - \mathrm{tr}(S)}$$

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				

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

5.3 A comparison of different hedonic price indexes

Figure 2: Chained hedonic imputation Paasche house price index

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 index for the price of land in 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

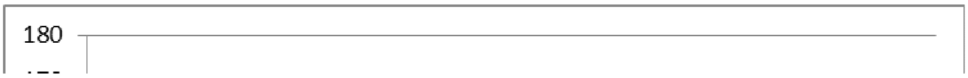
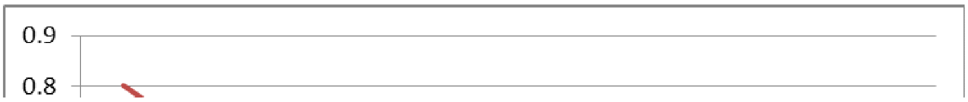


Figure 6: Estimates of value shares of land and structures, OLSD-based



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

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 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.

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 Un

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 Wei