Mixed Geographically Weighted Regression for Hedonic House Price Modelling in Austria

Marco HELBICH and Wolfgang BRUNAUER

1 Background

According to hedonic price theory (ROSEN 1974), real estate is valued for its utility-bearing characteristics. Because a property is fixed in space, a household implicitly chooses many different goods and services by selecting a specific object. From the methodological point of view, this can be explained with the hedonic price function *f*, describing the functional relationship between the real estate price *P* and object characteristics $X_{O1},...,X_{On}$ as well as neighbourhood characteristics $X_{N1},...,X_{Nm}$. Traditional approaches use a log-linear model structure (with the price and some of the continuous covariates logarithmically transformed), which reduces heteroscedasticity and nonlinearity. Nevertheless, locally varying equilibria or “submarkets” can be expected. If not accounted for, this leads to biased results and falsely induced spatial autocorrelation. Therefore, the literature provides a variety of local and global models (e.g. ANSELIN 1988, LESAGE & PACE 2009). One cutting-edge methodology that explicitly models heterogeneity is the geographically weighted regression (GWR, FOTHERINGHAM et al. 2002). Numerous applications (e.g. YU et al. 2007) show the usefulness of this technique, as discussed later on. The main purpose of this research is therefore to define a hedonic pricing model that explains transaction prices for family dwellings in Austria accurately, taking into account structural and locational differences as well as spatial heterogeneity in intercept and slope parameters.

2 Study Site and Data

The data set consists of 3,892 locations of family dwellings situated in Austria for the purchase period of 1998 to 2009 and is provided by the Bank Austria. For GWR analysis, a random sample of 35% (1,393 objects) of the population is used in order to make computation feasible. Additional to the transaction prices, the data comprise 24 structural attributes (e.g. condition of the house, quality of heating system, floor space), as well as characteristics of the surroundings (e.g. proportion of academics, purchase power index).

3 Methodology

The results of global models, in particular the feasible generalized least square and the simultaneous autoregressive error model, indicate serious problems concerning heteroscedasticity and autocorrelation, which advocates the use of GWR. Efficiency can be gained using a mixed GWR model (MGWR, FOTHERINGHAM et. al 2002), a semi-local approach where coefficients with small variation over space are kept constant over Austria. The decision whether an effect is global or local can be carried out on the basis of LEUNG's et al. (2000) statistic. A model with adaptive Gaussian kernel functions is used, which accounts for irregular densities of observations over space. MGWR models can be written as follows:

$$y_i = \sum_{j=1}^{k} a_j x_{ij} + \sum_{l=1}^{m} b_l(u_i, v_l) x_{il} + \varepsilon_i \tag{1}$$

where in this case y_i is the logarithmically transformed sales price of observation i, $i \in 1,\dots,n$, a_j the global coefficients of covariates x_{ij}, $j \in 1,\dots,k$ and $b_l(u_i, v_l)$ the local coefficients of covariates x_{il}, where the pair (u_i, v_l) are the coordinates of observation i. In matrix notation, this model is written as

$$\mathbf{y} = \mathbf{X}_1\mathbf{a} + (\mathbf{B} \circ \mathbf{X}_2)\mathbf{1} + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times k$ matrix of covariates with the respective $k \times 1$ vector of global coefficients $\mathbf{a}$, and $\boldsymbol{\varepsilon}$ is the usual *iid* vector of error terms. $\mathbf{B}$ is an $n \times m$ matrix whose i-th row is given by $\mathbf{b}(i) = (\mathbf{X}_2^{\mathbf{T}}\mathbf{W}(i)\mathbf{X}_2)^{-1}\mathbf{X}_2^{\mathbf{T}}\mathbf{W}(i)\mathbf{y}$, where $\mathbf{W}(i)$ is the diagonal spatial weighting matrix at point i. $\mathbf{B} \circ \mathbf{X}_2$ is the Hadamard product of the matrices

$$\mathbf{B} = \begin{bmatrix} b_0(u_1, v_1) & b_1(u_1, v_1) & \cdots & b_m(u_1, v_1) \\ b_0(u_2, v_2) & b_1(u_2, v_2) & \cdots & b_m(u_2, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ b_0(u_n, v_n) & b_1(u_n, v_n) & \cdots & b_m(u_n, v_n) \end{bmatrix}, \tag{3}$$

and $\mathbf{X}_2$, the $n \times m$ matrix of explanatory covariates with spatially varying coefficients. Here, $\mathbf{B}$ is multiplied entry-wise with the corresponding elements of $\mathbf{X}_2$, i.e. $(\mathbf{B} \circ \mathbf{X}_2)_{ij} = \mathbf{B}_{ij} \times \mathbf{X}_{2ij}$. $\mathbf{1}$ is an $m \times 1$ vector of ones. We write $\boldsymbol{\Gamma} = (\mathbf{B} \circ \mathbf{X}_2)\mathbf{1}$ and define $\mathbf{H}_{\mathbf{X}_2}$ as the partial hat matrix that projects the partial residuals of the response variable given the global part of the model onto $\boldsymbol{\Gamma}$, resulting in $\hat{\boldsymbol{\Gamma}}$, i.e.

$$\hat{\boldsymbol{\Gamma}} = \mathbf{H}_{\mathbf{X}_2}(\mathbf{y} - \mathbf{X}_1\mathbf{a}) \, . \tag{4}$$

The Frisch-Waugh-Lovell theorem states that pre-multiplying the equation with the orthogonal complement of this hat matrix leads to the same result for the parameters $\mathbf{a}$ as estimated in an equation with all covariates:

$$(\mathbf{I} - \mathbf{H}_{\mathbf{X}_2})\mathbf{y} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_2})\mathbf{X}_1\mathbf{a} + (\mathbf{I} - \mathbf{H}_{\mathbf{X}_2})\boldsymbol{\varepsilon} \, . \tag{5}$$

Using the estimated global parameters $\hat{\mathbf{a}}$ in turn, one can subtract $\mathbf{X}_1\hat{\mathbf{a}}$ from both sides of equation (2) and estimate a basic GWR model. Therefore, in contrast to the GWR without any fixed coefficients, MGWR is estimated in following steps: (a) Regress each vector $\mathbf{x}_j$, $j \in 1,\dots,k$ on $\mathbf{X}_2$ using GWR, obtaining the residuals $\tilde{\mathbf{x}}_j$ that form the columns of

$\tilde{\mathbf{X}}_1$. (b) Regress $\mathbf{y}$ on $\mathbf{X}_2$ using GWR, obtaining the residuals $\tilde{\mathbf{y}}$. (c) Regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_1$, which yields the correct coefficients $\hat{\mathbf{a}}$ for the non-varying part of the model. (d) Calculate the residuals $\tilde{\tilde{\mathbf{y}}} = \mathbf{y} - \mathbf{X}_1\hat{\mathbf{a}}$ and regress them on $\mathbf{X}_2$. This yields the correct local coefficients. An in-depth discussion can be found in Fotheringham et al. (2002). Common techniques to transfer the pointwise parameter estimations on non-observed locations are interpolation (e.g. Yu et al. 2007) or model estimation on a regular grid (Fotheringham et al. 2002). For our application, ordinary kriging was employed. All calculations are accomplished in the R environment for statistical computing.

4 Results

After some model selection procedures (minimizing AIC), the final model consists of seven global predictors and nine significant non-stationary variables. Table 1 gives an overview concerning the parameter estimations and confirms the pre-assumed relationships.

Tab. 1: Parameter estimations

		Global parameters			Local parameters		
		Estim.	Std. err.	t-val.	1. QT	Med.	3. QT
Non-stationary	Purchase power 2009 (M)***	0.005	0.001	4.497	0.000	0.003	0.006
	Share of academics 2001 (M)***	0.012	0.003	3.897	0.010	0.016	0.020
	Age index (M)***	-0.045	0.005	-8.635	-0.040	-0.030	-0.025
	Ln populat. density 2009***	0.066	0.009	7.459	0.038	0.075	0.083
	Condition house 3 (D)*	-0.044	0.019	-2.236	-0.098	-0.085	-0.004
	Attic 1 (D)**	-0.061	0.019	-3.137	-0.063	-0.039	-0.032
	Ln of plot space***	0.092	0.020	4.527	0.040	0.112	0.151
	Ln of total floor area***	0.466	0.029	15.906	0.375	0.423	0.518
	Age of building***	-0.006	0.001	-11.205	-0.006	-0.005	-0.005
Stationary	Intercept	-0.007	0.008	-0.828			
	Unemployment rate 2009 (M)	-1.232	1.735	0.478			
	Share academ. deviation from municipality mean 2001 (C)***	0.008	0,002	3.939			
	Quality heating syst. 3 (D)***	-0.132	0.038	-3.491			
	Quality bathroom 3 (D)	-0.017	0.035	-0.484			
	Existence cellar (D)***	0.117	0.022	5.368			
	Quality garage 3 (D)***	-0,094	0,020	-4.765			
	Terrace (D)*	0.044	0.019	2.257			

M = Municipal. level, C = Census tract level, D = Dummy, QT = Quantile, Signif.: '***' 0.001, '**' 0.01, '*' 0.05

For instance, the global predictor "quality of the garage" has a negative effect on the dwelling's price. A garage in bad condition reduces the price approximately about 9%. The same can be said for the quality of the heating system, whereas bad conditions yield a 13% reduction of price. Deeper insights can be gained from Figure 1, where exemplarily four kriged parameters are mapped. It can be clearly seen that there is significant spatial heterogeneity in the predictors. Thus, the relationship between these variables and the house price depends on the geographical location, which cannot be explored with a global stationary model. For example, the index of purchase power has a negative effect on the price in and around the metropolitan areas of Vienna and Salzburg, elsewhere the effect is positive.

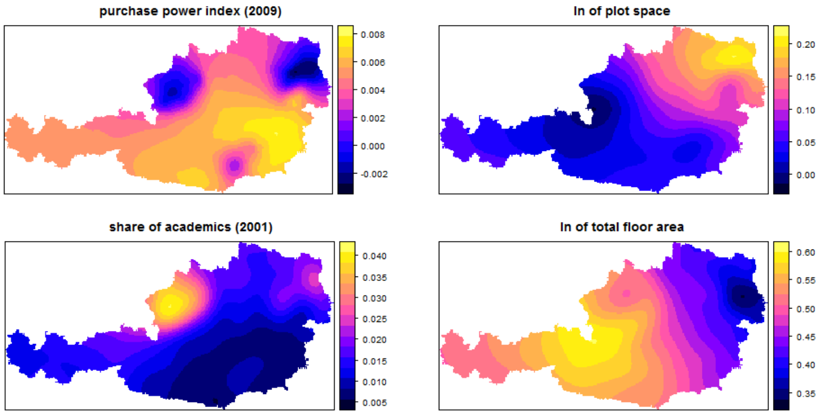

Fig. 1: Spatial distribution of four non-stationary parameters as kriged surfaces

The model fit, indicated by local R^2, varies between 0.25 and 0.50, whereas the lowest values are located in Vienna and its surroundings, as well as the south of Austria. The highest fit is achieved in northern areas. Finally, it is certainly worth noting that MGWR is a useful method to explore heterogeneity, although with limited applicability for large datasets. Hence, the application of computationally more efficient algorithms seems promising, particularly the tensor product smooths approach (WOOD 2006), where a penalized regression approach is adopted in which low-rank, scale-invariant tensor product smooths are constructed. The smooths can be written as components of (generalized) additive mixed as well as of standard (generalized) linear mixed models, allowing them to take advantage of the efficient and stable computational methods that have been developed for such models.

References

ANSELIN, L. (1988), Spatial Econometrics. Methods and Models. Dordrecht: Kluwer.

FOTHERINGHAM, S., BRUNDSON, C. & CHARLTON, M. (2002), Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley: Chichester.

LESAGE, J. & PACE, K. (2009), Introduction to Spatial Econometrics. Boca Raton: CRC.

LEUNG, Y., MEI, C.-L. & ZHANG, W.-X. (2000), Statistical Tests for Spatial Nonstationarity Based on the Geographically Weighted Regression Model. Environment and Planning A, 32, pp. 9-32.

ROSEN, S. (1974), Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82, pp. 34-55.

WOOD, S. (2006), Low-Rank Scale-invariant Tensor Product Smooths for Generalized Additive Mixed Models. Biometrics, 62, pp. 1025-1036.

YU, D., WEI, Y.D. & WU, C. (2007), Modeling Spatial Dimensions of Housing Prices in Milwaukee, WI. Environment and Planning B: Planning and Design, 34, pp. 1085-1102.