EKONOMETRIČNA ANALIZA CEN HIŠ V SLOVENIJI NA PODLAGI ZDRUŽEVANJA V SKUPINE

CLUSTER-BASED ECONOMETRIC ANALYSIS OF HOUSE PRICES IN SLOVENIA

Miroslav Verbič, Peter Korenčan

UDK: 330.43:332.74
Klasifikacija prispevka po COBISS.SI: 1.01
Prispelo: 29. 9. 2016
Sprejeto: 8. 4. 2017

DOI: 10.15292/geodetski-vestnik.2017.02.231-245
SCIENTIFIC ARTICLE
Received: 29. 9. 2016
Accepted: 8. 4. 2017

IZVLEČEK

Osrednji namen članka je proučiti oblikovanje cen hiš v Sloveniji s kombinacijo geografsko obtežene regresije in hierarhičnega združevanja v skupine. Prikazano je tudi, kako uporabiti geografsko obteženo regresijo za napovedovanje cen hiš, prilagojeno lastnostim regionalnih trgov. Z globalnim in regionalnimi regresijskimi modeli smo analizirali presečne podatke o prodajah nepremičnin iz leta 2015. Prostorski vplivi na vrednost nepremičnine so bili popisani na podlagi vrednostnih območij in regionalizacije. Ugotovili smo, da na vrednost hiš statistično značilno vplivajo lokacija, bližina železniške proge, priključek na plinovod, neto tlorisna površina, starost ter površina pripadajočih stavbnih in kmetijskih zemljišč.

ABSTRACT

The main purpose of the article is to investigate the formation of house prices in Slovenia with a combination of geographically weighted regression and hierarchical clustering. It also demonstrates how to operationalize the geographically weighted regression for house price forecasting, tailored to the regional market features. We analyzed data on house sales in Slovenia for 2015 by estimating a global and regional regression models. The effect of location on house value was described by employing house value zones and through regionalisation. Results reveal that location, railway proximity, access to gas network, house area, house age, and area of building and farming lot belonging to a house have a statistically significant effect on house prices.

KLJUČNE BESEDE

analiza cen hiš, geografsko obtežena regresija, združevanje v skupine, regionalizacija

KEY WORDS

analysis of house prices, geographically weighted regression, clustering, regionalisation

1 INTRODUCTION

The residential real estate market in Slovenia has been facing a period of stagnation for the last couple of years. After the crash in 2008, real estate prices dropped significantly, though the number of market transactions remained low until the end of 2013. At the beginning of 2014, we witnessed an increase in the number of sales, especially of apartments, and less significantly for houses, which lasted until 2015. That was probably the result of decreased apartment prices, which achieved in 2015 the lowest average values per square meter since the beginning of systematic monitoring of the market in 2007. On average, apartments were sold in 2015 for 1,440 €/m^2, which represents a 1% decrease from 2014 and a 21% decrease from 2007.

We can observe similar dynamics for houses, with some unique features. Namely, differences arise from the way houses were constructed in Slovenia. In rural areas, most of them were built by owners without the cooperation of construction companies, thus they differentiate widely and are tailored to owners' individual desires. We can also observe spatial heterogeneity in rural areas, which was caused by poor urban planning and "ignorance" of the law. As a result, well organized housing markets can be found in Ljubljana and Maribor only. However, we can still see that, at the national level, price bottom was reached in 2014, when houses were sold on average for 105,000 €. In 2015, the average price increased to 108,000 €, and is related to a 3% decrease in the number of transactions (GURS, 2016b).

In this article, we analyse the house prices in Slovenia for the year 2015, where various house and transaction features are used as determinants of house prices. We introduce two main hypotheses. First, the effect of location on house prices can be measured through the value zone in which the house is located and through the differences between the regression coefficients of the same explanatory variable across different regions. Second, house and transaction features, such as location, railway proximity, access to gas network, house area, house age, and area of building and farming lot belonging to a house, statistically significant affect the house prices.

House prices have been investigated for several decades, primarily by employing econometric models estimated by the least squares estimator. The main limitation of the method is that it produces average or global parameter estimations, which then have to be applied to the whole region that is being analysed (Fotheringham, Charlton and Brunsdon, 1998). This implies that the parameters are stationary over space, which has been proven false in researching various social phenomena (Chrostek and Kopczewska, 2013; Du and Mulley, 2011; Romih and Bojnec, 2008). Three main reasons for spatial non-stationarity are defined by Fotheringham et al. (1998) as: (1) random variation in data sample, (2) heterogeneous relationships across space, and (3) unavoidable misspecification of the model.

Even though the issues with global regression models are known, little is done (especially) with respect to the first and the third reason for spatial non-stationarity, which are caused by our limited understanding of the phenomena. On the other hand, significant gains were made in understanding heterogeneous relationships between variables across space. The first step in this direction can be traced to the introduction of the spatial expansion method (SEM) by Casetti (1972). His aim was to expand the least squares approach with parameters that are functions of subsequent variables. In this way, SEM offers researchers a possibility of including heterogeneous spatial attributes into the analysis. There were several attempts to use SEM models in analyzing house prices (*cf.* Gelfand, Kim, Simans and Banerjee, 2003). Complexity of the models is notable, which might explain the researchers transitioning to the geographically weighted regression (GWR) that evolved from SEM and was first discussed by Brunsdon, Fotheringham and Charlton (1996).

The novelty of the approach, employed in this article, is to combine the geographically weighted regression in investigating house prices with hierarchical clustering, in order to improve the explanatory and forecasting power of the models. Hierarchical clustering is a straightforward approach, which gives us an ability to identify clusters based on measures of similarity between residential units. The data used in this research document are house sales in Slovenia in the first half of 2015. They were gathered by the Surveying and Mapping Authority of the Republic of Slovenia, and were made available on request of the authors. Data contain information on sales terms, house attributes, maps of public infrastructure, and house value zones.

The article is structured as follows. In Section 2, we discuss the methods employed in our research, in particular the geographically weighted regression and the hierarchical clustering. Section 3 contains a detailed description of the data employed, in order to understand the current market conditions in Slovenia. Section 4 is devoted to reporting the main results of our analysis. Section 5 concludes with the key findings and indicates possible steps for further research.

2 METHODS

In this section, we first introduce and discuss the geographically weighted regression, which will give us the ability to adopt the models to local conditions. Then, we discuss ways to measure distance, which is a core element of the geographically weighted regression. Furthermore, we introduce and describe the clustering methods, which are essential in order to divide sample units into homogeneous regions.

2.1 Geographically weighted regression

The geographically weighted regression (GWR) is based on the weighted least squares (WLS) estimator, where the distances between the observed unit and all other units in the sample are used as weights. The method basically enables us to obtain the local parameter estimates instead of the global ones (Fotheringham et al., 1998). The data generating process for observation i, $i = 1, \ldots, N$, is thus the following:

$$y_i = \beta_0\left(u_i, v_i\right) + \sum_k \beta_k\left(u_i, v_i\right) x_{ki} + \varepsilon_i \tag{1}$$

where y is the dependent variable, x_k are the explanatory variables, β_k represent the regression coefficients, and ε is an IID (independent and identically distributed) disturbance term. Equation (1) also introduces two non-standard variables, u_i and v_i, which represent the coordinates of each sample point.

This means that instead of one global model we now get N local models, which are fitted for each point in the sample (Fotheringham et al., 1998; Lu, Charlton and Fotheringham, 2011; Du and Mulley, 2011). Parameter estimates are thus obtained as:

$$\widehat{\beta_i}\left(u_i, v_i\right) = \left(X'W\left(u_i, v_i\right)X\right)^{-1} X'W\left(u_i, v_i\right)y \tag{2}$$

where $W(u_i, v_i)$ represents the square diagonal matrix of graphically conditioned weights for pivotal point i. The matrix is presented in equation (3) and has number of rows and columns equal to N:

$$W\left(u_i, v_i\right) = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{iN} \end{pmatrix} \tag{3}$$

In order to calculate the weights, w_{ij}, we first have to introduce the measure of distance, d_{ij}, which can be described in different terms. Most commonly, the Euclidean distance is used, though it may not be the optimal in some cases, as nowadays people may perceive distance primarily as road distance or even as time spent to go from point i to point j. Lu et al. (2011) analysed the purchase prices of houses in London using the GWR method with both the Euclidean distance and the travelling time. They found that the latter provides better results, which corresponds to the perception that distance is measured by people as time spent on the way. Downside of the travelling time measure is in its complex calculation process, which needs a precise road network map and travelling speed for each segment of the network, thus researchers do not use it as widely as the Euclidean measures.

Relationship between w_{ij} and d_{ij} is not straightforward due to the assumption that the effect of distance is not linear. In general, there are five types of *kernel* functions (Lu et al., 2015), which all consist of two variables: distance (d_{ij}) and bandwidth (b). An important attribute of all *kernel* functions is also that they limit towards 0, when $d_{ij} \rightarrow \infty$.

The most used weight is the *Gaussian weight*, which is defined as:

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \tag{4}$$

with the main advantage being in its continuity, which makes it applicable for all sorts of real data that might produce errors with other approaches. A very similar approach is to use the *exponential weight*, which does not include division by two and squaring of the d_{ij}/b quotient. Some researchers use the *bi-square weight* (Brunsdon et al., 1996), due to its nullification of the out-of-bandwidth observations effect:

$$w_{ij} = \begin{cases} if\ d_{ij} > b \rightarrow \left[1-\left(d_{ij}/b^2\right)\right]^2 \\ else \rightarrow 0 \end{cases} \tag{5}$$

A modification of the latter approach produces the *tri-cube weight*, which uses cubing instead of squaring the elements. The last kernel function is extremely basic, and may be used only with great caution. It is called the *boxcar weight* and is defined as:

$$w_{ij} = \begin{cases} if\ d_{ij} > b \rightarrow 1 \\ else \rightarrow 0 \end{cases} \tag{6}$$

Bandwidth (b) can be determined in two ways. The first one is based on a broad theoretical basis of the right value. However, often this is not possible, thus the researchers came up with an econometric solution comprised of two steps (Brunsdon et al., 1996). The first step is to calculate fitted values with the GWR model, where the weights are calculated without bandwidth:

$$w_{ij} = \exp\left[- - \left(d_{ij} \right) \quad \right] \tag{7}$$

whereas in the second step, we employ the least squares method with *y* as the dependent variable and (previously) fitted values $\hat{y}(b)$ as an explanatory variable with bandwidth *b*:

$$b = \sum_{i} \left\{ y_i - \widehat{y_i}\left(b\right) \right\}^2 \tag{8}$$

However, this configuration may lead to a problem, since when *b* is decreasing, the effect of other observations is decreasing as well and the effect of observation *i* is increasing (Fotheringham et al., 1998), which can lead to “islands”. In order to counter that problem, the method of cross validation (CV) is used, based on the idea that in order to evaluate the effect of other observations on the pivotal one, we have to omit the effect of the latter observation (Brunsdon et al., 1996). Therefore, the equation for *b* is given as:

$$b = \sum_{i} \left\{ y_i - \hat{y}_{\neq i}\left(b\right) \right\}^2 \tag{9}$$

The cross validation technique was first proposed by Cleveland (1979) in terms of smoothing locally weighted robust regressions. In case of GWR, the CV technique is updated with kernel functions, introduced by Bowman (1984; *cf.* Fotheringham et al., 1998). This enables us to determine the effect of neighbouring elements based on distances between them.

Since 1990, the GWR method has become popular in different scientific areas where phenomena are spatially related. Further development of the method is focused on improving the weighting scheme and combining the method with other methods, such as linear mixed models (Lu and Zhang, 2012). The aim of these improvements is to maximize the explanatory power of such models. However, there is an issue of the GWR technique that remains unanswered – its forecasting capability – largely due to the fact that the GWR method results in a family of models, equal in number to the number of observations.

2.2 Distance metrics and clustering

Every point in the sample data is described with x and y coordinates, which allows us to calculate the distance between points. When real market data are used, the Minkowski approach to measuring the distance is used, as proposed by Lu, Charlton, Brunsdon and Harris (2016):

$$d = \left(\sum_{i=1}^{m} \left| u_i - v_i \right|^2 \right)^{1/p} \tag{10}$$

where u_i and v_i are vectors of dimension *m* defined in Euclidean space, and *p* is a positive real number. Based on Lu et al. (2016), we decided to use *p* = 2, which represents the standard Euclidean distance.

Next, the kernel function has to be chosen, which simulates the effect of neighbouring observations on the current one. Nowadays, the GWR method is often enhanced with *bi-square* (Fotheringham et al., 1998) and *tri-cube* kernels (Lesage, 1999). Nevertheless, these approaches have one disadvantage, which becomes obvious when applying them to heterogeneous spatial data. Namely, they both tend to produce islands that do not have enough neighbouring observations in bandwidth range (*b*) for employing the least squares estimator. This disadvantage is addressed with the introduction of adaptive bandwidth, which adapts to the pattern so that the radius is bigger where the observations are sparse and smaller where they are dense (Kupfer and Farris, 2006). This is achieved by determining the observations that have a significant effect on the given observation, producing *n* sub-samples of the same size *m*.

Although many authors find the *bi-square kernel with adaptive bandwidth* attractive (Fotheringham et al., 1998; Guo, Ma and Zhang, 2008; Kupfer and Farris, 2006; Lesage, 1999), it is still not prevalent due to low increase in explanatory power and significant increase in complexity compared to the Gaussian kernel. Therefore, researchers in real estate valuation still tend to use the Gaussian kernel (Bitter, Mulligan and Dall'erba, 2006). For these reasons, the Gaussian weight is chosen in the present article, as shown in equation (4). This allows us to incorporate all available data in a localised regression.

Subsequently, a family of regression coefficients is calculated. As the existing routines in the R software package return the descriptive statistics of a coefficient family separately for each coefficient, a new routine was developed, which returns a matrix of coefficients K_r:

$$K_r = \begin{pmatrix} b_{i0} & b_{i1} & \cdots & b_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n0} & b_{n1} & \cdots & b_{nk} \end{pmatrix} \tag{11}$$

where b_{i0} represents the regression constant of observation *i*, and b_{nk} represents the *k*-th coefficient of the *n*-th observation. Matrix K_r is the foundation for model agglomeration that follows, and enables us to observe the aspects of the localised regression coefficients.

Agglomeration of models is achieved by using the clustering method. In general, clustering is defined as a technique with which we divide a multivariate dataset into clusters or groups of similar units (Košmelj and Breskvar Žaucer, 2006). In terms of clustering GWR models, we divide localised regression models into groups based on their similarity. By clustering, we can get homogeneous groups or spatial regions where house prices are formed similarly with regard to the chosen explanatory variables.

The process consists of five steps (Košmelj and Breskvar Žaucer (2006): (1) choosing units of observation and explanatory variables; (2) standardising the variables if needed; (3) choosing the suitable distance between units; (4) using different classification methods; and (5) analysing the obtained results. In our case, the regression coefficients are the ones that are being standardised. The third step is skipped due to the usage of hierarchical clustering – a stepwise process where the closest units are joint till all are in one group (Murtagh, 2016). The similarity criteria used was proposed by Ward (1963), as these yield reasonable results and are known for its robustness. With the information obtained from the stepwise process, we can generate the dendrogram and determine the optimal number of cluster. These are then placed on the actual map and geographically defined as regions.

The last step of the analysis is to estimate the least squares models on regional data subsamples and perform model diagnostics to check for potential problems. In particular, each model is tested for heteroscedasticity, which introduces bias to the standard errors of regression coefficients (White, 1980). The problem is very common in housing econometrics due to the complexity of cross-section determinants affecting the real estate prices, and was already noted by some other authors analysing the Slovenian real estate market (Cirman et al., 2015; Romih and Bojnec, 2008).

3 DATA

The data for our research were obtained from the Slovenian Property sales register (GURS, 2016a). Our sample data contained information about house and apartment purchases in Slovenia between 1 February and 31 July 2015, the total sample size being equal to 5,891 transactions. We excluded all apartments (3,322), unfinished houses (231) and houses where only a portion was sold (1,055). We also investigated the variables and excluded all observations where the price per square meter (*Price*) was above 10,000 €/m^2 and/or the size was greater than 2,000 m^2 (90 transactions). This step was necessary in order to keep the sample homogeneous and to exclude possible errors in the database. The descriptive statistics are presented in Table 1. The final sample size amounted to 1,193 transactions.

Table 1: Descriptive statistics of the final sample (*n* = 1,193).
Source: GURS (2016a); own calculations.

Variable	Mean	Standard deviation	Minimum	Maximum
Price (€)	100,304.47	94,035.39	1,000	1,528,773
Location (value zone)	9.16	4.33	1	20
Railway proximity	0.02	0.13	0	1
Access to gas network	0.18	0.38	0	1
House age (years)	50.68	26.83	2	90
House area (m^2)	147.32	108.25	20	1,998
Farming lot (m^2)	1,959.22	10,799.65	0	184,190
Building lot (m^2)	886.42	1,124.89	0	10,920

Location was determined by a map of value zones, which was proposed by the Decree on determining real estate valuation models (Official Gazette of the Republic of Slovenia 95/2011; hereinafter referred to as the "Decree"). The Decree divides the country into 20 zones, which are internally homogeneous and differentiated by the average house price (see Figure 1). Previous research showed that this variable has an outstanding explanatory power and is highly significant in explaining house prices (*cf.* Romih and Bojnec, 2008).

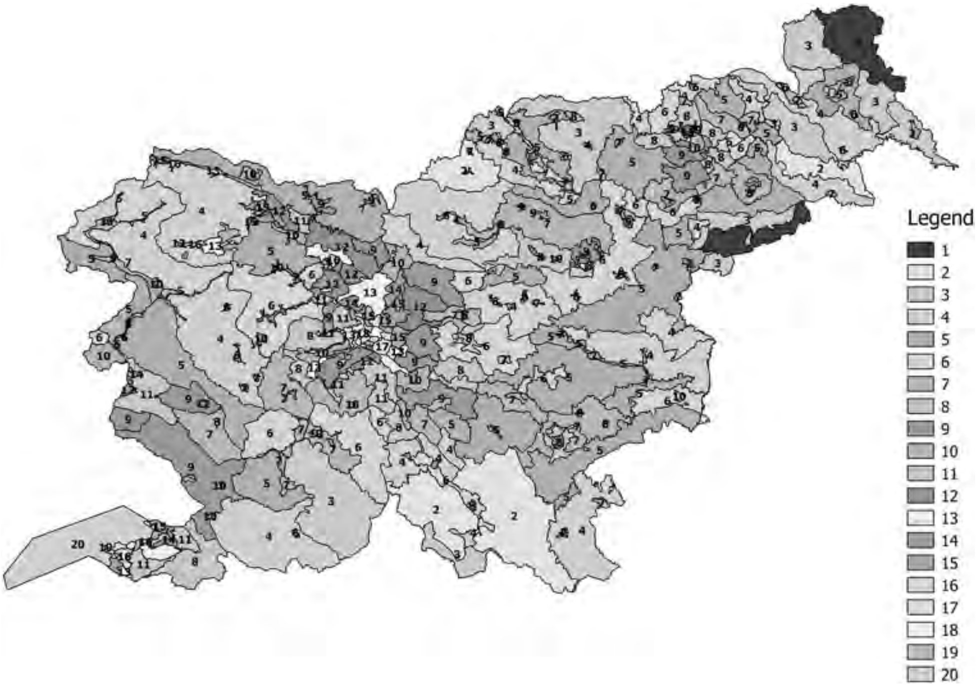

Figure 1: Map of house value zones.
Source: Decree on determining real estate valuation models (Official Gazette of the Republic of Slovenia 95/2011).

Railway proximity represents a dummy variable, proposed by the Surveying and Mapping Authority of the Republic of Slovenia in its document entitled Equations and value calculation using real estate valuation models (GURS, 2009). It suggests that a railway has a significant negative effect on house prices if the house is less than 75 meters away from active tracks. Therefore, the variable is defined with value 1 when the house is located inside the effect zone and 0 if it is outside. This variable is important due to the disturbances caused by railway traffic. In our sample, 22 units were positioned inside the 75 m buffer area. *Access to gas network* is a dummy variable that indicates whether a house is connected to a public gas network (1) or not (0). Connection to a gas network is important, because natural gas in one of the least expensive energy sources for heating. In our sample, 213 units were connected to a network. Variable *House age* represents the difference between the year of analysis and the year in which the house was build, though with a ceiling:

$$Age = \min \begin{cases} Age_{ef} \\ 90 \end{cases} \qquad (12)$$

The ceiling is set in order to reduce the age affect of very old houses that are still in good shape on their price. Variable *House area* represents the sum of net floor areas of all spaces in the house. Some researchers suggested that the effect of an area on the price is not linear, but rather parabolic (*cf.* Hu, Wang and Feng, 2013), therefore we also use the area squared in some model specifications. Usually houses come with premises that are divided into building lot and farming lot, based on the intended use. *Building lot* represents the sum of all plots that were sold with the house with the intended use defined as building land, whereas *Farming lot* represents the sum of all plots that were sold with the house for farming and thus cannot be build upon.

4 EMPIRICAL RESULTS

Hereinafter, we present the main findings obtained with a combination of geographically weighted regression for investigating house prices and hierarchical clustering for improving the explanatory and forecasting power of the models. This will enable us to better understand the differences between using only the GWR and adding the regionalisation techniques. By comparing the regression coefficients, we will be able to better understand the effects of local conditions on house price determination.

The dependant variable in our analysis is the selling price of a house, which will be explained by employing seven explanatory variables that were proposed by the Surveying and Mapping Authority of the Republic of Slovenia and were proven effective by other researchers (Romih and Bojnec, 2008; Cirman, Pahor and Verbič, 2015). These variables are: location, railway proximity, access to gas network, house area, house age, farming lot, and building lot. Due to increased explanatory power of the model, the dependant variable was defined as a natural logarithm of the price in Euros – a transformation that yielded good results in previous studies (Du and Mulley, 2011; Cirman et al., 2015).

4.1 Global model

The first step was to estimate the global model, based on the final sample of 1,193 transactions, and establish the point of comparison for subsequent research:

$$\begin{aligned}\ln\left(Price\right) = \beta_0 + \beta_1 Zone + \beta_2 Railway + \beta_3 Gas + \beta_4 Age + \beta_5 Area + \beta_6 Area^2 + \\ +\beta_7 Building\_lot + \beta_8 Farming\ \_\ lot + \varepsilon\end{aligned} \tag{13}$$

where β_j, $j = 0, \dots, 8$ are regression coefficients and ε is an IID-distributed disturbance term. The estimates of regression coefficients of the global model were calculated by using the basic R function *lm* from the package `stats` and are reported in Table 2.

Table 2: Regression coefficient estimates of the global model.
Source: GURS (2016a); own calculations.

Coefficient estimate	Value	Robust standard error	t–statistic	p–value
Intercept	10.01000	0.0657700	152.2460	0.0000
Location (value zone)	0.128900	0.0044810	28.7570	0.0000
Railway proximity	–0.230000	0.1236000	–1.8610	0.0630
Access to gas network	0.110400	0.0480300	2.2980	0.0217
House age	–0.010690	0.0006466	–16.5250	0.0000
House area	0.002841	0.0002651	10.7150	0.0000
House area squared	–0.000001	0.0000002	–6.7750	0.0000
Building lot	0.000115	0.0000154	7.4810	0.0000
Farming lot	0.000011	0.0000016	7.1230	0.0000
n	1,193			
R^2	0.61			

As can be seen from the regression results in Table 2, all the regression coefficients are statistically significant ($p < 0.10$), most of them highly statistically significant ($p < 0.001$). First, when a house is "promoted" to a higher value level (by one value level), on average, *ceteris paribus*, the house price increases approximately by 12.9%. This result can be logically corroborated in practice due to the fact that houses in the highest value levels (zones) are the most expensive. We can also infer that the house price decreases, on average, *ceteris paribus*, by approximately 23.0%, if the house is situated close to an active railway. This is reasonable from the owner's viewpoint due to increased sound pollution. On the other hand, access of a house to gas network increases, on average, *ceteris paribus*, the house price by approximately 11.0%. This is also reasonable, as natural gas in one of the least expensive energy sources for heating. Moreover, if the age of a house increases by one year, on average, *ceteris paribus*, the house price decreases approximately by 1.1%.

The effect of house area on the house price is indeed non-linear, as hypothesized earlier. Namely, an increase in the area of a house by 1 m^2, on average, *ceteris paribus*, results in an increase of the house price by approximately 0.3%. This effect becomes negative for larger houses (with area greater than 142 m^2), as indicated by the regression coefficient for house area squared, though this U-shaped effect is very low in strength (–0.000001). The non-linear effects of this kind were also confirmed by Cirman et al. (2015), analysing the market for apartments in the wider Ljubljana area, and Romih and Bojnec (2008), based on a nationwide sample of second-hand apartments.

Moreover, if the building lot increases by 1 m^2, on average, *ceteris paribus*, the house price increases by approximately 0.01%. The effect is small in strength, though still tenfold when compared to the effect of farming lot. Namely, if the farming lot increases by 1 m^2, on average, *ceteris paribus*, the house price increases approximately only by 0.001%. The latter two regression coefficients seem plausible for two reasons. First, because a unit change in these two explanatory variables is relatively small, and second, because it is extremely hard in Slovenia to change the intended land use to building land, which is a much more desirable commodity.

The explanatory power of the global model was rather good, with the multiple determination coefficient being equal to 0.61, the Akaike information criterion having a value of 2,067 and the Schwarz information criterion being equal to 2,118 (the latter two are useful first and foremost for model comparison). Nonetheless, we detected the problem of heteroscedasticity in our global model specifications (by employing Breusch–Pagan and White tests). The problem was dealt with in two ways. First, the detected heteroscedasticity was taken into account by applying a robust variance estimator for calculating the standard errors (see Table 2), and second, it was corrected for by employing the regional models (see Section 4.2).

4.2 Regional models

In order to evaluate the GWR model, we first determined the bandwidth value by using the R function *bw.gwr* from the package `GWmodel` (Gollini et al., 2015), which amounted to 253.88 km. Considering that the distance across Slovenia is roughly 258 km, the calculated bandwidth is quite large. Nevertheless, the most likely reason for this result is the inclusion of value zones into the analysis, as these account for a great amount of spatial variability in house prices. Consequently, the spatial heterogeneity decreases and significant differences occur only when observations are (very) far apart. This is in no way an issue for our analysis, as the Kernel function from equation (4) still evaluates the effects of neighbouring observations, only the decrease of effects is slower with greater distance.

Next, a family of GWR coefficients was calculated and clustered. This was achieved using a self-developed R function. It uses previously discussed bandwidth from the R function *gwr.Gauss* from the package `spgwr`, developed by Bivand and Yu (2006), to evaluate the weights matrices. The GWR family was evaluated using the R function *lm* for each set of weights, which yielded a matrix with 1,193 rows and 9 columns, representing regression coefficients for each data point in the sample. Afterwards, the clustering process was initiated, where the units were clustered based on proximity of the regression coefficients. The first step was to normalise the GWR family matrix values. Next, the function *dist* with method *euclidean* from the R package `stats` was used in order to get the matrix of distances between units of the GWR family. In the last step, we used the function *hclust* with the method *ward* from package `stats`, which returned three structures, represented in the dendrogram (Figure 2).

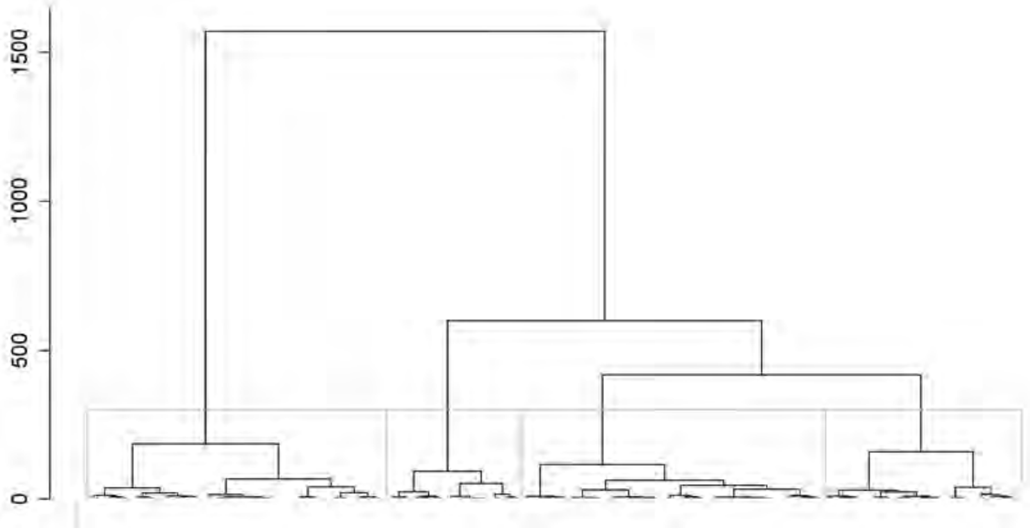

Figure 2: Dendrogram of clustered GWR coefficients.
Source: GURS (2016a); own calculations.

The dendrogram enables us to divide the localised models into four homogeneous groups, based on the stepwise process, representing: (1) Primorska and Notranjska regions, south-west (PN); (2) Gorenjska region and capital city of Ljubljana with its suburbs, north-west and central Slovenia (GLO); (3) Koroška and Savinjska regions, Zasavje and part of Dolenjska region, from the north to south-east (KSD); and (4) East of the country with Maribor as second largest city and Prekmurje region (MP). We can observe from Figure 3 that the groups stand for geographically coherent and spatially homogeneous regions. In order to evaluate the regional (subsample) model specifications, we again employ the least squares estimator. The estimates of regression coefficients of the four regional models are reported in Table 3.

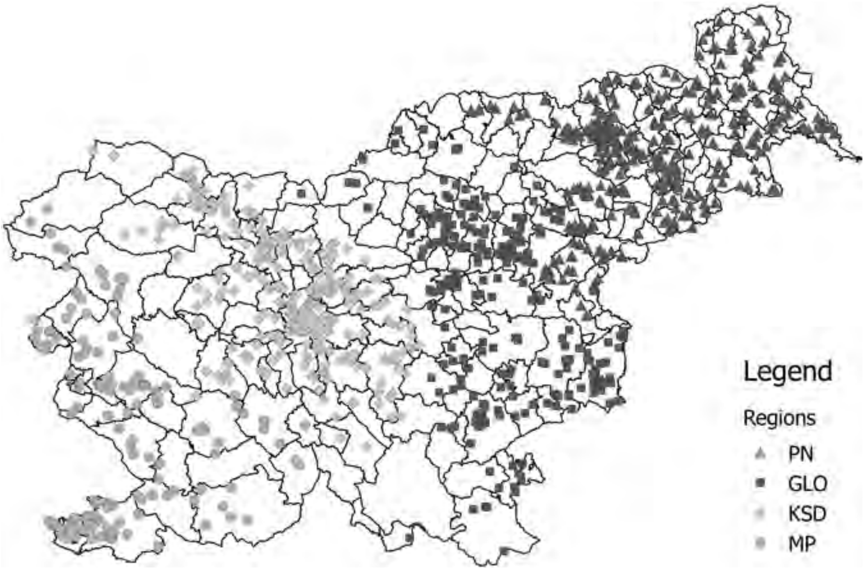

Figure 3: Spatial position of the clusters.
Source: GURS (2016a); own calculations.

As can be seen from the regression results in Table 3, when a house is "promoted" to a higher value zone (by one value zone), on average, *ceteris paribus*, the house price increases approximately by 12.0–15.5% (depending on the group of observations). The effect is highly statistically significant (p < 0.001) for all four groups of observations and in line with the result from the global model (12.9%). The negative effect of an active railway on the house price is statistically significant (p < 0.10) only for Gorenjska region and capital city of Ljubljana with its suburbs. On the other hand, access to gas network has a positive and statistically significant effect on the house price for Primorska and Notranjska regions and for Koroška and Savinjska regions, Zasavje and part of Dolenjska region. This explanatory variable is missing in the last model specification (MP), as none of the houses in that group of observations had a connection to a public natural gas network. The signs of the effects of the latter two explanatory variables are in line for all four groups of observations with the result from the global model.

Next, if the age of a house increases by one year, on average, *ceteris paribus*, the house price decreases approximately by 0.9–1.1% (depending on the group of observations). The effect is highly statistically significant for all four groups of observations and in line with the result from the global model (1.1%). The U-shaped effect of house area on the house price is also present in all four groups of observations, (highly) statistically significant and in line with the result from the global model. Namely, an increase in the area of a house by 1 m^2, on average, *ceteris paribus*, results in an increase of the house price by approximately 0.3–0.7%, whereas this effect becomes negative for large houses (with area greater than 423 m^2), as indicated by the negative regression coefficients for house area squared (see Table 3).

Moreover, if the building lot increases by 1 m^2, on average, *ceteris paribus*, the house price increases by approximately 0.01–0.02% (depending on the group of observations). The effect is (highly) statistically significant for all four groups of observations and in line with the result from the global model (0.1%). Again, even though small in strength, this effect is still tenfold when compared to the effect of farming lot. Namely, if the farming lot increases by 1 m^2, on average, *ceteris paribus*, the house price increases approximately by 0.001–0.002% (depending on the group of observations). The latter effect is statistically significant only for the first three groups of observations (see Table 3), though in line with the result from the global model (0.001%).

Table 3: Regression coefficient estimates of the regional models.
Source: GURS (2016a); own calculations.

Coefficient estimate		Regional models			
		PN	GLO	KSD	MP
Intercept	Value	9.689000	9.875000	10.08000	9.690000
	p–value	0.0000	0.0000	0.0000	0.0000
Location (value zone)	Value	0.121800	0.154500	0.120200	0.130500
	p–value	0.0000	0.0000	0.0000	0.0000
Railway proximity	Value	–0.134000	–0.623300	–0.148700	–0.610200
	p–value	0.5723	0.0898	0.2783	0.3924
Access to gas network	Value	0.197000	–0.043620	0.124100	–
	p–value	0.0171	0.7758	0.0532	–
House age	Value	–0.010610	–0.010810	–0.010920	–0.008845
	p–value	0.0000	0.0000	0.0000	0.0000
House area	Value	0.006836	0.002790	0.003386	0.004162
	p–value	0.0000	0.0000	0.0000	0.0001
House area squared	Value	–0.000008	–0.000001	–0.000004	–0.000003
	p–value	0.0000	0.0007	0.0073	0.0199
Building lot	Value	0.000092	0.000094	0.000192	0.000233
	p–value	0.0000	0.0010	0.0000	0.0016
Farming lot	Value	0.000009	0.000012	0.000023	0.000021
	p–value	0.0000	0.0000	0.0000	0.4416
n		382	251	388	172
R^2		0.63	0.46	0.58	0.56

Just as in the case of estimating the global model, we also checked for the presence of heteroscedasticity in our regional model specifications by employing Breusch–Pagan and White tests. The results of model diagnostics revealed that by employing the regional models, we corrected for the presence of heteroscedasticity (e.g. the *p*–value of the White test statistic amounted to 0.32–0.96, depending on the group of observations, which indicates not to reject the null hypothesis of homoscedasticity), without having to apply a robust variance estimator for calculating the standard errors. In addition, the results of the spatial Chow test provided evidence that the regional (localized) models are indeed different than the global model ($p < 0.05$).

5 CONCLUDING REMARKS

The article investigates house prices in Slovenia, where various house and transaction features, such as sales terms, house attributes, maps of public infrastructure, and house value zones, are used as determinants of house prices. The novelty of the approach is to combine the geographically weighted regression with hierarchical clustering in order to improve the explanatory and forecasting power of the models. The article demonstrates how to operationalize the GWR model for house price forecasting; either at the global level or at the local (regional) level. In addition, it demonstrates how to identify geographically meaningful regions that account for relief, major road networks, other infrastructure, and presence of towns.

The results of both the global and regional models demonstrate that house and transaction features, such as location, railway proximity, access to gas network, house area, house age, and area of building and farming lot belonging to a house, statistically significant affect the house prices. Access to gas network, building lot and farming lot all exhibited a positive linear effect on the house price, whereas railway proximity and house age exhibited a negative linear effect. The effect of house area turned out to be non-linear (U-shaped), though rather small. Moreover, the effect of location on house prices can indeed be measured through the value zone in which the house is located and through the differences between the regression coefficients of the same explanatory variable across different regions. Namely, when a house is "promoted" to a higher value zone, this results in a meaningful and statistically significant increase in the house price, whereas the differences in regression coefficients between the regional models are also sound and explicable.

However, there is still room for improvement. First, more testing should be done in future research with different types of distance measures, such are road distance and travelling time from one observation point to another. Second, one should attempt to improve the kernel weighting scheme, which can have an effect on the localised regression coefficients. And third, various non-hierarchical approaches and mixed approaches to clustering might yield a better spatial distribution of regions.

Literature and references:

Bitter, C., Mulligan, G. F., Dall'erba, S. (2006). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. Journal of Geographical Systems, 9 (1), 7–27. DOI: http://doi.org/10.1007/s10109-006-0028-7

Bivand, R., Yu, D. (2006). The spgwr Package. The Comprehensive R Archive Network (CRAN). ftp://ftp.auckland.ac.nz/pub/software/CRAN/doc/packages/spgwr.pdf

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71 (2), 353–360. DOI: http://doi.org/10.1093/biomet/71.2.353

Brunsdon, C., Fotheringham, A. S., Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis, 28 (4), 281–298. DOI: http://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Casetti, E. (1972). Generating models by the expansion method: Applications to geographical research. Geographical Analysis, 4 (1), 81–91. DOI: http://doi.org/10.1111/j.1538-4632.1972.tb00458.x

Chrostek, K., Kopczewska, K. (2013). Spatial prediction models for real estate market analysis. Ekonomia, 35, 25–43.

Cirman, A., Pahor, M., Verbič, M. (2015). Determinants of time on the market in a thin real estate market. Engineering Economics, 26 (1), 4–11. DOI: http://doi.org/10.5755/j01.ee.26.1.3905

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74 (368), 829–836. DOI: http://doi.org/10.1080/01621459.1979.10481038

Du, H., Mulley, C. (2011). Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. Journal of Transport and Land Use, 5 (2), 46–59. DOI: http://dx.doi.org/10.5198/jtlu.v5i2.225

Fotheringham, A. S., Charlton, M. E., Brunsdon, C. (1998). Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. Environment and Planning A, 30 (11), 1905–1927. DOI: http://doi.org/10.1068/a301905

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical

Association, 98 (462), 387–396. DOI: http://doi.org/10.1198/016214503000170

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. Journal of Statistical Software, 63 (17), 1–50.
DOI: http://dx.doi.org/10.18637/jss.v063.i17

Guo, L., Ma, Z., Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: A case study. Canadian Journal of Forest Research, 38 (9), 2526–2534. DOI: http://doi.org/10.1139/X08-091

Hu, G., Wang, J., Feng, W. (2013). Multivariate regression modeling for home value estimates with evaluation using maximum information coefficient. In: Lee, R. (Ed.), Software engineering, artificial intelligence, networking and parallel/distributed computing, Springer, Berlin, pp. 69–81.

Košmelj, K., Breskvar Žaucer, L. (2006). Metode za razvrščanje enot v skupine; osnove in primer. Acta agriculturae Slovenica, 87 (2), 299–310.

Kupfer, J. A., Farris, C. A. (2006). Incorporating spatial non-stationarity of regression coefficients into predictive vegetation models. Landscape Ecology, 22 (6), 837–852. DOI: http://doi.org/10.1007/s10980-006-9058-2

Lesage, J. P. (1999). A family of geographically weighted regression models. In: Anselin, L. (Ed.), Advances in spatial econometrics, Springer-Verlag, Berlin, pp. 241–264.

Lu, B., Charlton, M., Fotheringham, A. S. (2011). Geographically weighted regression using a non-Euclidean distance metric with a study on London house price data. Procedia Environmental Sciences, 7, 92–97.
DOI: http://doi.org/10.1016/j.proenv.2011.07.017

Lu, J., Zhang, L. (2012). Geographically local linear mixed models for tree height-diameter relationship. Forest Science, 58 (1), 75–84.
DOI: http://doi.org/10.5849/forsci.09-123

Lu, B., Harris, P., Charlton, M., Brunsdon, C., Nakaya, T., Gollini, I. (2015). GWmodel: Geographically-weighted models (Version 1.2-5). The Comprehensive R Archive Network (CRAN).
http://cran.r-project.org/web/packages/GWmodel/index.html

Lu, B., Charlton, M., Brunsdon, C., Harris, P. (2016). The Minkowski approach for choosing the distance metric in geographically weighted regression. International Journal of Geographical Information Science, 30 (2), 351–368. DOI: http://doi.org/10.1080/13658816.2015.1087001

Murtagh, F. (2016). Hclust: Hierarchical clustering. The Comprehensive R Archive Network (CRAN).
http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

Romih, M., Bojnec, Š. (2008). Višina in oblikovanje cen rabljenih stanovanj v Sloveniji. Management, 3 (2), 165–184.

GURS. (2009). Equations and value calculation using real estate valuation models. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia.

GURS. (2016a). Property sales register data for year 2015. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia.

SGURS. (2016b). Report on the Slovenian real estate market for year 2015. Ljubljana: Surveying and Mapping Authority of the Republic of Slovenia.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58 (301), 236–244.
DOI: http://doi.org/10.1080/01621459.1963.10500845

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48 (4), 817–838.
DOI: http://doi.org/10.2307/1912934

Verbič M., Korenčan P. (2017). Cluster-based econometric analysis of house prices in Slovenia.
Geodetski vestnik, 61 (2), 231-245. DOI: 10.15292/geodetski-vestnik.2017.02.231-245

Assoc. prof. Miroslav Verbič, Ph.d
University of Ljubljana, Faculty of Economics
Kardeljeva ploščad 17, SI-1000 Ljubljana, Slovenia
e-mail: miroslav.verbic@ef.uni-lj.si

Peter Korenčan, M.Sc
Akustika Group
Vojkova cesta 58, SI-1000 Ljubljana, Slovenia
e-mail: peter.korencan@akustikagroup.si