Spatial Hedonic Modeling of Housing Prices Using Auxiliary Maps

Branislav Bajat, Milan Kilibarda, Milutin Pejović, and Mileva Samardžić Petrović

Abstract The latest applications of hedonic dwelling price models have included recent advances in spatial analysis that control for spatial dependence and heterogeneity. The study of spatial aspects of hedonic modelling pertains to spatial econometrics, which is relevant to this study because it clearly accounts for the influence and peculiarities related by space in real estate price modeling analysis.

The research reported herein introduces regression-kriging as a geostatistical method for obtaining econometric models in the analysis of real estate. The aim of this study is to compare the efficacy of regression-kriging (RK) with common regression and geographically weighted regression (GWR) methods of econometric modelling.

The spatial predictors, given as raster maps, were used as auxiliary inputs necessary for regression modeling. In addition to standard environmental predictors, some socio-economic data such as distribution, ages and income of inhabitants, were prepared in the same manner enabling their use in a GIS supported environment. Based upon global and local spatial analysis (Moran's indices), we inspected spatial pattern and heterogeneity in model residuals for all considered methods. The obtained results indicate a similar spatial pattern of model residuals for RK and GWR methods. A spatial-econometric hedonic dwelling price model was developed and estimated for the Belgrade metropolitan area based on cross-sectional and georeferenced transaction data.

Keywords Hedonic price model • Spatial econometrics • Geographically weighted regression • Regression kriging

B. Bajat (✉) • M. Kilibarda • M. Pejović • M. Samardžić Petrović
Faculty of Civil Engineering, Department of Geodesy and Geoinformatics, University of Belgrade, Bulevar kralja Aleksandra 73, 11000, Belgrade, Serbia
e-mail: bajat@grf.bg.ac.rs; kili@grf.bg.ac.rs; mpejovic@grf.bg.ac.rs; mimas@grf.bg.ac.rs

J.-C. Thill (ed.), *Spatial Analysis and Location Modeling in Urban and Regional Systems*, Advances in Geographic Information Science,
https://doi.org/10.1007/978-3-642-37896-6_5

Introduction

Nowadays, the residential market is a major component of the overall real estate market. Over time, methods and research related to this field of economics has shifted from classical econometrics to spatial econometrics (Anselin 1988).

The appearance of hedonic price models, derived mostly from Lancaster's (Lancaster 1966) consumer theory and Rosen's (Rosen 1974) model, became the milestone in econometric theory related to the real estate market. A hedonic price model decomposes the price of a good into separate components that determine the price. Basically, the hedonic equation is a regression of expenditures (rents or values) on housing characteristics of the unit that determine that rent or value. Other pricing models related to hedonic price indices include repeat-sales models (Wang and Zorn 1997) or hybrid models, which combine the elements of hedonic price and repeat-sales models (Quigley 1995). Meese and Wallace (1997) provide comprehensive research comparing the advantages and limitations of all mentioned models. The main drawback of the conventional hedonic model is that it is not capable of taking into account of spatial effects on housing prices even when locational variables are taken into consideration.

One of most important issues related to observed data in hedonic modelling is spatial autocorrelation. Basu and Thibodeau (1998) outlined two main reasons for spatial autocorrelation of housing prices. The first reason that housing prices are found to be spatially autocorrelated is that most of the dwellings in neighborhoods were built at the same time with similar structural characteristics such as dwelling size, year built, interior and exterior design features, etc. The second reason that housing prices are found to be spatially autocorrelated is a consequence of sharing the same neighborhood amenities such as proximity to public transportation, schools, markets, etc. Since hedonic house price parameters are usually estimated using ordinary least squares procedures – which assume independent observations from residuals that are spatially autocorrelated – the resulting parameter estimates often produce incorrect confidence intervals for estimated parameters and for predicted values. The importance of spatial relationships was recognized in recent hedonic studies by introducing spatial lag and spatial error models (Anselin and Bera 1998).

This problem could also be solved using spatial statistical techniques like GWR or RK that incorporate the observed spatial relationships between sample data. Geostatistics has become an essential tool in diverse environmental studies performed during the last few decades. It is imposed particularly in the field of spatial data analysis and in the prediction of numerous natural phenomena. Spatial econometrics, geostatistics, and spatial statistics share many similarities since these fields all deal spatial autocorrelation and spatial heterogeneity (Anselin 1999). Increased interest in use of geostatistics has resulted in numerous improvements and modifications that are essentially extensions to the fundamental kriging theory. Extended versions of kriging have been adopted to deal with non-normality (lognormal kriging, disjunctive kriging), while others address nonstationarity, i.e.

varying trend or drift (universal kriging, kriging with external drift, IRF-k kriging and stratified kriging) (McBratney et al. 2000). The common characteristics of all geostatistical applications are that they were initially used for spatial modelling of diverse natural (i.e. non-anthropogenic) phenomena. Although hedonic regression models (Can and Megbolugbe 1997; Kim et al. 2003; Osland 2010) prevail in real estate appraisal applications, the role of geostatistics has increased in importance recently (Dubin 1998; Yoo and Kyriakidis 2009; Fernández-Avilés et al. 2012).

Most of those spatial statistical techniques are already available in most geographical information systems (GIS) operational environments. The use of GIS technology in spatial econometrics studies started in the mid-1990s. Soon after, the advantages of using GIS applications for hedonic price modelling were recognized in a number of studies (Lake et al. 1998; Anselin 1998; Lovett and Bateman 2001).

In this paper, a geostatistical method regression-kriging (RK) is presented as a method for spatial prediction and mapping of housing prices. Although the RK technique has not been extensively used in hedonic price modeling, there are certain examples where forms of RK were applied under different names. In the literature, the terms used for geostatistical methodologies can be confusing due to the different terms used for the same or very similar techniques (Hengl 2009). Yoo and Kyriakidis (2009) used the term area-to-point kriging with external drift (A2PKED). Chica-Olmo (2007) tested the performance of two co-kriging methods for prediction of housing location prices, in which the authors used a heterotropic version of co-kriging that is very similar to RK.

Usually, most input data for hedonic modelling is acquired from multiple listings databases. In this study, both transaction data and explanatory data are organized as auxiliary maps in a raster format. The majority of required spatial data is already available in GIS formats, thereby minimizing the effort required for pre-processing. The results of spatial prediction are produced as a raster GIS layer in the same resolution as input maps; the resulting map may be of interest to appraisers, real-estate companies and governmental agencies, as it provides an overview of location prices.

The performance of an alternative method, Geographically Weighted Regression (GWR), was also examined in this study; this method is intended for the local analysis of relationships in multivariate data sets. The comparison of the proposed approach using GWR reveals RK as the model of choice for spatial prediction of housing prices in combination with auxiliary data.

The results of spatial prediction are exported as a standard raster GIS layer, or in HTML format enabling simple creation of rich interactive Web-based maps. The user-created maps can assist in facilitating communication between appraisers, real-estate companies, governmental agencies, and other interested parties.

The overall objectives of this study were: (1) to compare two alternative regression techniques (OLS in implicit and error model form and GWR) against the RK geostatistical model for evaluating dwelling prices; (2) to compare the performance of all models using global statistics; and (3) to evaluate the performance of all models in terms of spatial distribution and clustering of residuals using local indicators of spatial association (LISA).

The next section contains a brief description of the theoretical foundations of hedonic price models, geographically weighted regression and regression-kriging methodologies, as well as some details about the *R* software environment used in this study. Included in this study is a concise representation of real estate market characteristics in the city of Belgrade as well as the description of the transaction and auxiliary data layers used. The important issues related to multicollinearity and attribute selection are discussed in the following section, along with detailed explanatory data analyses and mapping. The final section concludes the study.

Methodology

Hedonic Price Models

The basic hedonic price function can be represented as (Can and Megbolugbe 1997):

$$\mathrm{Y} = f\left(\mathrm{S}\beta, \mathrm{N}\gamma\right) + \varepsilon \tag{1}$$

where **Y** is a vector of observed housing values; **S** is a matrix of structural characteristics of properties; **N** is a matrix of neighborhood characteristics, including measures of socioeconomic conditions, environmental amenities and public accommodations for area residents; β and γ are the parameter vectors corresponding to **S** and **N**; and ε is vector of random error terms. The basic form of hedonic regression assumes that each parameter is fixed in space, which means that each identified attribute has the same intrinsic contribution throughout the study area. The given formula can be expressed like a common regression function:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \tag{2}$$

where $\mathrm{Y}_{n\times 1}$ represents vector of observed sale prices of n dwellings, $\mathrm{X}_{n\times k}$ is a vector of k explanatory variables characterizing housing units. $\beta_{k\times 1}$ is a vector of unknown coefficients and $\varepsilon_{n\times 1}$ is a vector representing the error term. By using ordinary least squares (OLS), the unknown coefficients are solved as:

$$\widehat{\beta} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{3}$$

The basic assumption for OLS usage is the indepedence of observations, which is often violated due to spatial autocorrelation in data, leading to a biased estimation of standard errors of model parameters and misleading significance tests. The above given regression formula has particular modified versions that are often used for house price modelling: the spatial-lag model (also known as spatial autoregressive model) and spatial error model (Dubin 1988), (Kim et al. 2003), (Osland 2010):

$$\begin{aligned} &\text{Spatial lag model :} && \mathrm{Y} = \mathrm{X}\beta + \phi\mathrm{W}\mathrm{Y} + \varepsilon \\ &\text{Spatial error model :} && \mathrm{Y} = \mathrm{X}\beta + \xi + \lambda\mathrm{W}\varepsilon \end{aligned} \tag{4}$$

where ε is the vector of errors terms and W represents the weights matrix that specifies the assumed spatial structure or connections between the observations. The elements of the weights matrix can be based upon contiguity (i.e., shared borders) or distance. The parameter φ is often referred to as the spatial correlation parameter giving the intensity of the dependence between neighboring prices. The term ξ represents random error and λ is a spatial autoregressive parameter.

If data exhibits a spatial lag process the target variable is affected by the values of the target variables in nearby places. The OLS hedonic model omits φwY, and this becomes part of the error which leads to biased parameters estimates (Anselin 1988). If spatial lag is in true functional form (Anselin and Bera 1998), then parameter estimates should be inefficient as well. Nevertheless, some studies showed that although spatially correlated errors are presented in data, non-spatial hedonic models (implicit OLS) may provide results useful for policy analysis (Mueller and Loomis 2008).

Geographically Weighted Regression

The geographically weighted regression is a relatively new method used in spatial modeling and was developed as an alternative format for spatial analysis that is local rather than global in its analytical design (Fotheringham et al. 2002). Although this method is useful in a wide range of applications, its widest practical application still is in the mass assessment of real property (Crespo et al. 2007; Yrigoyen et al. 2008; Hanink et al. 2010). Increasing application of this technique is made possible by GIS database processing tools and publicly available databases on the Internet. GWR represents the extension of a conventional multiple regression framework by addressing the issue of non-stationary processes (Fotheringham et al. 2002):

$$y_i = \beta_0 (u_i, v_i) + \sum_{k=1}^{m} \beta_k (u_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{5}$$

where (u_i,v_i) are the coordinates for i-th point; $\beta_k(u_i,v_i)$ are the realizations of continuous function $\beta_k(u,v)$ at the same location; $x_{i1}, x_{i2}, \ldots x_{im}$ are the explanatory variables at point i; and ε_i is the error term.

$$\widehat{\beta}(i) = \left(\mathrm{X}^T \mathrm{W}(i) \mathrm{X}\right)^{-1} \mathrm{X}^T \mathrm{W}(i) \mathrm{y} \tag{6}$$

W(i) is a matrix of weights for particular location i, such that observations nearer to i are given greater weight than observations further away.

$$\mathrm{W}(i) = diag\,[w_{i1}, w_{i2}, \ldots, w_{in}] \tag{7}$$

w_{ij} is the weight related to data point j for the estimate of the local parameters at location i. Several types of parameterized weight functions may be used. A common choice also used here is the Gaussian distance-decay curve has the form:

$$w_{ij} = \exp\left[-\frac{d_{ij}^2}{2b^2}\right] \tag{8}$$

where d_{ij} is a distance between location i and location j, while parameter b is a range to be determined. The weight's value would decay gradually with distance to the extent that when $d_{ij} = b$, the weight reaches the value of 0.5. In the event that the spatial distribution of sampled variables is spatially homogeneous, this parameter is taken as a constant value. However, a spatially variable (adaptive) parameter of the range should be used in the event that the spatial distribution of the variables is heterogeneous.

To calculate the parameters associated with a weighting function, such as the bandwidth and *Nth* nearest neighbors considered, the GWR methodology utilizes a calibration process. This calculates the parameter so as to form an appropriate trade-off between bias and standard error in the prediction of the overall model. Commonly used approaches include minimizing the cross-validation scores (CV) or Akaike Information Criterion (AIC_c) (Fotheringham et al. 2002). The optimal values for b and N reported here were obtained by minimizing the cross-validation scores.

Regression-Kriging

Regression-kriging (RK) is a geostatistical technique that combines the regression of the target variable on explanatory variables with kriging of the regression residuals. In the literature, this interpolation technique is termed as kriging with external drift (KED). Let the measured values of the target variables be symbolized as $Y(s_i)$, $i = 1, \ldots n$, where s_i represents spatial location and n number of realized measurements. The system of equations from which the estimated values of target variables $\hat{Y}(s_0)$ are obtained is:

$$\begin{aligned} &\hat{Y}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \\ &\hat{Y}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) \cdot e(s_i); \\ &q_0(s_0) = 1 \end{aligned} \tag{9}$$

where $\hat{m}(s_0)$ is the fitted deterministic part; $\hat{e}(s_0)$ is the interpolated residual; $\hat{\beta}_k$ is estimated deterministic model coefficient; and w_i represents ordinary kriging weights resolved by the spatial structure of residuals $e(s_i)$ (Hengl 2009). The essential difference between RK and KED is explained as follows: while KED

weights are solved within extended matrix taking into consideration trend and residuals at the same time, the RK drift model coefficients are computed separately and the residuals interpolated by ordinary kriging (OK) are summed back to the drift estimates using simple summing of predicted drift and residual surfaces. Despite the differences in the computational steps used, the resulting predictions and prediction variances are the same, given the same point set, auxiliary variables, regression functional form, and regression fitting method (Hengl et al. 2007). Regression coefficients $\hat{\beta}_k$ could be obtained by different fitting methods such as ordinary least squares (OLS) or generalized least squares (GLS) which is more often recommended:

$$\hat{\beta}_{GLS} = \left(\mathrm{q}^{\mathrm{T}} \cdot \mathrm{C}^{-1} \cdot \mathrm{q}\right)^{-1} \cdot \mathrm{q}^{\mathrm{T}} \cdot \mathrm{C}^{-1} \cdot \mathrm{Y} \tag{10}$$

where $\hat{\beta}_{GLS}$ is the vector of estimated regression coefficients; **C** is the covariance matrix of residuals; **q** is the matrix of predictors at the sampling location; and **Y** is the vector of measured values of target variable. The estimated $\hat{\beta}_{GLS}$ coefficients basically present a special case of geographically weighted regression coefficients. The predicted variable value $\hat{Y}(s_0)$ at the location s_0, obtained by regression-kriging is commonly written in matrix notation:

$$\hat{Y}_{RK}(\mathrm{s}_0) = \mathrm{q}_0^{\mathrm{T}} \cdot \hat{\beta}_{GLS} + \lambda_0^{\mathrm{T}} \cdot \left(y - \mathrm{q} \cdot \hat{\beta}_{GLS}\right) \tag{11}$$

where $\mathbf{q_0}$ is the vector of $p + 1$ predictors and λ_0 is the vector of n kriging weights used for interpolation of residuals.

RK explicitly separates trend estimation from residual interpolation, thereby allowing the use of arbitrarily complex forms of regression rather than using the simple linear techniques that can be used with KED. Besides, RK allows the separate interpretation of the two interpolated components, which reinforces the advantage of the RK approach (Hengl et al. 2007).

Spatial Residual Analysis

We used global and local analysis methods (i.e., global and local Moran's I indices) to evaluate spatial distribution and heterogeneity in model residuals. Spatial autocorrelation of residuals was calculated by *Moran's I* statistic (O'Sullivan and Unwin 2003):

$$MI = \frac{n}{\sum_{i=1}^{n}(e_i - \overline{e})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(h)\,(e_i - \overline{e})\left(e_j - \overline{e}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{12}$$

Where e_i and e_j are calculated values of the residuals at the locations i and j; the term $\bar{e}$ is a mean value; and w_{ij} is the spatial weight within a given distance or bandwidth that was determined as $h = 750$ m, according to the variogram of the OLS model residuals (Fig. 5). The weighted matrix was row standardized in order to overcome the problem of uneven spatial distribution of observations, which causes up-weighted values for locations with more neighbors. The value of the Moran's I statistic ranges from near $+1$ indicating clustering of the e values to near -1 indicating dispersed pattern of the e values. In the Global Moran's I statistic, the results of the analysis are always interpreted within the context of its null hypothesis, which states that the variable (residuals in our case) being analyzed is randomly distributed among the locations in our study area; or better said, the spatial processes promoting the observed pattern of values is random chance. The results of Moran's I statistic with significant p-values and positive Z-scores indicates spatially clustered datasets. However, at the same time, significant p-values and negative Z-scores depict that spatial pattern is more spatially dispersed than what would be expected to result from random spatial processes.

Local Indicators of Spatial Association (LISA) (Anselin 1995) that are based on the local Moran's I test enable the assessment of significant local spatial clustering around an individual location- thereby providing details of (1) the degree of spatial clustering; (2) an estimate of detailed variations of clustering in the locally defined area; and (3) identification of the locations of the spatial clusters. The local version of Moran's I at location i is given by:

$$MI_i = \frac{(e_i - \bar{e})}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{ij}(h)\left(e_j - \bar{e}\right) \tag{13}$$

where w_{ij} is the spatial weight within a given distance or bandwidth (h), as stated in Eq. 12 and S_i^2 is calculated as:

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n} w_{ij}}{n-1} - \bar{e}^2 \tag{14}$$

This local indicator represents a disaggregated measure of autocorrelation that depicts the extent to which the residuals for particular areal locations are similar to, or differ from, neighboring locations. The *Local Moran's I* statistic is used to detect possible non-stationarity of the data – i.e., a clustered pattern – among sub-regions. A positive local MI_i indicates a cluster of similar residual values of the same sign around i, while a negative MI_i indicates that the value of the residual at location i has a sign opposite that of its neighbors. If values for e_i and e_j are close to $\bar{e}$, then local MI_i indicates the absence of spatial autocorrelation.

Case Study: Belgrade Metropolitan Area

Belgrade, the capital of the Republic of Serbia, is situated at the confluence of the Sava and Danube Rivers. The administrative boundary encompasses a 3223 km^2 area of nearly 2 million inhabitants. Its territory is divided into 17 municipalities, comprised of 157 settlements. The urban core, an area of 360 km^2, includes the 10 urban municipalities which constitute the study area for this research. According to official census data records, there are approximately 470,000 households with 1,300,000 inhabitants in the study area, which represents metropolitan area of the city (Statistical Office of the Republic of Serbia 2011).

Due to political issues, and the fact that most apartments in large cities were under social ownership, the number of individual transactions was negligible prior to privatization. Between 1991 and 2000, most housing units were privatized. The housing market in Serbia started developing rapidly and housing prices spiked after privatization in the 1990s. Housing prices peaked in 2008, just prior to the present financial crisis. As of the time of publication, this market remains depressed and housing prices are falling along with declining construction activity.

According to local property experts, the housing prices are expected to fall by another 20% in the year 2013. Average housing prices in Belgrade vary considerably depending upon apartment location and structural characteristics; in some cases, average prices may be 60% higher in one Serbian municipality or town, as compared to another (Cvijanović 2006). However, housing studies regarding Belgrade and other cities in Serbia are limited because of the short history of the free housing market.

The original dataset used in this study consists of 747 records of apartment transactions referring to real estate sales in the year 2010. The dataset used was provided by several real estate trading companies because a unique database of real estate transactions has not been compiled for the Belgrade area. Selected transaction records include total transaction value (EUR), covered flats size (m^2) and addresses. Since additional information regarding internal living space, age of building, and parking amenities was available only for some records, these attributes were not included in the analysis. A geographic information system (GIS) was used to match street addresses of the transactions with the official dataset of building geographic coordinates in order to geocode observations into the study area (Fig. 1).

Target Variable

Housing price (expressed in EUR/m^2) is the target variable to be spatially evaluated. A study area map with corresponding housing price ranges is shown in Fig. 1 and the descriptive statistics of housing price are displayed in Fig. 2. The distribution is positively skewed, with a mean of EUR 1633 EUR/m^2.

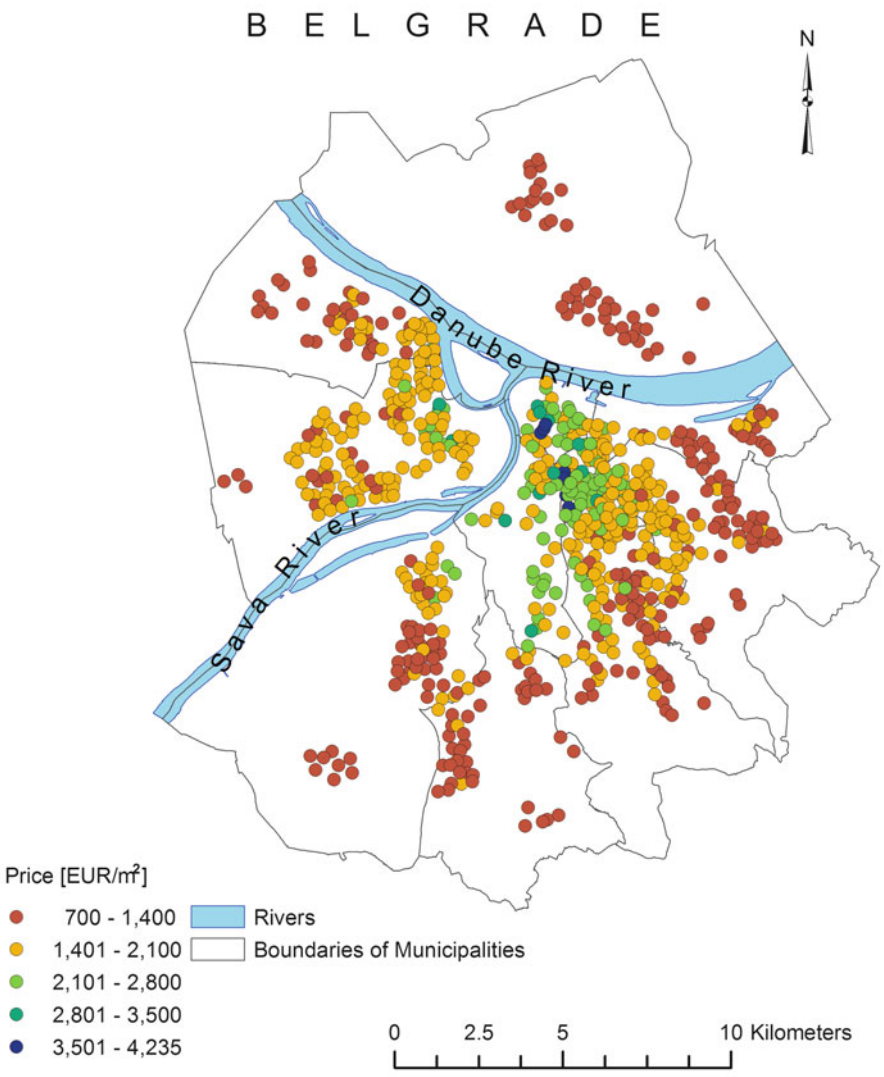

Fig. 1 Locations and values of observed transactions in Belgrade in the year 2010

Mapped housing price values (Fig. 1) indicate that the most expensive apartments are located in the city center, while the prices are lower in peripheral areas. The variability of the prices is more pronounced in the central urban core, where the price depends upon the specific location of the building, the quality of interior infrastructure and other amenities (Cvijanović 2006).

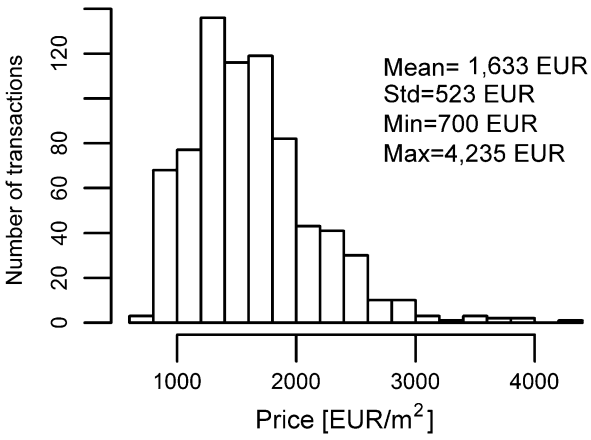

Fig. 2 Histogram of observed housing prices

Exogenous Variables

In applying the hedonic price model to the real estate market, the determinants of housing prices can be divided into four groups (Lake et al. 1998): (1) structural variables (e.g., age, the number of rooms in each house, etc.); (2) accessibility variables (e.g., the proximity of schools, bus routes, railway stations, shops, parks, and the Central Business District); (3) neighborhood variables (e.g., local unemployment rates); (4) environmental variables (e.g., road noise and visibility impact). The accessibility characteristics primarily consist of site-related factors.

In this study we were confined to predictors that incorporate accessibility, neighborhood and environmental variables (Table 1). The explanatory variables referring to accessibility and environment could be considered as spatial determinants referred to as the distance variables (Koramaz and Dokmeci 2012). They were arranged as input maps/grids of 20 m resolution by using a proximity function within the SAGA (System for Automated Geoscientific Analyses) GIS environment (http://www.saga-gis.org/). The values assigned to grid cells are calculated Euclidean distances between each cell and the input features (roads, schools, parks, etc.); these values are included in each cell in the grid map. Researchers have previously used travel times to the central business district and to highways, shopping centers, schools, and universities, instead of Euclidean distances to these features, in order to compute accessibility measures for hedonic models (Des Rosiers et al. 2000; Adair et al. 2000). Two neighborhood variables (illiteracy and income) are based upon census data (Statistical Office of the Republic of Serbia 2011). The illiteracy layer/map was generated as a factorial variable referenced to each municipality, where each cell represents the proportion of illiterate inhabitants in a particular municipality in regard to the whole city of Belgrade. The income variable represents the average income in each municipality so that every grid cell within each municipality area has the same value.

Table 1 The list of explanatory variables used in study with corresponding VIF (variance inflation factor) values

Variables	Description	Type	VIF
dist_ Airport	Prox. (Euclidean distance) to airport	Accessibility	4.138
dist_ Highway	Prox. to highway	Environmental	7.022*
dist_ Culture	Prox. to museums, theatres	Accessibility	4.708
dist_ Main road	Prox. to main roads/ boulevards	Environmental	2.951
dist_ Sciences facilit	Prox. to University/science facilities	Accessibility	6.018*
dist_ Schools	Prox. to elementary/high schools	Accessibility	2.045
dist_ Parks	Prox. to parks/playgrounds	Accessibility	8.110*
dist_ Market	Prox. to green markets	Accessibility	3.201
dist_ Industry	Prox. to industrial objects	Environmental	1.567
dist_ River	Prox. to river banks	Environmental	7.264*
dist_ Recreation	Prox. to big green areas/forest	Accessibility	1.458
dist_ Sport	Prox. to sport stadiums	Accessibility	1.903
dist_ public transport	Prox. to station of public transport	Accessibility	1.470
dist_ Shopping	Prox. to Shopping centers	Accessibility	9.458*
dist_ Main streets	Prox. to main streets	Accessibility	3.292
dist_ Religious	Prox. to Religious facility	Accessibility	1.498
dist_ Kindergarten	Prox. to kindergarten	Accessibility	1.837
dist_ Hospital	Prox. to ambulance/hospitals	Accessibility	1.813
dist_ Railway	Prox. to railway	Environmental	3.985
Illiteracy	Percentage of inhabitants who are illiterate	Neighborhood	2.688
Income	Average income in municipality	Neighborhood	5.724*

*VIF > 5; variable indicates high multicollinearity

R Language Environment

The *R* open source language (Development Core Team 2008) contains the base system that allows statistical computation, linear algebra computation, graphics creation and the like. Most of the computations related to spatial autocorrelation estimation and model testing are utilized through the *spdep* package (Bivand et al. 2008). It includes a number of features such as tools for the creation of spatial weights matrix, a collection of tests for spatial autocorrelation, including global and local Moran's I and and Getis/Ord G statistics, and functions for estimating spatial simultaneous autoregressive (SAR) lag and error models. There is also a set of developed *R* packages that are especially interesting for geoscientists.

All utilized methods were implemented using the open-source R statistical computing environment with *gstat* and *spgwr* packages (Bivand et al. 2008) intended for modelling and prediction, as well as the *sp.* package which provides classes and methods for dealing with spatial data in *R* (Pebesma 2004). The results obtained in *R* can easily be converted into any of the standard GIS formats, which enables the manipulation and analysis of the results in commercial GIS packages afterwards.

There are also several open-source and commercially available software packages with associated GWR methods. Unfortunately, GWR is a time-consuming computational procedure, especially in the case of large data sets. However, it is possible to solve this problem by using grid computing (Harris et al. 2010). The *spgwr* package used in *R* environment has also been adopted for use on grid based systems.

The recently developed *R* package *plotGoogleMaps (*Kilibarda and Bajat 2012*)*, designed to automatically create web maps by combining users' data and Google Maps layers as a base map, map was also used in this study to improve insight into predicted housing price layers.

Model Fitting and Evaluation

Before regression analyses were performed, an indicator of multicollinearity between exogenous variables was examined. The variance inflation factor (VIF) test (Fox 2008) indicates the presence of multicollinearity between predictors (Table 1). Principal components analysis (PCA) is often used with the aim of transforming a dataset with many correlated variables into a dataset consisting of a smaller number of uncorrelated variables, known as principal components (PCs) (Lake et al. 1998). PCA can assist when multicollinearity exists between predictor variables and can also assist in collapsing a large set of variables into a more efficient set of uncorrelated components. However, the main drawback of using PCA is that the newly generated components complicate the interpretation of the influence of the original variables.

After the performing PCA upon the set of explanatory variables, 21 PC predictors were derived. Because the predictors are now known to be independent, we can reduce their number by using step-wise regression (Draper and Smith 1981) based on AIC or *t-statistics* (Table 2).

By looking at the specific *t*-values of coefficients, we can infer which predictors are significant and useful in further analysis. Because the target variable is positively skewed, we used lognormal transformation to improve linearity. The results obtained show that eight predictors are highly significant, seven are marginally significant and six predictors can be removed from the list.

In order to check for the presence of spatial correlation we performed the Lagrange Multiplier test (Anselin 1988; Bivand et al. 2008).This test is often used in spatial econometrics as a supplement to Moran's I and is designed to determine whether the input data generate a spatial lag (LMlag) or spatial error (LMerr) model (Eq. 4). In addition, the robust LM error (RLMerr) statistics tests for error dependence in the possible presence of a missing lagged dependent variable while its counterpart, the robust RLMlag statistics test, works the other way round. Based on the obtained test results (Table 3), it appears that the spatial correlation problem is more of the "spatial error type" rather than of the "spatial lag type".

Nevertheless, model fitting statistics indicates that the introduction of the OLS modified ,spatial error data "version (OLS.ERR) could not be of benefit to improve

Table 2 Results of step-wise regression analysis of predictors

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	Signif. code
(Intercept)	7.05202	0.02392	−294.85	<2.00E-16	***
PC1	0.09051	0.00875	10.35	<2.00E-16	***
PC2	0.04272	0.00837	5.1	4.30E-07	***
PC3	0.03479	0.00679	5.12	3.90E-07	***
PC5	−0.04602	0.01229	−3.75	0.00019	***
PC6	0.02117	0.01214	1.74	0.08148	.
PC8	0.03983	0.01304	3.05	0.00234	**
PC9	−0.07971	0.02301	−3.46	0.00056	***
PC10	−0.17372	0.01599	−10.87	<2.00E-16	***
PC11	0.0473	0.02013	2.35	0.01905	*
PC12	−0.07523	0.01819	−4.14	4.00E-05	***
PC13	0.03499	0.01782	1.96	0.05004	.
PC15	−0.05511	0.02258	−2.44	0.01488	*
PC17	0.18078	0.02624	6.89	1.20E-11	***
PC19	0.0357	0.02291	1.56	0.11954	
PC21	0.09761	0.04082	2.39	0.01705	*

[a]Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '•' 0.1 '' 1,
[b]Residual standard error: 0.184 on 731 degrees of freedom,
[c]Multiple R-squared: 0.646, Adjusted R-squared: 0.636,
[d]F-statistic: 88.4 on 15 and 731 DF, p-value: < 2.2e-16.

Table 3 Lagrange Multiplier test on OLS residuals

Test	LMerr	RLMerr	LMlag	RLMlag
Statistic	65.43107	60.40537	6.161991	1.136293
p-value	5.55E-16	7.77E-15	0.013052	0.286437

Table 4 Model fitting statistics for utilized modeling techniques

Model	R^2	SSE	RMSE	F-Test[a]	p.value
OLS	0.56	89077395	349.0801		
OLS.ERR	0.56	90550113	351.9539	0.986253	0.850
RK	0.63	75277050	320.9021	1.182	0.022
GWR	0.61	79413549	329.1511	1.123563	0.111

[a] Hypothesis test for testing the improvement of model fitting over OLS

the performance of implicit OLS (Table 4). Both models had the same R^2 values (0.56), while the error sum of squares (SSE) and the root mean square error (RMSE) values as well as the Fisher statistics for homogeneity of variance indicate better model fitting performance of implicit OLS over the spatial error data OLS model. For that reason we left the implicit OLS model as the benchmark model in further analysis.

Even though quantitative evaluation measures demonstrate the similar performance for RK and GWR methods, the RK method provides slightly better

performance. Both methods achieved improvement in performance, this performance was not significantly improved over the results achieved via standard OLS hedonic modeling.

Results and Discussions

The model's residuals were obtained through the leave one out cross-validation process of all referred spatial models. Table 5 and Fig. 3 show that the model's residuals have similar characteristics. All models demonstrate that the dataset is characterized by positive skewness and large positive kurtosis. Figure 4 shows strong linear relationship within OLS, RK and GWR residuals.

Table 5 Characteristics of model residuals for utilized modeling techniques

	Mean	Std	Skewness	Kurtosis	Min	Median	Max
OLS	29.758	344.267	1.393	3.761	−733.956	−18.767	2020.549
RK	25.219	316.655	1.220	3.547	−885.708	−8.546	1874.712
GWR	31.070	324.786	1.167	3.551	−892.572	−6.445	1915.185

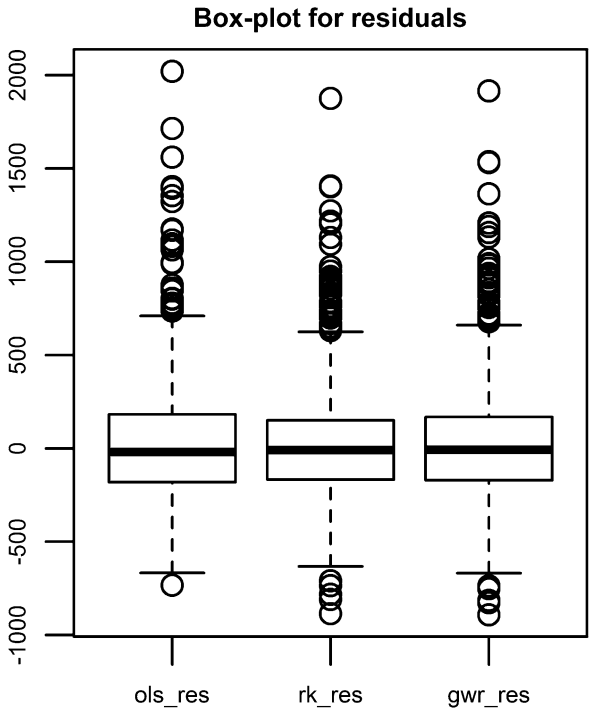

Fig. 3 Box plot of the model's residuals

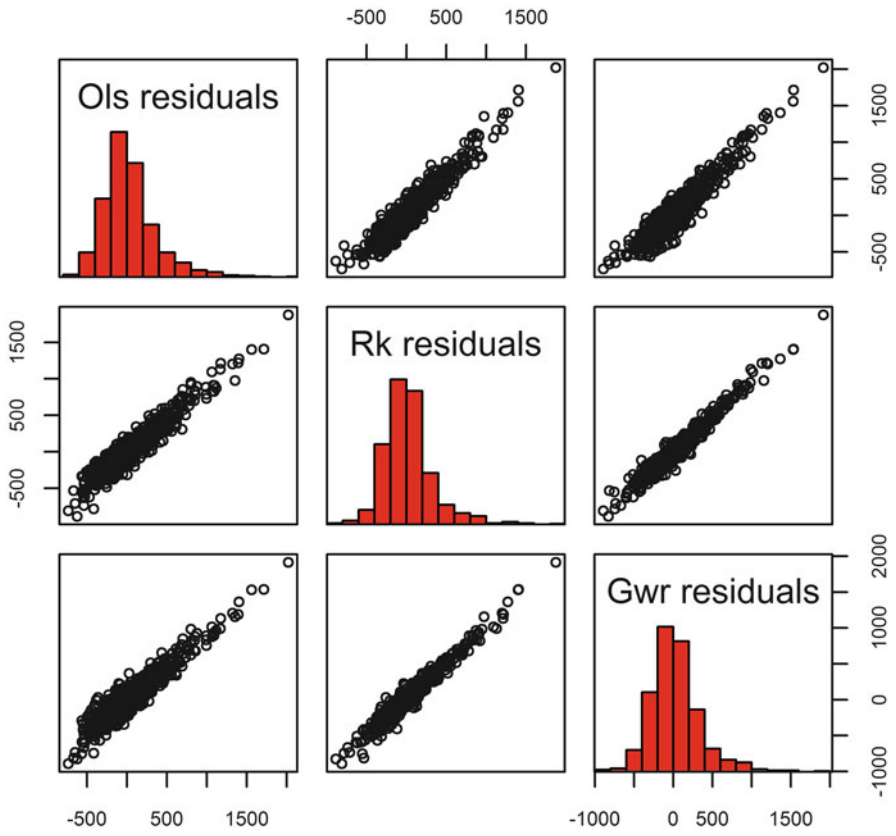

Fig. 4 Matrix plot of residual relationship between referred models

A semivariogram of OLS residuals was created in order to examine the spatial distribution of the model's residuals. Specifically, the semivariogram determines the range of the residuals, i.e. the distance at which the spatial correlations between observations fall to zero (Fig. 5). Based upon this information, we built a spatial (row standardized) weights matrix W (Eqs. 12 and 13) in which all buildings that are located within 750 m from one another are considered as neighbors. At same time the experimental variogram of GWR residuals exhibits practically no autocorrelation.

The global Moran's *I* (Eq. 12) was computed for all three models (Table 6). The global *MI* for the OLS residuals is significantly positive (Z.value > 1.96) indicating that model residuals tend to be similar across the space. In contrast, the global *MI* for RK's and GWR's are significantly negative (Z.value < −1.96), which means their residuals are clustered with dissimilar values (positive are surrounded with negative, and vice versa).

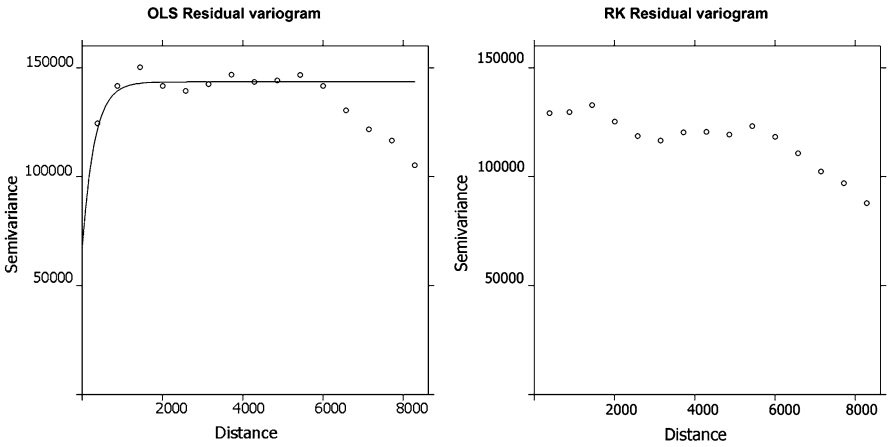

Fig. 5 Semivariograms of the OLS (left) and GWR (right) residuals

Table 6 Global Moran (*MI*) for utilized modeling techniques

Model	MI	Z.value[a]	Z.value[b]
OLS	0.111	6.484	6.468
RK	−0.050	−2.831	−2.824
GWR	−0.041	−2.274	−2.269

[a]Standard normal test based on randomization assumption
[b]Standard normal test based on normal assumption

Table 7 Local Moran indexes (MI_i) for utilized modeling techniques

	Mean	Std	Skewness	Kurtosis	Min	Median	Max
OLS	0.112	0.531	2.534	20.293	−2.702	0.021	4.570
RK	−0.050	0.333	−1.762	29.343	−3.267	−0.008	2.313
GWR	−0.041	0.345	−0.594	26.449	−2.993	−0.007	2.747

The local indicator of spatial association MI_i was used to determine "hot spots" (positive autocorrelation or similarity) and "cold spots" (negative autocorrelation or dissimilarity) of residual values of all utilized models (Anselin 1995). Table 7 indicates that local MI_i of RK and GWR residuals have similar characteristics while MI_i of OLS residuals are considerably different. Obviously, RK and GWR produced more frequently negative local MI_i with stronger linear relationship in contrast to OLS (Fig. 6).

Similar patterns were obtained for calculated Z-values of the local Moran (MI_i) (Table 8).

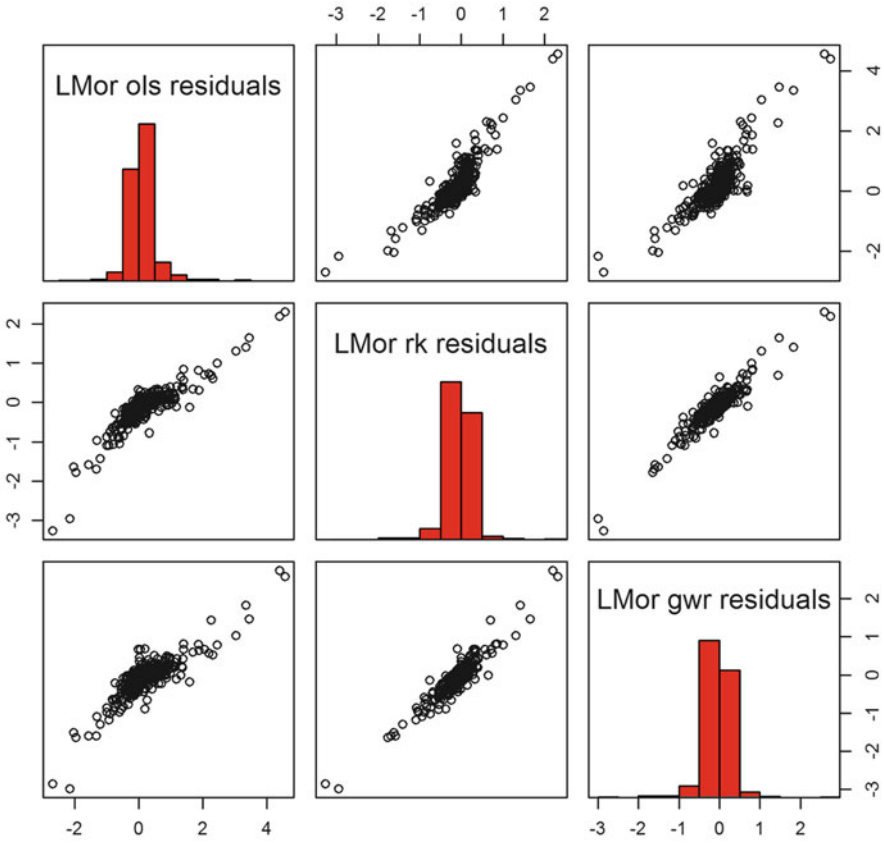

Fig. 6 Matrix plot of local Moran's indexes relationship between referred models

The local Z-values for the local MI_i were also evaluated for the significance level 0.05 ($Z_{\alpha=}1.96$). Table 9 indicates that RK and GWR produced a similar percentage of significant Z-values, where higher percentages among significant values are negative (76% and 71%, respectively). These results indicate the clustering of the model residuals, whereby a large residual tends to be surrounded by smaller neighboring residuals and vice versa when a small residual is surrounded by larger ones. On the other hand the majority of significant positive Z-values indicate opposite behavior of the OLS model (71%) – i.e., it generates more clusters of either positive or negative model residuals. That means that OLS produced sub-areas with underestimated (positive residuals) or overestimated (negative residuals) prediction.

Table 8 Z-values of the local Moran (MI_i) for utilized modeling techniques

	Mean	Std	Skewness	Kurtosis	Min	Median	Max
OLS	0.374	1.841	2.563	18.48	−7.549	0.072	15.324
RK	−0.183	1.106	−1.257	22.32	−10.329	−0.023	7.758
GWR	−0.139	1.166	−0.424	23.29	−10.481	−0.018	8.777

Table 9 Significant Z-values for local Moran's indexes

	n = 747	Among the significant Z-values	
Model	No. of sign. $\|Z\| \geq 1.96$ (%)	$Z \leq -1.96$ (%)	$Z \geq 1.96$ (%)
OLS	94 (12)	27 (29)	67 (71)
RK	45 (6)	34 (76)	11 (24)
GWR	51 (7)	36(71)	15 (29)

Number in parenthesis is the percentage

The visual inspection of the mapped residuals may give us more detailed insight into the performance of the model in the case of particular observations (Figs. 7, 8 and 9). Spatial patterns of Z-values in terms of size, sign and clustering are apparently similar for GWR and RK.

Generated local *MI* Z-values can also be completely mapped by *plotGoogleMaps* package (Kilibarda and Bajat 2012) in HTML format (available at http://osgl.grf.bg.ac.rs/en/materials/hedonic/), which has become a standard medium for cartographic communication.

Finally, based upon the estimated model parameters, we produced raster maps that depict predictions for dwelling prices over the whole case study area. The raster maps were generated in the same resolution (20 m) as input maps/grids (Figs. 10 and 11). Visual inspection of spatially predicted maps clearly emphasizes the limitation of the GWR method (Fig. 10). At first glance, the artefacts (areas with high price values) generated at the peripheral areas (north-western and northern) are easily distinguished on the GWR map. Those artefacts are caused by the spatial pattern of observed transactions (Fig. 1) because most of the transactions are grouped in the inner part of the city and peripheral parts are apparently lacking data. The peripheral parts of the case study area did not have any data on transactions and therefore the predicted values are a result of extrapolation.

Conclusions

This chapter demonstrates how GWR and RK techniques- which are already widely used for spatial prediction in various environmental studies – be useful tools for the prediction and mapping of housing prices. The primary objective of this research was to compare the performance of recently introduced spatial modelling methods in the hedonic prediction of residential property values in city of

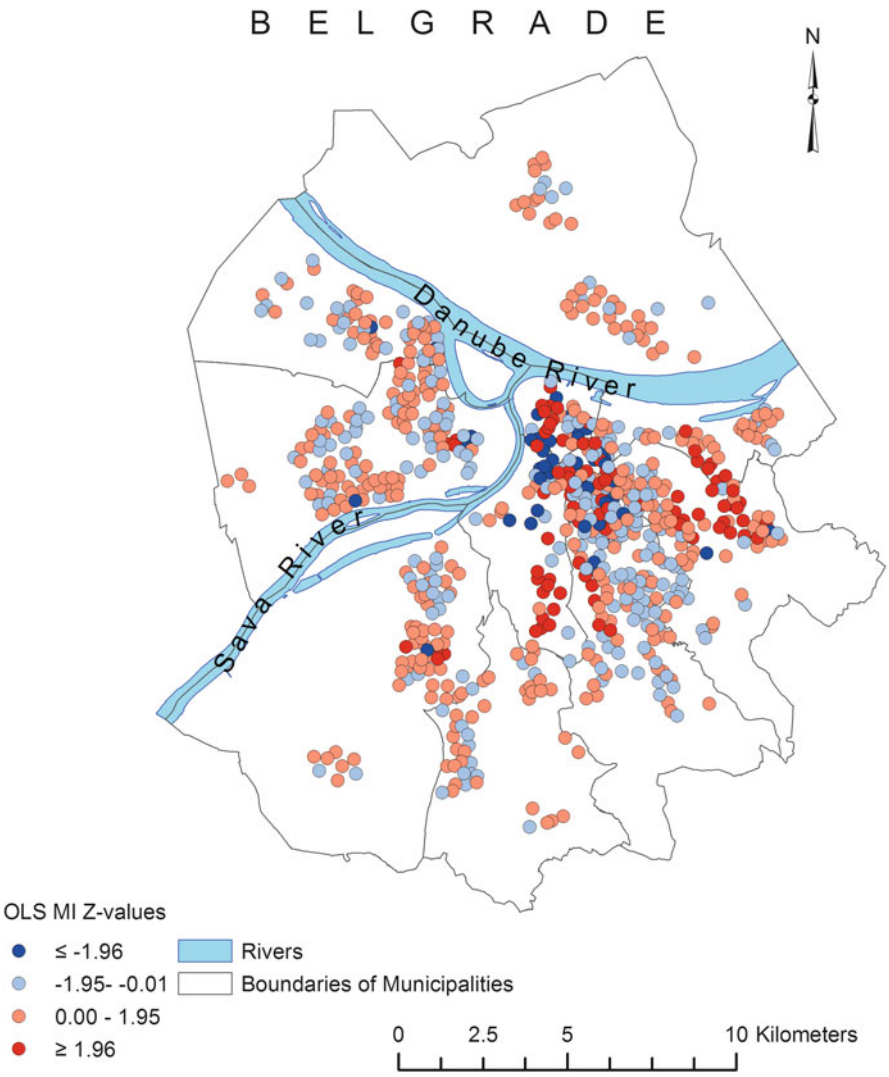

Fig. 7 Mapped local MI values of OLS residuals

Belgrade, based upon examination of spatial distribution and heterogeneity in model residuals. Spatial models including geographically weighted regression (GWR) and regression-kriging (RK) were estimated in contrast to standard hedonic price modelling based upon an ordinary least squares (OLS) technique. Although similar values for quantitative evaluation measures were obtained for by each model, spatial patterns of residuals for the RK and GWR models were found to be different in contrast to the OLS model. The values erroneously predicted by the GWR

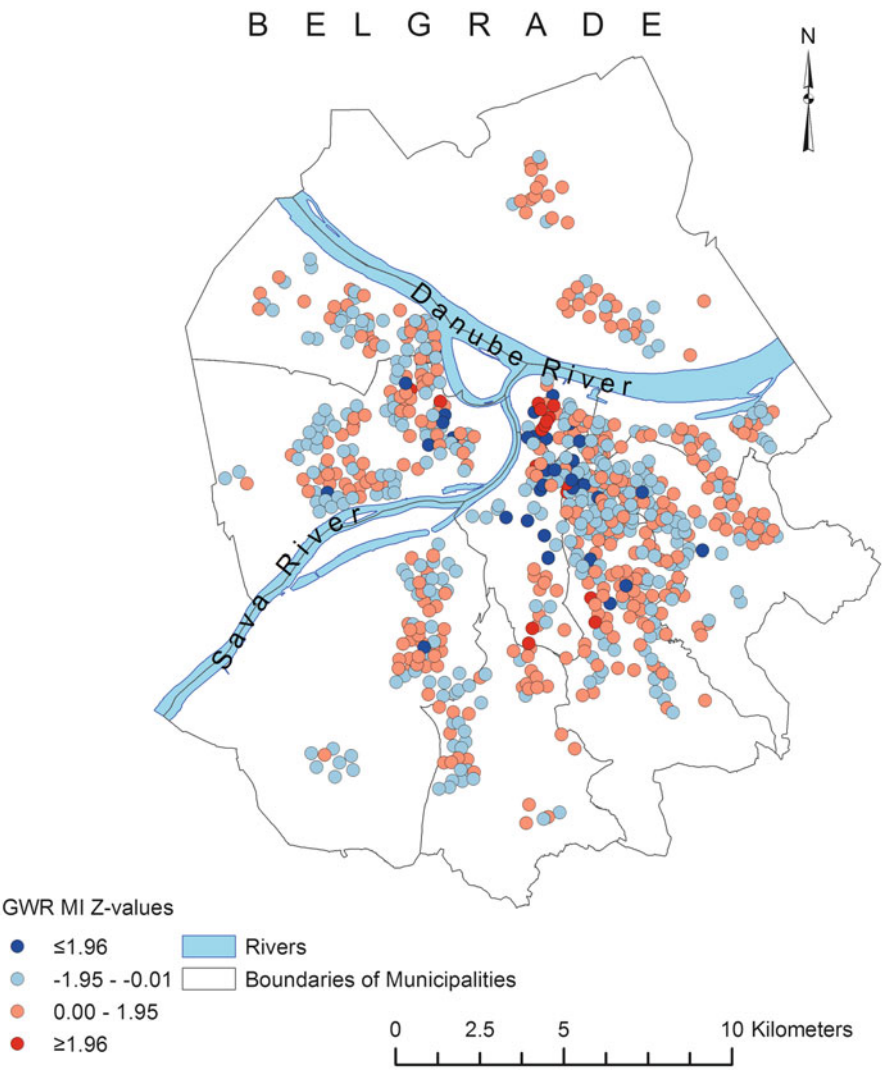

Fig. 8 Mapped local MI values of GWR residuals

model (located in the peripheral part of the city) indicate that this technique cannot correctly handle the samples pattern used in this study.

Coupling this methodology with GIS and the Web 2.0 environment facilitates large-scale housing price appraisal in the framework of collaborative GIS – thereby enabling platforms for evaluation and spatial decision support. The use of a Web-based GIS tools enables authorities to combine the different spa-

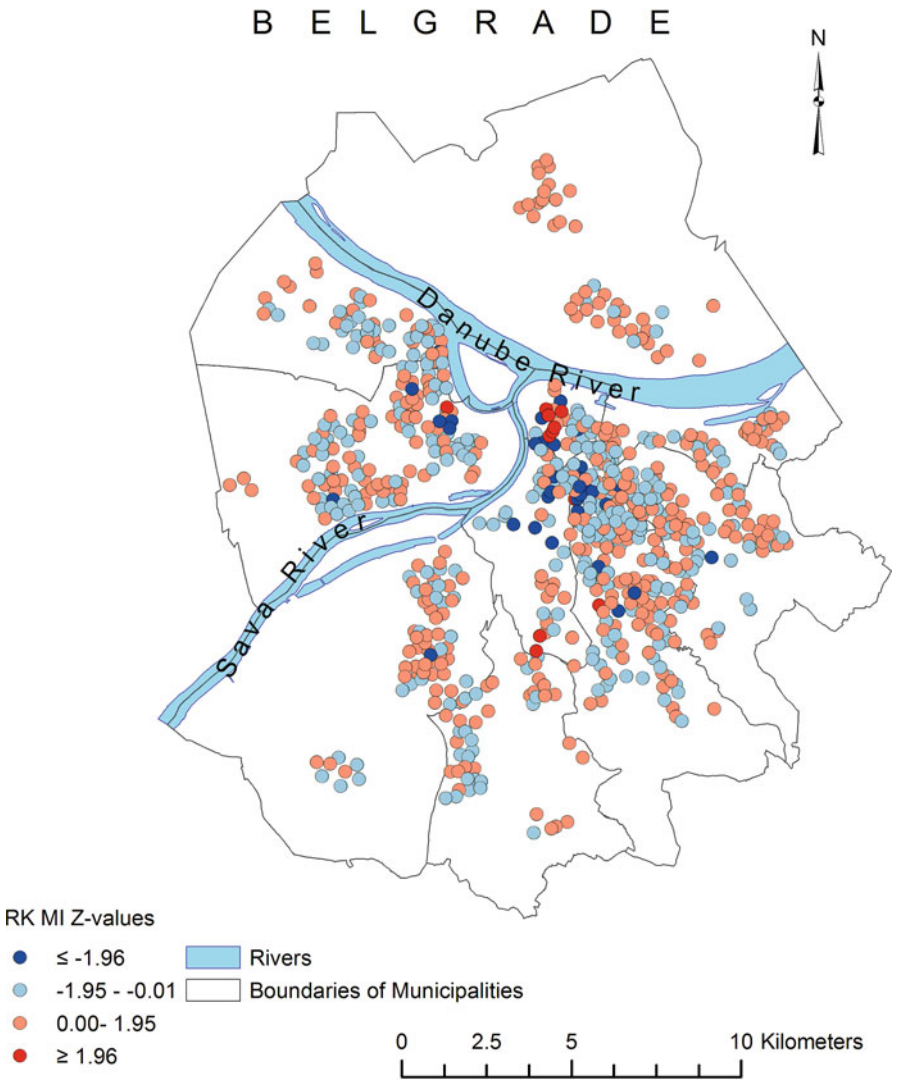

Fig. 9 Mapped local MI values RK residuals

tial layers, particularly socioeconomic datasets provided as raster maps, to spatially model distribution of housing values. This methodology provides a reliable view to spatial distribution of housing price and can be useful in hedonic price modelling.

Further research could focus on the application of spatiotemporal geostatistical techniques in hedonic price modelling. By combining the growing number of transaction database records dating from different time periods with the great variety

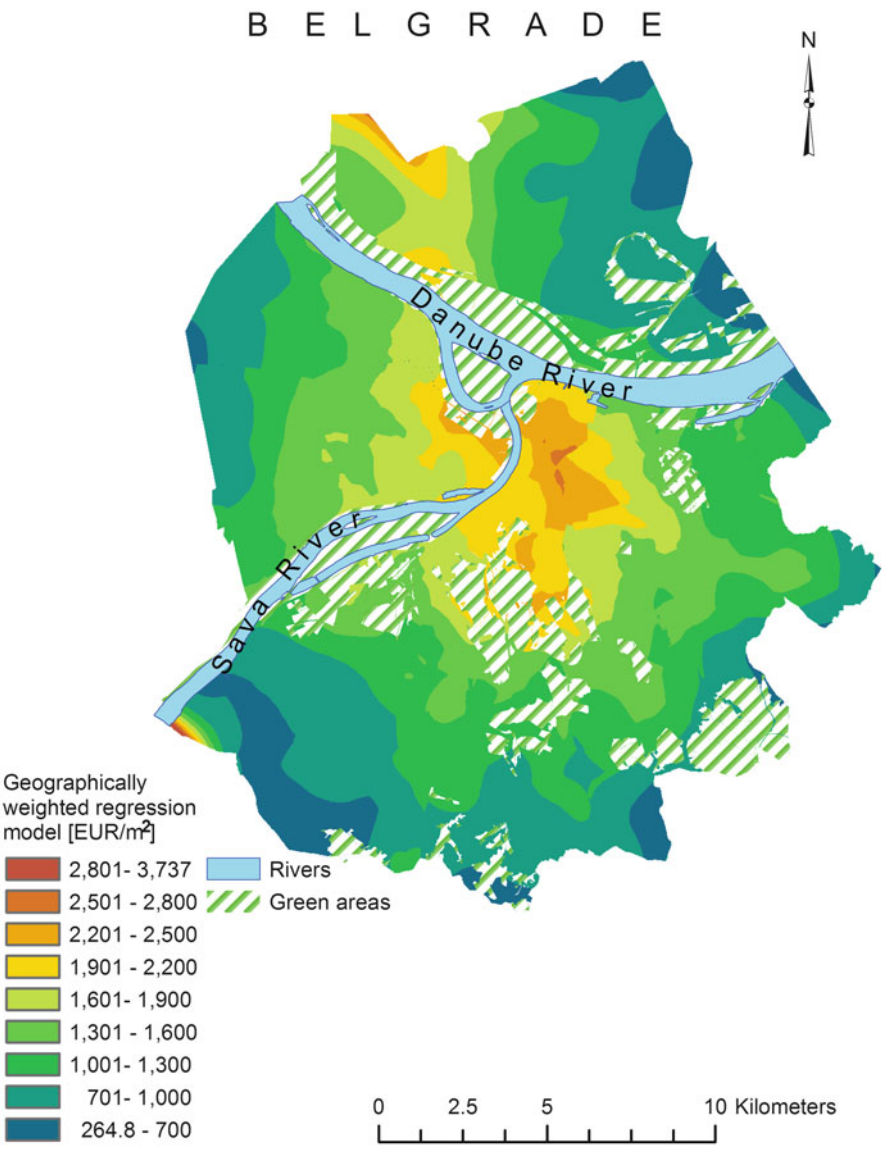

Fig. 10 The map of housing prices predicted by GWR (also available in HTML format at: http://osgl.grf.bg.ac.rs/static/materials/bajat/GWR_RK_htm/gwr.htm)

of publically accessible spatial data in GIS formats, housing price models can be developed for spatial and temporal domains. The application of GWR and RK should be reinforced in hedonic price modelling in regard to the latest developments in the field of spatio-temporal modelling (Huang et al. 2010; Cressie and Wikle 2011).

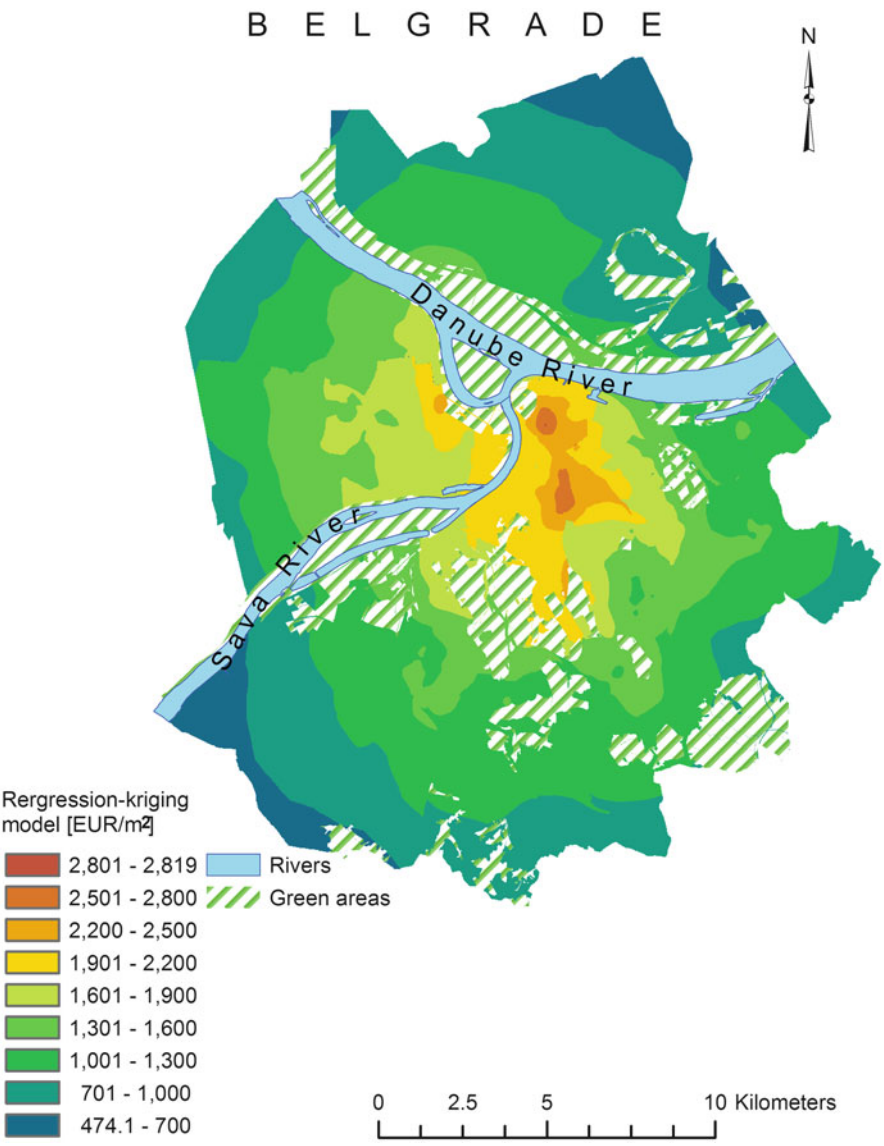

Fig. 11 The map of housing prices predicted by RK (also available in HTML format at: http://osgl.grf.bg.ac.rs/static/materials/bajat/GWR_RK_htm/rk.htm)

Acknowledgments This work was supported by the Ministry of Science of the Republic of Serbia (Contracts No. III 47014, TR 36035 and TR 36009).

References

Adair, A., McGreal, S., Smyth, A., Cooper, J., & Ryley, T. (2000). House prices and accessibility: The testing of relationships within the Belfast urban area. *Housing Studies, 15*(5), 699–716.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis, 27*, 93–115.

Anselin, L. (1998). GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research, 9*(1), 113–133.

Anselin, L. (1999). Spatial econometrics www.csiss.org/learning_resources/content/papers/baltchap.pdf. Accessed 20 Sept 2011

Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & D. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237–290). New York: Marcel Dekker.

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics, 17*(1), 61–85.

Bivand, R., Pebesma, E., & Rubio, V. (2008). *Applied spatial data analysis with R*, *Use R Series*. Heidelberg: Springer.

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics, 14*, 203–222.

Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: Cokriging. *Journal of Real Estate Research, 29*(1), 95–114.

Crespo, R., Fotheringham, A. S., & Charlton, M. E. (2007). *Application of geographically weighted regression to a 19-year set of house price data in London to calibrate local hedonic price models. Proceedings of the 9th International Conference on Geocomputation 2007*. Maynooth: National University of Ireland Maynooth.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken: John Wiley.

Cvijanović, D. (2006). Tržište stambenih nekretnina u Srbiji. (in Serbian) Kvartalni monitor 5: 66–77. http://www.fren.org.rs/sites/default/files/qm/036_km5-00-ceo.pdf. Accessed 20 June 2011

Des Rosiers, F., Thériault, M., & Villeneuve, P. (2000). Sorting out access and neighborhood factors in hedonic price modeling. *Journal of Property Investment & Finance, 18*(3), 291–315.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: John Wiley & Sons.

Dubin, R. A. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *The Review of Economics and Statistics, 70*, 466–474.

Dubin, R. A. (1998). Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics, 17*(1), 35–59.

Fernández-Avilés, G., Minguez, R., & Montero, J. (2012). Geostatistical air pollution indices and spatial hedonic and models: the case of Madrid Spain. *Journal of Real Estate Research, 34*(2), 243–274.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: Wiley.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Los Angeles: Sage.

Hanink, D. M., Cromley, R. G., & Ebenstein, A. Y. (2010). Spatial variation in the determinants of house prices and apartment rents in China. *The Journal of Real Estate Finance and Economics, 45*(2), 347–363.

Harris, R., Singleton, A., Grose, D., Brunsdon, C., & Longley, P. (2010). Grid-enabling geographically weighted regression: A case study of participation in higher education in England. *Transactions in GIS, 14*(1), 43–61.

Hengl, T. (2009). *A practical guide to geostatistical mapping* (2nd ed.). Amsterdam: University of Amsterdam. www.lulu.com.

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computational Geosciences, 33*(10), 1301–1315.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science, 24*(3), 383–401.

Kilibarda, M., & Bajat, B. (2012). plotGoogleMaps: The R-based web-mapping tool for thematic spatial data. *Geomatica, 66*(1), 37–49.

Kim, C. W., Phipps, T. T., & Anselin, L. (2003). Measuring the benefits of air quality improvement: A spatial hedonic approach. *Journal of Environmental Economics and Management, 45*, 24–39.

Koramaz, T. K., & Dokmeci, V. (2012). Spatial determinants of housing price values in Istanbul. *European Planning Studies, 20*(7), 1221–1237.

Lake, I. R., Lovett, A. A., Bateman, I. J., & Langford, I. H. (1998). Modelling environmental influences on property prices in an urban environment. *Computers, Environment and Urban Systems, 22*(2), 121–136.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*, 132–157.

Lovett, A. A., & Bateman, I. J. (2001). Economic analysis of environmental preferences: Progress and prospects. *Computers, Environment and Urban Systems, 25*, 131–139.

McBratney, A. B., Odeh, I. O. A., Bishop, T. F. A., Dunbar, M. S., & Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma, 97*, 293–327.

Meese, R. A., & Wallace, N. E. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression and hybrid approaches. *The Journal of Real Estate Finance and Economics, 14*(1–2), 51–73.

Mueller, J. M., & Loomis, J. B. (2008). Spatial dependence in hedonic property models: Do different corrections for spatial dependence result in economically significant differences in estimated implicit prices? *Journal of Agricultural and Resource Economics, 33*(2), 212–231.

O'Sullivan, D., & Unwin, D. (2003). *Geographical information analysis*. Hoboken: John Wiley & Sons.

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modelling. *Journal of Real Estate Research, 32*(3), 289–320.

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computational Geosciences, 30*, 683–691.

Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics, 4*(1), 1–12.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy, 82*, 35–55.

Statistical Office of the Republic of Serbia. (2011). *2011 census of population, households and dwelings in the republic of Serbia-first results*. Belgrade: Statistical Office of the Republic of Serbia.

Wang, F. T., & Zorn, P. M. (1997). Estimating house price growth with repeat sales data: What's the aim of the game? *Journal of Housing Economics, 6*, 93–118.

Yoo, E. H., & Kyriakidis, P. C. (2009). Area-to-point Kriging in spatial hedonic pricing models. *Journal of Geography, 11*, 381–406.

Yrigoyen, C. C., Otero, J. V., & Rodríguez, I. G. (2008). Modeling spatial variations in household disposable income with geographically weighted regression. *Estadística Española, 50*(168), 321–360.