Spatial Prediction Models for Real Estate Market Analysis

Krzysztof Chrostek[*], Katarzyna Kopczewska[**]

Abstract

The econometric modeling of real estate prices is an important step in their valuation. As shown in the theory and practice of valuation, the most important determinant of these prices is location. Therefore, models comprising the spatial components give better estimates than a-spatial models. The purpose of this paper is to compare the quality of prediction for several models: a classical linear model estimated with OLS, linear OLS model including geographical coordinates, Spatial Expansion model, spatial lag and spatial error models, and geographically weighted regression. The evaluation will be based on the calibrated models for the real estate market data in Wroclaw in 2011. The study confirms that the inclusion of the spatial aspect of the analysis may result in improvement in the quality of models. Best fit to the data among the presented methods has proved a geographically weighted regression.

Keywords: spatial modeling, geographically weighted regression, spatial heterogeneity, housing market.
JEL Code: C21, R31.

[*] Consultant at Business and Decision Poland.
[**] Assistant Professor at University of Warsaw, Faculty of Economic Sciences, e-mail: kkopczewska@wne.uw.edu.pl.

Introduction

Economic models are theoretical constructs used to represent economic phenomena. Omission of spatial aspect is a common simplification of the economic modelling. Both in microeconomic and macroeconomic theory, best known models are characterized by a lack of spatial approach. On the other hand, attempts to take account of the spatial aspects often lead to the abstract model, the results of which differ considerably from reality (Fujita et al. 2000).

The econometric modeling ignores the spatial component, most often through the use of Classical Linear Regression Model (CLRM) estimated with Ordinary Least Squares (OLS) method, applied to the data of a spatial nature, results in unfulfilled assumptions about the correctness of the functional form or spherical random errors, occurring due to the spatial autocorrelation. This results in biased estimators. The answer to these problems resides in the spatial modeling. Two basic models, the spatial error model and spatial lag model, can resolve respectively the problem of autocorrelation in random components, occurring as a result of the existence of omitted variables and spatial characteristics being taken into account in the function form with respect to the impact the neighboring locations have on dependent variable.

However, basic models of spatial error and spatial lag do not take into account the possibility of variation of parameters in space because of aggregating all locations. Spatial parameters are obtained as the point estimate not a distribution, and instead of different results in different geographical locations, all parameters are global. Fotheringham et al. (2002) indicate that it is often difficult to justify such a methodology. In their book, they illustrate problems with this approach by providing example from climate data studies where assigning a single value for specific phenomenon for the entire area of the United States and omitting the spatial aspect would result in information loss. They further suggest that aforementioned problem of information loss may not be solved simply by adding only the set of discrete variables characterizing different areas. For this reason, they suggest using models that allow for the modelling heterogeneity of parameters according to their location.

In the present paper four classes of models are to be compared: a-spatial CMLR, CMLR including geographical coordinates as an explanatory variable, basic model of spatial dependence with homogeneous spatial coefficient (spatial error model and the spatial lag model), and more advanced models with the heterogeneous spatial components - *Spatial Expansion Method and Geographically Weighted Regression.* In the spatial error model, a spatial autocorrelation error is added by including a spatial weights matrix in the error component. In the spatial lag model the spatial lag dependent variable is added to a set of explanatory variables, which should be understood as a weighted average of the neighboring location. The Spatial

Expansion Method (SEM) requires the polynomials of geographical coordinates and their interactions to be added to a linear model. Using Geographically Weighted Regression method (GWR) allows for the variation of dependence over space to be reflected in the diversity of local parameters of the model. It addresses the problem of heterogeneity of the model parameters by applying the regression model similar to kernel estimation. Both, in the case of GWR and kernel estimation, results of estimation in one point are determined by other observations. There is one difference; in the case of kernel estimation weights depend on the position in the "attribute space", but in geographically weighted regression they are based on the location in the geographic space (Brudson et al., 1996).

The purpose of this paper is to compare the effectiveness of models mentioned before in the prediction of actual data. It is done to answer the question of whether the inclusion of geographical information affects the accuracy of predictions for the econometric model. Research hypothesis states that the accuracy of predictions is indeed affected by the inclusion of spatial information in the model and taking into account the variability of coefficients in space further improves model quality. Empirical verification of models was based on a dataset of buy/sell transactions of housing units in Wroclaw in 2011. Constructed models assume that the price of the property is determined by such features as flat area, number of rooms, floor, building type, year of construction, and the presence of the garage and location.

1. Spatial modeling of real estate prices – research overview

Models taking into account the spatial heterogeneity

In the econometric literature on real estate prices models, spatial heterogeneity of parameters has been present for a long time. Casetti (1972) proposed the Spatial Expansion Method (SEM), where parameters of the global model are functions of other variables, which allows for the examination of trends in the parameters over space. It allows for the modeling of heterogeneity due to the fact that the model coefficients are different for each observation. In practice, the initial model is proposed first, and then it is further expanded to form a terminal model. When the terminal model has the correct model specification, then the initial model estimators are biased due to omitted variables. The SEM model was developed into model GWR (Fotheringham et al., 1998). One of the early uses of the model is a study by Brown and Jones (1985), who use the SEM approach for analysis of migration between Costa Rica cantons in 1968 - 1973. Gelfand et al. (2003) proposed SEM in the Bayesian approach, applied to the valuation of single-family homes in the United States. They show that the inclusion of a spatial process for all variables leads to the best results.

Foster and Gorr (1986) propose the use of adaptive spatial filtering. The method involves the adjustment of parameters in space relative to adjacent values. Jones (1991) proposes the use of multilevel modeling, which is characterized by adding an additional random factor to the model dependent on the area where the observations are located. These random factors are added to each model parameter. The method was used by them to model transaction prices of 918 houses in Southampton in 1986-1990, taking as explanatory variables the age of the house, its type, number of bathrooms, and the presence of central heating and/or a garage. Yet another method is kriging (Krige, 1951), which allows for the prediction of unknown values at certain points in space. One of the possible approaches in kriging is a model in which covariance between observations of the relative position is additionally explained (Fotheringham et al., 2002). An overview of local spatial analysis methods was presented by Fotheringham and Brunsdon (1999), including such methods as point pattern analysis, geographically weighted regression, random coefficients models, the spatial expansion method, adaptive filtering, autoregressive models, and local forms of spatial interaction models.

One of the most interesting methods is geographically weighted regression (GWR) (Brudson et al., 1996; Fotheringham et al., 2002). As the method is based on kernel estimation, this technique allows for the investigating spatial heterogeneity of parameters. In addition to the basic model commonly used, its extensions are: a) mixed GWR model for which only a part of the parameters is a variable, while the part is constant for all locations, b) a method which aims to reduce the influence of outliers in the calibration c) a geographically weighted regression model, which takes into account the spatial heteroscedasticity of errors. GWR allows researchers to model the spatial non-stationarity of economic processes.

Models price and location of the property and business

In the literature one can find examples of use of the above described models for the analysis of real estate markets. Brudson et al. (1996) used the GWR to model house prices sold in London in 1991. The basic model, using ordinary geographically weighted regression, includes 12,493 observations and makes the price of the property dependent on its size, the date of construction, type (detached, terraced single-storey, flat, semi-detached), the presence of a garage, the presence of central heating, number of bathrooms, the percentage of people in the region working in higher positions, percentage of unemployed people living in the region, and a number of interactions between the size of the property and its type.

GWR model was used also by Deller and Sunder-Stukel (2012) to explain decisions on the location of headquarters of the credit unions in nearly 3,000 counties in the United States in 2007-2008. The authors explain the relationship

between socioeconomic factors and the concentration of the credit union and answer the question whether this concentration is affected by the same factors that affect the concentration in the presence of banks, or because of divergent purposes that banks and credit unions serve, their occurrence do not coincide with each other. Explanatory variables are population density, dummy variables for metropolitan counties and for counties that are not adjacent to any metropolitan county, percentage of Afro-American population, the percentage of Hispanic population, percentage of the population over 25 years with a bachelor's degree; percentage of the population born outside the U.S., poverty rate, the percent change in the number of households, percentage of houses occupied by the owners, population to employment ratio, the unemployment rate, the ratio of population to the number of property owners, the number of banks per capita, the number of loans and deposits per capita, and the number of different types of organizations and associations. In the verification of hypotheses authors use both spatial and a-spatial models.

They used non-spatial: CLRM and tobit models, and spatial: the spatial error model, spatial lag model, the spatial Tobit, and GWR. The quality of the models was compared with the use of residuals analysis (sum of squared errors, root mean square error, and mean absolute error) and correlation of observed and predicted values. None of the models proved to be of more use than the other in all applied criteria. Nevertheless in any of these criteria GWR model proved to be the best or the second-best model. Based on the obtained results, the authors state that the high concentration of credit unions is negatively associated with the concentration of banks, which confirms the hypothesis that credit unions are established in areas where there are few banks. The results also allowed for the questions regarding the impact of socio-economic variables on the location of the credit union to be answered.

GWR models and Spatial Expansion Methods were used by Bitter et al. (2006), in explaining the prices of detached houses in Tucson based on 11,732 transactions in 2000. The model was estimated on the basis of 90% of the observations, while the remaining 10% were used to assess *ex ante* errors. The logarithm of house prices is explained by variables such as the size of the patios, dwelling area, presence of refrigerated air conditioning, presence of a swimming pool, number of rooms divided by dwelling size, structural quality of the dwelling, age of the dwelling, number of floors, number of bathrooms divided by the number of rooms, interior quality of the dwelling, presence of a garage. These variables, however, were reduced with the use of principal components analysis to the seven variables that were included in the model. The authors compare the results of several methods: a) regression, which includes the latitude and longitude and its interaction, up to the third order, b) the *spatial expansion method,* for which the third order polynomial was used for the coordinates c) the *spatial expansion method* with an additional

variable representing the spatial lag - where the lag is defined as the weighted average of the 15 nearest neighbors of the observation d) geographically weighted regression. The highest forecast accuracy on the test sample was given by GWR models (based on two criteria: in the case of the smallest number of observations matched values have been exceeded by more than 10% and more than 20% of the observed ones) and the *spatial expansion* model with spatial lag (based on criteria of the highest $R^{2)}$. The authors also note that for most locations geographically weighted regression results provide more accurate predictions than the results of models based on the *expansion method.*

The problem of spatial autocorrelation and heterogeneity when modeling the price of over 68 thousand real estate in Milwaukee in 2003 was raised by Yu, Wei and Wu (2007). According to authors the sources of spatial relations include similar behavior of the neighboring owners in regard to renovation and improvements of buildings, as well as similar surroundings of neighboring properties. Their model explains the valuation of the property and is used for data from the Milwaukee Master Property database. The variation of the dependent variable is explained by floor area, the presence of air conditioning, the presence of a fireplace, the number of bathrooms, age of the property, and type of surface. The purpose of the latter variable is to examine the impact of environmental degradation on the valuation of the house. In modeling, the authors used the following methods: the ordinary least squares, spatial error and spatial lag model, and geographically weighted regression. Models were compared using AIC statistics for samples used for estimation, and the relative error statistics and the root mean square error for data both *in-sample* and *out-of-sample.* The authors drew conclusions from models that all the analyzed attributes affected the price of homes significantly, and in accordance with intuition. Floor area, the presence of air-conditioning and fireplace, and the number of bathrooms have a positive effect and the level of degradation of the environment and the age of the house negatively affect the value of the residence. The authors also conclude that in general spatial models fit the data better than a simple regression. Best performance *out-of-sample* was achieved by spatial error model followed by GWR. Another finding of the study is the fact that geographically weighted regression captured the notion that parameters of analyzed phenomenon vary depending on the observation's position.

Among research for Polish real estate market on the spatial heterogeneity of parameters, there is a study for 3800 transactions in Olsztyn in 2003-2009 (Cellmer, 2010) and a study of 276 prices of residences in Kraków's Krowodrza in 2004-2005 (Kulczycki and Ligas, 2007). In both models, the price of real estate is explained with a linear trend, i.e., a variable representing the number of months that have elapsed since the first observation. Kulczycki and Ligas (2007) further add variables defining the size of flats, the percentage of depreciation and the number of variables expressed on ordinal scales containing information about the surroundings, and standard

equipment and the availability of transportation infrastructure. However, variables expressed on an ordinal scale were included with no breakdown by individual levels, and therefore assumed that for each pair of adjacent levels difference in impact these levels have on the dependent variable is the same. For the calculation of parameters both CMLR and GWR were used. The authors state that improvement of model quality is easier and more efficient by including spatial components than when using more complex non-linear a-spatial factors.

GWR was used also by Ilnicki et al. (2011), who presented an analysis of spatial relationships between the population and the presence of stores in Wroclaw. A model was built on 527 observations using Thiessen polygons (or Voronoi polygons) obtained for centroid census enumeration. The endogenous variable in the model is the number of stores per square kilometer, and the exogenous variables are as follows: population density, distance from the centre of the city, latitude, longitude, road density, the number of dwellings per square kilometer, and the number of buildings per square kilometer. The authors, however, reduce the model to the significant variables in the regression model: number of buildings per square kilometer and population density. Additionally, they examine how the results of the estimation affect the distance from the centre. To obtain the results they use several different methods: the ordinary least square method and geographically weighted regression with the spatial weights matrix based on the criterion of cross-validation and Akaike information criterion. The authors compare the results of estimation of six models. Based on the conventional R^2, models including spatial aspect performed better. The best of them was the model which takes into account the distance from the centre and includes a neighborhood matrix obtained with AIC criterion. The study proves that the deployment of shops is significantly correlated with the population density, the number of buildings, and their distance from the centre. The authors emphasize that by using GWR it became possible to also present the results of the estimation of the maps.

2. Functional form of models with the spatial component

Six models listed below were used in the analysis of the research problem of property valuation.

The first model is a classical linear regression model (CLRM), which completely ignores the spatial aspect. It is used as the base model in assessing the quality of improvement in extended models. This is given by:

$$y = BX + \varepsilon \tag{1}$$

where y is the dependent variable, X is a matrix of explanatory variables, β the vector coefficients of the model, and the random component $\varepsilon \sim IID\ N\ (0, \sigma)$. It is a

non-spatial model. If spatial dependence exists which is not included in this model, OLS estimators are inefficient and biased.

The second model further extends CLRM by polynomials of variables latitude and longitude of up to the third degree. This model is given by:

$$y = BX + \theta_1 x^1 + \theta_2 x^2 + \theta_3 x^3 + \vartheta_1 y^1 + \vartheta_2 y^2 + \vartheta_3 y^3 + \mu_{1,1} x^1 y^1 + \mu_{2,1} x^2 y^1 + \mu_{1,2} x^1 y^2 + \varepsilon \quad (2)$$

where x^n are polynomial longitude coordinates (also called *easting*) and y^n are polynomial latitude coordinates (also called *northing*), $x^n y^m$ are mixed polynomials, θ and υ and μ are model coefficients. It takes into account the effects of localization in a simplified manner, generating global spatial factors.

The third and fourth models used the spatial error model and the spatial lag model to enable exploration of spatial dependence. While the analysis of spatial dependence is not the purpose of this study, these methods are provided in order to compare their results with the results of geographically weighted regression. For calculation of the models, spatial weights matrix under the criterion of inverse distance was applied. These models have the form as follows:

$$y = BX + u \text{ and } u = \lambda Wu + \varepsilon \quad (3)$$

and

$$y = BX + \rho Wy + \varepsilon \quad (4)$$

where u is the spatial random error, λ and ρ are the parameters of spatial autocorrelation, W is a spatial weights matrix, Wu is an error spatial lag, and Wy is an explained variable lagged spatially. Spatial effects allow for assessment of importance of the neighborhood for a given process. Spatial lags, understood as the average of the values in neighboring locations, when weighted with spatial weights, can detect clusters of similar observations as well as the local outliers.

The fifth model is the *Spatial Expansion Method,* which allows for the modeling of spatial heterogeneity of model parameters because the parameters are functions of certain variables. It is one of the local spatial analysis models used in the situation of systematic variability in the regression coefficients. This method consists of adding to the polynomial model variables denoting latitude and longitude, and their interactions. The global model has the form (Fotheringham et al., 1998):

$$y_i = \alpha + \beta x_{i1} + \cdots + \tau x_{in} + \varepsilon_i \quad (5)$$

where x_n are the explanatory variables, and β, α, τ ... the model parameters.

The basic model extension in order to account for the heterogeneity of the parameters takes the form:

$$\begin{cases} \alpha_i = \alpha_0 + \alpha_1 x_i + \alpha_2 y_i \\ \beta_i = \beta_0 + \beta_1 x_i + \beta_2 y_i \\ \tau_i = \tau_0 + \tau_1 x_i + \tau_2 y_i \end{cases} \tag{6}$$

It is also possible to take into account the extension of a more complex form with polynomials of coordinates up to second degree:

$$\begin{cases} \alpha_i = \alpha_0 + \alpha_1 x_i + \alpha_2 y_i + \alpha_3 x_i^2 + \alpha_4 y_i^2 + \alpha_5 x_i y_i \\ \beta_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 y_i^2 + \beta_5 x_i y_i \\ \tau_i = \tau_0 + \tau_1 x_i + \tau_2 y_i + \tau_3 x_i^2 + \tau_4 y_i^2 + \tau_5 x_i y_i \end{cases} \tag{7}$$

where x_i^n is a polynomial longitude coordinates, and y_i^n is a polynomial of the coordinates of latitude.

The sixth, and the last, model is a geographically weighted regression (GWR). In contrast to models that assign global parameters to variables, this method allows for variation of parameters depending on the location. It can be presented as follows (Fotheringham et al., 2002):

$$y_i = \sum_k \beta_k (xx_i, yy_i) x_{ik} + \varepsilon_i \tag{8}$$

The parameters β are as follows (Fotheringham and Brunsdon, 1999):

$$\beta(xx_i, yy_i) = (X' W(xx_i, yy_i) X)^{-1} X' W(xx_i, yy_i) \mathrm{y} \tag{9}$$

where $W(xx_i, yy_i)$ is a weighting function.

One of the key issues in GWR modeling is the need to select the weighting function, which will correspond to a process which generates the data. Weights which change with a Gaussian kernel are often selected for weighting functions. Therefore it has a form:

$$w_{ij} = \exp[-\frac{1}{2} (\frac{d_{ij}}{b})^2] \tag{10}$$

where w_{ij} is the weight of a data point j in regression point i, d_{ij} is the distance between points i and j, and b is the smoothing parameter (*bandwidth*).

The selection of smoothing parameter is possible based on the calibration, in the case of which the sum of squares of deviations between fitted values and realized values is minimized. Optimization is achieved by cross-validation, wherein for each observation its theoretical values were calculated from the model, which omitted the variables of that observation.

In assessing the quality of models, it is impossible to apply the R^2 criterion as the analyzed models have different numbers of degrees of freedom. It is also not

always possible to use the information criterion AIC, as for the GWR models log-likelihood functions are not calculated. Therefore, several criteria are established to replace the Akaike information criterion. One of these was proposed by Fotheringham et al. (2002).

In addition to assessing the goodness-of-fit, it is important to check whether the parameters derived from the model estimation are indeed able to predict the dependent variable. Comparisons beyond the estimated sample are possible by cross-validation, run with a bootstrapping approach. This involves the division of trials into *n* sub-samples (e.g. 10) and counting each model *n* times, using a different set of *n*-1 (e.g. 9) samples each time and excluding one sub-sample. The outcome of each model creates an *out-of-sample* prediction for the sub-samples not included in the analysis and determines the *out-of-sample* random errors and *ex-post* statistics.

Models can be compared using several methods: a) MAE (*mean absolute error*) - the average absolute error, defining an average deviation of the predicted value from the empirical one; b) RMSE (*root mean square error*) - as the MAE, measuring the average deviation of the predicted value from the empirical one, but assigning higher weights to greater deviations; c) ME (*mean error*) – useful in deciding whether given models consistently lead to an overestimation or underestimation of the endogenous variable, and d) MAPE (*mean absolute percentage error*) - the mean relative error, comparing errors with the observed values. These are typical measures of assessment predictions.

3. Estimation of real estate prices in Wroclaw in 2011

Characteristics of the dataset

The modeling of housing prices in Wroclaw in 2011 was conducted using data from the AMRON Center. The system of Analysis and Monitoring Real Estate Market (AMRON) is a standardized database of prices and real estate values, and is owned by the Polish Bank Association. The database contains data on the characteristics of various property, including their location and transaction prices. The analyzed database consisted initially of 5602 observations and estimates containing approximately 90% of all real estate transactions in this market. The data collected in the database originated from three sources: the appraisers, the banks, and the Polish Bank Association. After removing observations for where complete information was not given as required from inputs to the model, there were 598 observations remaining for analysis.

Modeling was performed using the statistical package R. Estimation of the spatial models of dependency was run with package *spdep* (Bivand, 2013) and GWR with package *spgwr* (Bivand and Yu, 2013). To visualize location of the observation on the map of Wroclaw, a shapefile contour map from the Geoportal website was

applied. The acquisition of geographic coordinates of observation was possible due to address variables in AMRON database. Transformation of text data into geographic coordinates was possible with the use of R and package SmarterPoland (Biecek, 2013), based on the geocode application created by Google. Some coordinates were not unique (e.g. in the case of data on apartments in the same building), therefore a random variable from uniform distribution in the range (-1e-7, 1e-7) was added to longitude and latitude. This was accomplished to fulfill the requirements of models to have observations in different locations - this is necessary in the case of the spatial weights matrix under criterion of inverse distance and in GWR which uses a weighing function, depending on the distance between the points.

In the model construction, the continuous and integral variables describing the property price, their area, floor and utility room surface, and discrete variables such as the year of construction, the presence of a garage and the type of building, were used. The **transaction price** of real estate housing in PLN is the explained variable in the analysis. This variable has not been subjected to logarithmic transformation, which would treat the parameters as semi-elasticities, because there is no single answer to the question of whether such a transformation is reasonable (Freeman, 2003) and most empirical studies use price as an endogenous variable in the unprocessed form. However, log transformation can flatten u-shaped relationship between price and size of the property. This non-linear pricing can appear due to demand externalities (Oren et al., 1982). **Flat area** in square meters is a natural determinant of property price in most of the analyzed models (e.g. Fotheringham et al., 2002, Bitter et al., 2006). **Floor,** on which the dwelling is located, determines the choices of buyers and empirical distributions show that the housing prices of similar size properties vary between floors. The model also takes into account the second power of this variable to capture possible non-linear relationships. **Usable area per room** is understood as the usable area divided by number of rooms, which allows for diversifying housing with larger and smaller rooms. This was added to the model under the influence of Bitter et al. (2006), where numbers of rooms divided by area was used, however, in order to facilitate interpretation, an inversion of this variable has been chosen for this study. The table below shows the characteristics of continuous variables.

Table 1. Descriptive statistics of continuous variables

Variable	Average	Median	St.dev.	Minimum	Maximum	Skewness	Kurtosis
Price	347900	325000	128171.25	80000	1474000	2.22	11.64
Size	56.88	53.4	19.52	19	150.1	1.5	3.97
Floor	2.8	2	2.39	0	10	1.09	0.81
Usable area per one room	24.48	23.6	6.4	11	62.08	1.67	5.05

Including a garage in the price of the flat significantly and positively affects the price of the property, which is confirmed by the analysis of the variance test. The variable for **the year of construction** may have an ambiguous effect on the valuation of the property (Fotheringham et al., 2002). On the one hand, newer buildings can have a higher quality, and their depreciation may be lower. On the other hand, in some cases property value may increase with age, as older buildings can be more valued because of their architecture or better location. Because of the ease of interpretation, as well as the accuracy of the dataset, the variable has been divided into intervals, depending on the date of construction: before 1940 (56 observations), between 1940 and 1980 (176 observations), between 1980 and 2000 (94 observations) and after 2000 (272 observations). ANOVA indicates that there is a significant correlation between the values of this variable and those of housing prices. The model takes into account **the type of building** in which the dwelling is located. These categories are multifamily low buildings (165 observations), multifamily high buildings (314 observations), townhouses (37 observations), skyscrapers (45 observations) and apartments (7 observations). For the analysis, only the two biggest groups were included - multifamily low buildings and multifamily high buildings. Other types were combined in one degree variable, which forms the baseline in the model. Finally, the model was built with the below listed variables (see Table 2).

Table 2. Variables used in the models

Variable	Description
Price	housing transaction price (in Polish PLN)
Size	usable floor space (in square meters)
Usable area per one room	floor area per room (in square meters)
Floor	floor on which the flat is located
$Floor^2$	variable floor squared
Multifamily low building	dummy variable=1 for flats in a multifamily low building
Multifamily high building	dummy variable=1 for flats in multifamily high building
Year 1940_1980	dummy variable=1 for the property built between years: 1940-1980
Year 1980_2000	dummy variable=1 for the property built between years: 1980-2000
Year 2000	dummy variable=1 for the property built after 2000
Garage	dummy variable=1 for units with garage

Table 3 shows the results of different model estimation methods: OLS, OLS with added geographical coordinates, the spatial error model, and the spatial lag model. The results of S*patial Expansion* models are not presented, because this method generates a large number of parameters - six parameters per one variable with polynomials of up to second degree, and ten parameters at third degree. Table

3 also does not provide the results of the estimation of 9 parameters corresponding to the coordinate polynomials in model using the method of least squares with the added coordinates.

Table 3. The results of OLS estimation models, OLS with coordinates, the spatial error model and the spatial lag model

	OLS		**OLS with coordinates**		**Spatial error model**		**Spatial lag model**	
Variable	Coefficients	P-value	Coefficients	P> | t |	Coefficients	P> | z |	Coefficients	P> | z |
Constant	3016.20	0.8710	-19720.00	0.2556	-4889.02	0.7754	-47235.74	0.0070
Size	5108.30	<2e-16	5327.00	<2e-16	5269.80	<2.2 E-16	4521.81	<2.2 E-16
Floor	2409.80	0.4814	-827.80	0.7918	444.28	0.8815	3596.74	0.2507
Floor^2	-234.20	0.5519	124.30	0.7291	55.80	0.8720	-296.15	0.4104
Usable area per one room	155.30	0.7427	-210.40	0.6260	-417.72	0.3103	-227.84	0.5986
Multifamily low building	10451.00	0.2861	21140.00	0.0204	19035.10	0.0230	8527.54	0.3411
Multifamily high building	9132.90	0.3143	4390.00	0.5988	11661.84	0.1264	10636.57	0.1993
Year 1940_1980	13262.50	0.2488	22850.00	0.0336	20806.63	0.0471	12548.66	0.2323
Year 1980_2000	23492.40	0.0625	38730.00	0.0011	32124.97	0.0043	21097.02	0.0669
Year 2000	59689.10	1.39 E-07	87880.00	2.45 E-15	66383.00	0.0000	50444.80	0.0000
Garage	19431.50	0.0078	13230.00	0.0483	16320.36	0.0043	19251.47	0.0038

In the case of the CLRM estimated with OLS, a significance level of 5% was proven by size, year of building after year 2000 and the presence of a garage, for model with R^2 adjusted = 71.74%. Test statistics F=152.6 (p-value ≈ 0) indicates that all the variables included in the model are jointly significant.

Estimates of OLS model, including polynomials of geographical coordinates, are consistent with the results of the standard OLS model. In addition, the variables of multifamily low buildings and year of construction proved to be significant. Statistic F=105.4 (p-value ≈ 0) rejects the hypothesis of insignificance of all variables in the model. Adjusted R^2=76.96% was found to be higher than in the base model. This means that even a very simplified model captures the spatial relationship of the real estate market and improves goodness-of-fit.

In the spatial error model, the sign of coefficients and the significance of variables proved to be consistent with OLS estimates. Significant spatial error autocorrelation coefficient $\lambda = 0.5272$ indicates a strong dependence of neighborhood and the existence of exogenous shocks or omitted variables in the process. Wald test $W = 261.28$ (p-value <5%) confirms the correctness of the model. Pseudo-R^2 Nagelkerke is 0.7969 (Nagelkerke, 1991), which makes the goodness-of-fit acceptable and better than in previous models.

The spatial lag model results appeared to be similar to the basic CLRM model. Significant spatial autocorrelation coefficient of the explanatory variable $\sigma = 0.4782$ indicates a strong link between neighboring observations. Wald test $W = 261.28$ (p-value <5%) indicates the total significance of all the variables used. The pseudo-R^2 Nagelkerke was 0.7579, worse than in the spatial error model.

In the Spatial Expansion model, using the interactions of mentioned variables with polynomials of coordinates up to the third degree, each variable or interaction is proved to be significant at the 0.1 level at least once. Type of building, differentiating apartments in high multifamily buildings from the other apartments, proved to be the only insignificant variable. Adjusted R^2 was 0.8058, better than in the spatial error model. The high values of these statistics are, however, result of over-fitting. This is confirmed by the analysis of quality measures and capacity of predictive models. To avoid overtraining, an analogous model was estimated additionally with polynomials of geographical coordinates of the second degree only. In this limited model, the same variable for an apartment in high multifamily buildings as well as its interactions became insignificant. Adjusted R^2 is 0.7869.

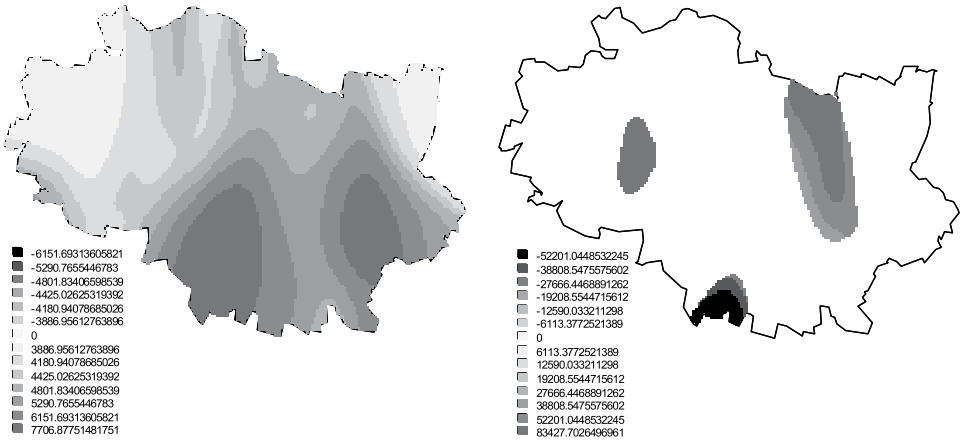
Coefficients of GWR model are presented in Table 4. In the case of many variables, its influence on the price of the flats proved to be unequal at different points in space. The direction of the relationship is fixed only for the parameters indicating the area and group of newest flats. Adjusted R^2 is 0.7869.

Table 4. GWR model estimation results

Variable	min.	1 quarter	median	3 quant	max.
Constant	-152700.00	-41120.00	-19510.00	22280.00	96570.00
Size	3397.00	4700.00	5372.00	6036.00	7579.00
Floor	-22630.00	-6,214.00	2391.00	5846.00	21470.00
Floor^2	-5883.00	-744.50	-220.40	698.20	3355.00
Usable area per one room	-3788.00	-352.60	-21.27	314.90	2699.00
Multifamily low building	-98700.00	6219.00	15730.00	25490.00	66950.00
Multifamily high building	-79280.00	-10680.00	-4503.00	15740.00	65160.00
Year 1940_1980	-51010.00	13660.00	30190.00	43730.00	110900.00
Year 1980_2000	-25840.00	22830.00	41270.00	47620.00	219500.00
Year 2000	12860.00	66100.00	83520.00	100400.00	132200.00
Garage	-50180.00	1936.00	12430.00	17100.00	44860.00

Graphical analysis of the significance and value was carried out for all parameters of GWR model. Figure 1 shows an example of graphs. Only the variable area coefficients were significant at all points of the study area. Other variables of significance are partial, suggesting their ambiguous relationship with the price of residential property. For some variables the reason may be the lack of certain features in a particular area.

Figure 1. The values of parameters in the GWR: a) variable *size*, b) a variable *multi-family low building*



Evaluation of the quality of models is positive (see Table 5). Comparisons were made on the basis of statistics of goodness-of-fit and statistics assessing the accuracy of predictions for models. The procedure is to enable the verification of the hypothesis; that the inclusion of spatial factor analysis allows researchers to obtain results of higher quality.

Table 5. Quality statistics for estimated models

Model	AIC	RMSE	IEA	MAPE	ME
OLS	15020.49	69137.76	45330.29	13405	-277.70
OLS with coordinates	14909.80	186316.31	136581.32	43.243	-48201.74
Spatial Expansion 3	14883.86	427021.53	248835.42	80.890	-35272.87
Spatial Expansion 2	14903.03	216028.60	116107.02	36.476	15045.79
Spatial error model	14835.00	68692.86	44905.84	13186	4162.96
Spatial lag model	14940.28	71521.84	47542.07	14191	701.43
GWR	14868.15	60728.96	41904.06	12961	-2124.72

Akaike information criterion (AIC) calculated for the sample *in-sample* is the lowest in the case of spatial error model. The second best model according to this criterion is the GWR. The least preferred model is basic OLS. It is worth not-

ing here that in the case of models estimated using GWR, maximum likelihood logarithms are not calculated. There are several versions of the AIC measure for GWR. As mentioned earlier, in this study it was decided to use the criterion proposed by Fotheringham et al. (2002). It is possible, however, to give a different version of the measure, which would indicate a better fit for the GWR model than that obtained using the spatial error model. These problems do not occur in the case of the other criteria for which statistics are calculated for samples not included in the counting of parameters.

To prove overfitting in OLS models with coordinates added, *Spatial Expansion Method* with polynomials up to the third degree are the statistics counted *out-of-sample.* The values of these statistics for these models are significantly higher than for the other four methods. The GWR model is characterized by the least absolute forecast error (MAE statistics). It also performs better than other models when higher weights are applied to the highest error (RMSE) and in a ratio of value of the prediction error to the dependent variable. The next best models according to these criteria are the spatial error model, OLS without coordinates, and the spatial lag model. An average error indicates, however, that for the data used in the analysis, GWR and OLS models underestimate the price of the property, while the spatial dependence models overestimates these values. Mean errors obtained from the basic OLS model are closest to zero.

Summary

All applied models confirmed that the price of a property depends mainly on its usable floor area, which is consistent with intuition. Similar conclusions have been obtained in other studies analyzing real estate prices, for example, in Yu, Wei and Wu (2007). This relation proved to be significant in GWR model for all locations which are the subject of analysis.

This study did not confirm the theory proposed by Fotheringham et al. (2002); that the age of a building can unequally affect the value of the apartment, depending on the location. All models showed that the later the property was built, the higher its price was. It is possible, however, that this is a unique feature of Wroclaw (and there are no places where old buildings are higher valued) and that in the case of analysis for other cities this relation could demonstrate the opposite results.

In comparisons of models based on the criteria of goodness-of-fit of models to actual data and based on the Akaike information criterion from the results of *in-sample*, it can be concluded that the models which take into account spatial elements are preferred over non-spatial models. Best performance was found in the cases of the spatial error model and GWR, which was also confirmed by cross-validation. The best models according to the size of the errors in the analysis of *out-of-sample* proved to be GWR, spatial error, OLS, and finally spatial lag. We found

that the addition of variables indicating the coordinates to the OLS model led to its overtraining. Excessive fitting was found both in cases where the coordinates interacted with independent variables (SEM) and in cases where only coordinates without interactions were added. Based on the results of analysis it can be confirmed that the inclusion of spatial elements in the real estate valuation models has a positive effect on the quality of the model. It should be noted, however, that in order to improve the model it is not enough to apply only a simple method of adding coordinates' variables to OLS model, as these new variables can lead to the overtraining of the model. The analysis in this paper suggests that spatial relationships are more complex on the real estate market in Wroclaw. Top matches to the data seem to be the model estimated using GWR. These results are consistent with the results of studies comparing GWR with other models based on other empirical data (Bitter et al., 2006; Deller and Sunder-Stukel, 2012).

One of the possible ways to explain why the GWR model performed better than the OLS model is the difficulty in construction of an equation that represents the interdependence between the actual characteristics of the dwelling and its price. In the case where all necessary features associated with the housing environment affecting the price would be included in the model then using the methods of differentiating the spatial parameters would be redundant. However, this would require a much broader set of data being available for the analysis. Smoothing methods based on spatial parameters (as geographically weighted regression), however, permit some way to overcome these obstacles.

The main objective of this study was to compare the quality of models, and not to create a model explaining the price of housing. It is therefore possible that an improvement in the model results could be achieved by appropriately adjusting the parameters to obtain more suitable models. For models of spatial dependence, other spatial weights matrices could possibly be used. In the case of GWR model, a change of the weighting function and the method of optimization as well as application of the adaptation process in the selection of the number of nearest neighbors, which would be taken into account when weighting geographically, could be considered.

More advanced tools could still be used in order to further improve performance and fit of the model. Harris et al. (2010) conducted a simulation to evaluate the different spatial models. Based on these results they drew conclusions that methods based on universal kriging and a hybrid model combining GWR with kriging can give more accurate predictions than GWR.

Further analysis comparing spatial models could also include the model described in the work of Gelfand, Kim, and Sirmans (2003), using the Bayesian approach. Further research could also be directed at the analysis of time-space relationships. Models that take into account both spatial heterogeneity of parameters and the time factor, were proposed, inter alia, by Huang, Wu and Barry (2010), and aforementioned Gelfand, Kim, and Sirmans (2003).

References

Biecek P. (2013) SmarterPoland: A set of tools developed by the Foundation SmarterPoland.pl. R package version 1.2. http://CRAN.R-project.org/package=SmarterPoland

Bitter, Mulligan, Dall'erba (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)

Bivand R. (2013) spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-68. http://CRAN.R-project.org/package=spdep

Bivand R., Yu, D. (2013) spgwr: Geographically weighted regression. R package version 0.6-24. http://CRAN.R-project.org/package=spgwr

Brown, Jones (1985) Spatial variation in migration processes and development: a Costa Rican example of conventional modeling augmented by the expansion method. Demography, 22 (3)

Brunsdon, C., Fotheringham, AS, Charlton. ME (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis, 28

Brunsdon, C., Fotheringham, AS, Charlton, ME (1998) Geographically weighted regression - modeling spatial non-stationarity. Journal of the Royal Statistical Society, Series D-The Statistician, 47 (3)

Brunsdon, C., Fotheringham, AS, Charlton, ME (1998) Spatial nonstationarity and autoregressive models. Environment and Planning A, 30

Brunsdon, C., Fotheringham, AS, Charlton, ME (1999) Some notes on parametric signficance tests for Geographically weighted regression. Journal of Regional Science, 39

Brunsdon, C., Fotheringham, AS, Charlton, ME (2000) Geographically weighted regression as a statistical model. Working paper, Department of Geography, University of Newcastle.

Casetti, E. (1972) Generating models by the expansion method: applications are geographic research. Geographical Analysis, 4

Cellmer R. (2010) Spatial analysis of the dynamics of changes in real estate prices premises. Acta Scientiarum Polonorum, Administratio Locorum, 9 (3)

Deller, S. Sundaram-Stukel, R. (2012) Spatial patterns in the location Decisions of U.S. credit unions. The Annals of Regional Science, 49 (2)

Foster, SA, Gorr, WL (1986) An Adaptive Filter for Estimating spatially Varying Parameters: Application to Modeling Police Hours in Response to Calls for Service. Management Science, 32

Fotheringham, AS, Brunsdon, C. (1999) Local forms of spatial analysis. Geographical Analysis, 31

Fotheringham, AS, Charlton, ME, Brunsdon, C. (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis, Environment and Planning A, 30

Fotheringham, AS, Charlton, ME, Brunsdon, C. (2002) Geographically weighted regression: the analysis of spatially varying relationships.

Freeman (2003) The measurement of environmental and resource values. Resources for the Future, Washington DC

Fujita, M., Krugman, P., Venables, A. (2000) The Spatial Economy: Cities, Regions, and International Trade.

Gelfand, AE, Kim, HJ, Sirmans, CJ, Banerjee, S. (2003) Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association, 98

Harris, P., Fotheringham, AS, Crespo, R., Charlton, M. (2010) The use of Geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Mathematical Geoscience, 42

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)

Ilnicki, D., Janc, K., flange, M., Szymanowski, M. (2011) Features distribution of stores in the metropolitan for example Wroclaw - the use of geographically weighted regression.

Jones (1991) Specifying and estimating multi-level models for geographical research. Transactions of The Institute of British Geographers, 16

Jones, JP, Casetti, E. (1992) Applications of the expansion method.

Krige (1951) A statistical approach that some mine valuations and allied problems at the Witwatersrand. Master's thesis of the University of the Witwatersrand.

Kulczycki, M., Ligas, M. (2007) Geographically weighted regression as a tool for the analysis of real estate market, Geomatics and Environmental Engineering, 1 (2)

Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. Biometrika 78: 691-692

Oren S.S., Smith S.A., Wilson R.B. (1982) Non-linear pricing in markets with interdependent demand, Marketing Science, vol.1, no 3, Summer 1982

Yu, DL, Wei, YD, Wu, CS (2007) Modeling spatial dimensions of housing prices in Milwaukee, WI. Environment and Planning B: Planning and Design, 34