

Radosław Cellmer

The use of the geographically weighted regression for the real estate market analysis

Folia Oeconomica Stetinensia 11(19)/1, 19-32

2012

Artykuł został opracowany do udostępnienia w internecie przez Muzeum Historii Polski w ramach prac podejmowanych na rzecz zapewnienia otwartego, powszechnego i trwałego dostępu do polskiego dorobku naukowego i kulturalnego. Artykuł jest umieszczony w kolekcji cyfrowej bazhum.muzhp.pl, gromadzącej zawartość polskich czasopism humanistycznych i społecznych.

Tekst jest udostępniony do wykorzystania w ramach dozwolonego użytku.

**THE USE OF THE GEOGRAPHICALLY WEIGHTED REGRESSION
FOR THE REAL ESTATE MARKET ANALYSIS**

Radosław Cellmer, Ph.D.

*University of Warmia and Mazury in Olsztyn
Faculty of Geodesy and Land Management
Department of Real Estate Management and Regional Development
Prawochenskiego 15, 10-720 Olsztyn, Poland
e-mail: rcellmer@uwm.edu.pl*

Received 17 September 2012, Accepted 15 November 2012

Abstract

The article presents a method for developing geographically weighted regression models for analyzing real estate market transaction prices and evaluating the effect of selected property attributes on the prices and value of real estate. The property attributes were evaluated on a grading scale to determine the relative (percentage) indicators characterizing the relationships on the real estate market. The market data were analyzed to evaluate the influence of infrastructure availability on the prices of land in Olsztyn. The results were used to assess the effect of every utility service on the property transaction prices.

Keywords: market analysis, geographically weighted regression, spatial models.

JEL classification: R32.

1. The use of regression models in analyses of transaction prices

A variety of qualitative and quantitative methods are used to evaluate the effect of property attributes on the prices and value of real estate. Qualitative methods, which are particularly effective on weakly developed markets, may produce subjective evaluations and opinions. Quantitative methods generate more subjective results, but their application necessitates the fulfillment of many formal and statistical requirements.

Regression analysis is one of the basic tools for modeling relationships between a dependent (explained) variable and one or more independent (explaining) variables. The simplest form of regression is expressed by the following linear model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

and if more than one explaining variables exist:

$$Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \varepsilon, \text{ for } i = 1, 2, \dots, n \quad (2)$$

where: Y – the explained variable which corresponds to the location i , X_i c explaining variables for the same location, ε – the error (residual), and $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ – modeled parameters (coefficients). The modeled parameters are generally determined by the least squares method where the sum of the squared differences between the observed value y_i and its estimator \hat{y}_i is minimized. The above can be written as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (3)$$

where: $\hat{\beta}$ – the vector of estimated parameters, X – the matrix of explaining variables, and Y – the vector of observed values. Multiple regression models for real estate market analyses have been discussed by numerous authors¹. Although multiple regression models are a convenient analytical tool, they are not often deployed in practice because they have to meet a series of formal requirements at the development stage².

The classical regression models used in real estate market analyses do not directly account for potential interactions (spatial autocorrelations) at the level of a given phenomenon in space, and they assume that the price-shaping process in geographic space is constant³. In this case, the significance of parameters in classical regression models is not affected by the spatial structure of the studied phenomenon, which could lead to the misinterpretation of results⁴, in particular on the assumption that real estate markets are characterized by spatial heterogeneity. The above is illustrated by Simpson's paradox⁵ which shows that non-spatial models may give a misleading

picture of market relations. A regression dependence between unit prices and area in square meters is shown in Figure 1.

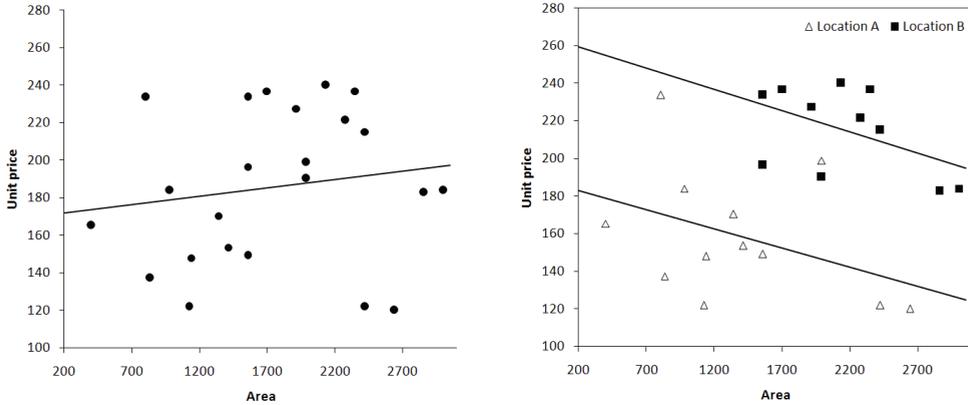


Fig. 1. Simpson's paradox on the example of correlations between property area and unit price
Source: own study based on Charlton and Fotheringham (2009).

An analysis of combined data from various locations points to an insignificantly positive correlation, whereas the dependencies observed in separate locations actually differ.

2. Geographically weighted regression

Various methods of accounting for the spatial structure of the studied phenomenon in regression models have been discussed in literature⁶. One of the reviewed methods proposes to assign weights to observations which, owing to their location in space, could have a theoretically greater impact on the analyzed phenomenon than other observations. The above can be expressed with the use of the geographically weighted regression.

The geographically weighted regression (GWR) is used on the assumption that the modeled parameters can be estimated separately at every point in space for which the values of the explained variable and explaining variables are known. The interactions between the studied objects in space are often characterized by the observation that elements found close to one another are more similar than objects that are further apart⁷. The above principle can be used to estimate the modeled parameters in a given location on the assumption that the observations made at points closer to the studied object will have greater weight than more distant observations⁸. A standard GWR model equation will take on the following form:

$$Y = \beta_0(x_i, y_i) + \beta_1(x_i, y_i) \cdot x_i + \varepsilon_i \quad (4)$$

or, for many independent variables:

$$Y = \beta_0(x_i, y_i) + \sum_{i=1}^n \beta_j(x_i, y_i) \cdot X_i + \varepsilon_i, \text{ for } j = 1, 2, \dots, n \quad (5)$$

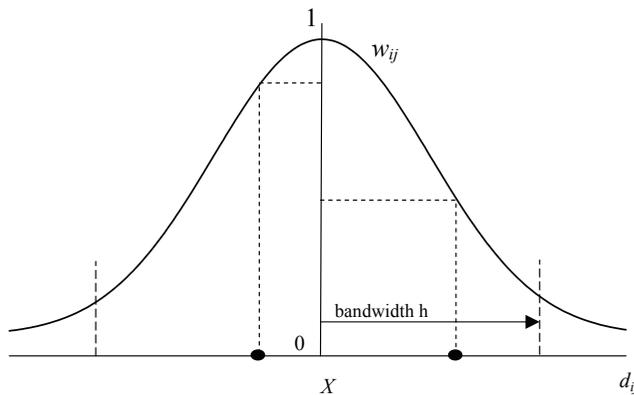
The value of the modeled parameters is determined by location which, in this case, is expressed by the coordinates (x_i, y_i) . The parameters of the GWR model are estimated similarly to classical models, but the weights of observations determined by location are taken into account:

$$\hat{\beta}(x_i, y_i) = (X^T W_{(i)} X)^{-1} X^T W_{(i)} Y \quad (6)$$

where $W_{(i)}$ is the matrix of weights that are a function of the distance between the location described by the coordinates (x_i, y_i) and the location of every observation point. The above matrix takes on a diagonal form:

$$W_{(i)} = \begin{bmatrix} w_{i1} & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & w_{in} \end{bmatrix} \quad (7)$$

and its elements can be expressed in various ways. To determine weights, we use spatial kernel functions which decrease along with an increase in distance from the point at which the geographically weighted regression model is estimated (Figure 2).



X – regression point, \bullet – data point, d_{ij} – distance between i and j , w_{ij} – weight for distance d_{ij}

Fig. 2. A spatial kernel function

Source: own study.

Weights are generally determined using functions with Gaussian-like distribution⁹, for example:

$$w_{ij} = e^{-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2} \quad \text{or} \quad w_{ij} = \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2 \quad (8)$$

where: d_{ij} is the distance between the locations i and j , and h is the bandwidth. The bandwidth indicates the spatial range of observations, which implies that $w_{ij} = 0$ for $d_{ij} > h$. The greater the bandwidth, the more the results of the GWR model are likely to approximate the global multiple regression model.

The application of the GWR model produced a series of fields mapped by the estimated parameters. The spatial variation in the value of those parameters suggests that the effect that the explained variables have on the explaining variable is characterized by a local variance and that the studied phenomenon is spatially heterogeneous¹⁰. Since the parameters are estimated only at the selected points in space, the local variance at any given point may be determined by spatial interpolation, and the result may be presented in the form of a map, such as an isarithmic map.

In a classical linear model of multiple regression the statistical tests of significance are commonly used to describe how well the model fits the data. Beside many criterions of the model evaluation, the coefficient of determination R^2 (or adjusted R^2) is also used. Models with a higher number of parameters are generally characterized by higher goodness of fit. The situation is somewhat more complex in the GWR models where the effective number of parameters is analyzed. A hat matrix S is multiplied by the empirical values of the explained variable to produce theoretical values¹¹:

$$\hat{y} = Sy \quad (9)$$

where S is a hat matrix expressed as:

$$S = X(X^T W_{(i)} X)^{-1} X^T W_{(i)} \quad (10)$$

The trace of the matrix S (the sum of the elements on the main diagonal) in the global model represents the number of parameters. In the GWR model, the effective number of parameters is calculated using the following formula:

$$2tr(S) - tr(S^T S) \quad (11)$$

The effective number of parameters in the model is determined by the number of explaining variables and the bandwidth and in most cases it is not an integer. The model's goodness of fit is generally evaluated with the use of the corrected Akaike information criterion which is expressed by the following formula¹²:

$$AIC_C = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left(\frac{n + \text{tr}(S)}{n - 2 - \text{tr}(S)} \right) \quad (12)$$

where:

- n – the number of observations in a data set,
- $\hat{\sigma}$ – the estimator for the standard deviation of residuals,
- $\text{tr}(S)$ – the trace of the matrix S .

The AIC_c criterion is applied to compare models with different numbers of explaining variables¹³ and to determine the “optimal” bandwidth (e.g. in the ArcGIS application). The bandwidth parameters which correspond to the lowest value of AIC_c are then applied to estimate the model.

Most property attributes have a qualitative character which impairs the effectiveness of statistical tools used in the analyses of prices on the real estate market. The above particularly applies to the location attribute which cannot be directly expressed in numbers without subjective evaluation. The geographically weighted regression does not directly determine the effect of location on property prices, but it may be used to build models which account for location not as an explaining variable, but as weights influenced by the distance between the evaluated properties.

The use of geographically weighted regression models in the analyses of transaction prices on the real estate market is broadly discussed in literature¹⁴. According to the referenced studies, the geographically weighted regression models are generally more accurate in illustrating the correlations on the property market than the global models. Spatial variations in the value of the estimated regression parameters often indicate that non-spatial attributes have a non-stationary effect on transaction prices.

The procedure of preparing data for a GWR analysis of transaction prices is generally similar to that applied in global models. Attributes should be evaluated on an interval scale at least to ensure that the correlations between the values of property attributes reflect price changes in a linear fashion (i.e. that they are correlated with prices). The explained variables should be weakly correlated. According to Bitner¹⁵, attribute evaluation scales should begin from zero and it is recommended that property parameters are expressed on a continuous, normalized scale [0; 1]. The above is not an absolute condition, but it facilitates the interpretation of the modeled parameters. In the

geographically weighted regression (i.e. $\beta_1, \beta_2, \beta_3$, etc.), the parameters are interpreted similarly to those in global models as a quota share per unit in which the explaining variable is expressed. In the discussed case, the value of a parameter is related to a point in space where the model is estimated. The constant value in model β_0 is interpreted as a theoretical value of the explained variable (transaction price) on the assumption that the value of all explaining variables is zero.

The modeled parameters are estimated independently at different regression points, therefore they may be characterized by a significant spatial variation. An analysis of a relatively large local market produces information about the absolute (quota) effect of the analyzed parameters on property prices in each location. The results can be averaged to obtain information about the attributes' average effect on prices (and the results are similar to those produced by a global model), but they are unlikely to be reliable. If the attributes' influence on prices were to be expressed in relative terms (as the percentage of the price), it could be applied to the entire market, provided that the analyzed correlation were spatially stationary. In global models, a relative effect can be determined in a simple way using the below formula:

$$p_i \% = \frac{\beta_i}{\bar{Y}} \cdot 100\% \quad (13)$$

In the GWR, the modeled parameter has a local character, therefore the average should make a reference to the local weighed average with the use of the same weighting method as that applied in the estimated model. The average is simple to calculate, but this option is rarely available in applications for developing geographically weighted regression models. In a GWR model the relative effect can be determined based on information about the theoretical value of an explained variable at a given point:

$$p_i(x_u, y_u) \% = \frac{\beta_i(x_u, y_u)}{\beta_0(x_u, y_u) + \sum_{i=1}^n \beta_i(x_u, y_u) X_i} \cdot 100\% \quad (14)$$

The constant value in the model can also be applied in the calculations on the condition that it corresponds to the theoretical value of the explaining variable (price) at a given point. The above condition can be met by applying evaluation scales where explaining variables can take on zero value. If the scale were chosen in such a way as to ensure that the zero value corresponded to a typical property, the resulting relative effect would have a universal character. The above influence at a given point can be represented as follows:

$$p_i(x_u, y_u) \% = \frac{\beta_i(x_u, y_u)}{\beta_0(x_u, y_u)} \cdot 100\% \quad (15)$$

A selected attribute's effect on the prices applicable to the entire market can be evaluated using a weighted average (owing to differences in the "quality" of each model):

$$p_i \% = \frac{\sum_{u=1}^n [v_u(x_u, y_u) \cdot p_i(x_u, y_u) \%]}{\sum_{i=1}^u v_u(x_u, y_u)} \quad (16)$$

where $v_u(x_u, y_u)$ is the weight of a given observation which may be determined by, for example, the average error of a determined parameter or a local determination coefficient.

3. The use of the geographically weighted regression to evaluate the effect of public infrastructure on transaction prices of land

An attempt was made to evaluate the usefulness of the geographically weighted regression in analyses investigating the effect of public infrastructure on the prices and value of land. A market data survey was carried out in the city of Olsztyn. The information about transaction prices was provided by the Register of Property Prices and Values kept by the Olsztyn City Office. The analyzed transactions involved privately-owned, undeveloped land plots, zoned for housing construction and traded in 2009–2011. The acquired data were processed, non-market transactions were excluded, and a final database of 298 transactions was compiled for the needs of the analysis. The location of traded properties is presented in Figure 3.

In addition to the location, a variety of other factors affect transaction prices and may be included in the model. Due to space constraints this article focuses solely on the effect that public infrastructure has on property prices. The omission of other explaining variables could affect the results, but the sole aim of this study was to verify the effectiveness of the geographically weighted regression method.

The location of public infrastructure utilities was determined based on the updated digital map of Olsztyn. The values of attributes characterizing public utilities were calculated in a spatial analysis (buffer zones were created around sections of infrastructure networks) using the ArcGIS software. For every utility service (w, e, s, g), the value of the explaining variable was set at zero on land plots situated within a 10 m radius from the infrastructure network. Land plots located within a 50 m distance from a utility point were assigned the value of 0.5. In the remaining cases, the value of the explaining variable was set at 1. It was assumed that a typical property has full utility access (the value of attributes for every utility service was zero).



Fig. 3. Spatial location of land plots traded in Olsztyn in 2009–2011 (represented by dots)

Source: own research.

The global model was estimated by the least squares method to produce the following results (Table 1).

Table 1. Results of a multiple regression analysis (global model)

$R^2 = 0.287$ $F(4.293) = 29.451$ $p < 0.0001$ Standard error of estimate: 88.346			
	β	St. error β	p
Y-intercept	291.818	7.840	0.000
Water	-91.413	23.692	0.000
Electricity	8.279	22.377	0.712
Sewerage	-8.849	21.819	0.685
Gas	-37.680	23.833	0.115

Source: own study.

Mains water supply was the only significant attribute at the significance level < 0.05 . This parameter had a nearly 40% influence on the average unit price of PLN 230.72/m².

When applied to the real estate market, global multiple regression models rarely meet formal and statistical requirements. The presumed linear character of correlations, mutual correlations between explaining variables and low stability of estimated relations produce results that are not highly reliable. Similar problems may be encountered when the GWR models are applied. In this case, however, numerous models are developed, and at least some of them meet standard requirements and can be reliably applied to evaluate the effect of selected property attributes on property prices.

The geographically weighted regression models were estimated with the use of the ArcGIS application. The relevant weights were calculated using the formula (8), and the bandwidth was determined based on the AIC_c criterion. The results are presented in Table 2.

Table 2. General results of the geographically weighted regression analysis

	β_0	β_1	β_2	β_3	β_4	Local R ²
Min	237.217	-131.125	-46.701	-243.402	-222.077	0.055
Max	364.714	86.689	112.388	54.417	118.085	0.582
Average	299.958	-65.456	-12.864	-47.313	-20.550	0.338

Source: own study.

From a group of 298 models for calculating the effect of public utility services on transaction prices, only those models that met basic evaluation criteria were selected, i.e. a negative sign before parameter β , relative to the adopted evaluation scale (the higher the grade, the fewer public utilities in the analyzed land plot). Absolute values of the modeled parameters were not compared or averaged due to a relatively large area of the analyzed market and, consequently, significant variations in land prices. A quota increase in price resulting from the availability of public utilities would differ in land plots with unit prices of PLN 100/m² and PLN 300/m². For this reason, the price-forming effects of infrastructure were determined with the use of relative indicators calculated according to the formula (15). An analysis of every GWR model produced a series of independent results. They were used to develop empirical probability distributions (histograms) of relative indicators which characterized the effect of every public utility on transaction prices (Figure 4).

The relative effect of every public utility service was determined using the formulas (15) and (16), and the inverse of the relative average error of the estimated parameter was adopted as the weight. The results of the analysis are presented in Table 3.

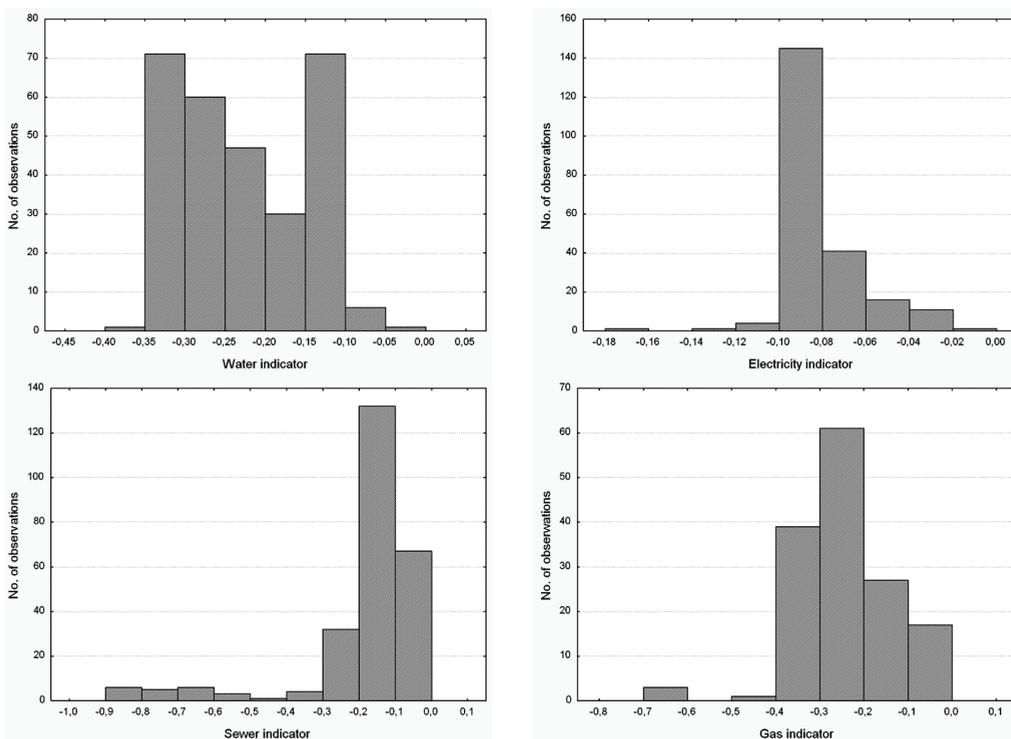


Fig. 4. Histograms of relative indicators characterizing the effect of public utility services (water, electricity, sewer and gas) on transaction prices

Source: own study.

Table 3. The effect of public utility services on transaction prices in percentage terms

Utility	Relative effect on transaction prices (%)
Water	-26.21
Electricity	-8.95
Sewerage	-27.34
Gas	-30.05

Source: own study.

Negative values are a reflection on the adopted evaluation scale. For example, the price of a land plot without mains water supply is likely to be more than 26% lower than the price of a property with direct access to this public utility service. The total effect of the analyzed utilities on transaction prices will reach:

$$\prod_{i=1}^k (1 + p_i) = 0.34 \quad (17)$$

The above implies that the prices of land plots where public utility services are not available will be approximately 66% lower than the prices of property with full utilities access.

Conclusions

Transaction prices and price change trends on the real estate market are significantly influenced by local factors, in particular the location of property. The impact of location on real estate prices obstructs the analyses of cause-and-effect relationships between property attributes and transaction prices. The geographically weighted regression supports the modeling of local correlations, and it largely eliminates the influence of location on property prices.

The evaluation of the effect of public utility services on property prices indicates that the geographically weighted regression is a useful tool for the real market analyses. The discussed example serves only a methodological purpose, which is why other factors such as shape and area were not taken into account. Models where the applied weights were determined by spatial correlations between the studied objects could produce superior results to classical regression models.

The article focused primarily on an utility service treated as an example of a factor having effect on prices. In the process of the real estate market analysis it is necessary to take into consideration the fact that price is being shaped under the influence of many other factors which are not related to location, such as the economic, demographic or social ones. When we treat location as one of the main price influencing factors, we also have to consider advantages and inconveniences connected for example with access to public transportation, schools, shops or green areas.

The geographically weighed regression helps evaluate the selected attributes' effect on transaction prices. The discussed method can also be used to analyze the spatial variation of different phenomena on the property market, so it is a valuable tool in the real estate management. The information about spatial variation patterns present on the market is particularly useful in the process of planning local development strategies, zoning plans and real estate tax planning.

Notes

- ¹ Bruce, Sundell (1977); Mark, Goldberg (1988); Czaja (2001); Hozer (2001); Wang, Wolverton (2002); Lis (2005); Adamczewski (2006); Bitner (2007); Barańska (2008); Sawiłow (2010).
- ² Hozer (2001).
- ³ Kulczycki, Ligas (2007).
- ⁴ Charlton, Fotheringham (2009).
- ⁵ Simpson (1951).
- ⁶ Swamy (1971); Caseti (1972); Anselin (1988); Haining (2003).
- ⁷ Tobler (1970).
- ⁸ Charlton, Fotheringham (2009).
- ⁹ Ibidem.
- ¹⁰ Ibidem.
- ¹¹ Brunson et al. (1999); Brunson et al. (2000).
- ¹² Akaike (1973); Hurvich et al. (1998).
- ¹³ Brunson et al. (2000).
- ¹⁴ Fotheringham et al. (2002); McCord (2012); Kulczycki, Ligas (2007).
- ¹⁵ Bitner (2007).

References

- Adamczewski, Z. (2006). *Elementy modelowania matematycznego w wycenie nieruchomości*. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.
- Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In: B. Petrov and F. Csaki (Eds.), *2nd Symposium on Information Theory*, Budapest: Akademiai Kiado (pp. 267–281).
- Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic Publishers.
- Barańska, A. (2008). Kryteria stosowania modeli stochastycznych w predykcji rynkowej wartości nieruchomości. *Zastosowania Metod Statystycznych w Badaniach Naukowych*. StatSoft, www.statsoft.pl/czytelnia.
- Bitner, A. (2007). Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości, *Acta Scientiarum Polonorum, Administratio Locorum*, 7, 1, 41–53.
- Bonnafous, A. & Kryvobokov, M. (2011). Insight into apartment attributes and location with factors and principal components. *International Journal of Housing Markets and Analysis*, 4, 2, 155–171. DOI: 10.1108/17538271111137930.
- Bruce, R.W. & Sundell, D.J. (1977). Multiple regression analysis: history and application in the appraisal profession. *Real Estate Appraiser*, Jan/Feb, 37–44.

- Brunsdon, C., Fotheringham, S. & Charlton, M. (1999). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science*, 39, 3, 497–524.
- Brunsdon, C., Fotheringham, S. & Charlton, M. (2000). *Geographically Weighted Regression as a Statistical Model*, Spatial Analysis Research Group, Department of Geography, University of Newcastle-upon-Tyne, UK.
- Caseti, E. (1972). Generating models by the expansion method: applications to geographic research, *Geographical Analysis*, 4, 81–91.
- Charlton, M. & Fotheringham, S. (2009). *Geographically weighted regression*. National Centre for Geocomputation, Maynooth, Ireland.
- Czaja, J. (2001). *Metody szacowania wartości rynkowej i katastralnej nieruchomości*. Kraków: Wydawnictwo Akademii Górniczo-Hutniczej w Krakowie.
- Fotheringham, S., Brunsdon, C., Charlton, M. (2002). *Geographically Weighted Regression – the analysis of spatially varying relationships*. New York: John Wiley & Sons, Ltd.
- Haining, R. (2003). *Spatial analysis of regional geostatistics data*. Cambridge University Press.
- Hozer, J. (2001). Regresja wieloraka a wycena nieruchomości. *Rzeczoznawca Majątkowy*, 2, 13–14.
- Hurvich, C.M., Simonoff, J.S. & Tsai, C.L. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of Royal Statistical Society*, B, 60, 271–293, DOI: 10.1111/1467-9868.00125.
- Kulczycki, M. & Ligas, M. (2007). Regresja ważona geograficznie jako narzędzie analizy rynku nieruchomości, *Geomatics and Environmental Engineering*, 1, 2, 59–68.
- Lis C. (2005). Ekonometryczne modele cen transakcyjnych lokali mieszkalnych, *Zeszyty Naukowe Uniwersytetu Szczecińskiego* 415. *Prace Katedry Ekonometrii i Statystyki* 16, 161–174.
- Mark, J. & Goldberg, M.A. (1988). Multiple regression analysis and mass assessment. A review of the issues. *Appraisal Journal*, 56, 89–109.
- McCord, M., Davis, P.T., Haran, M., McGreal, S. & McIlhatton, D. (2012). Spatial variation as a determinant of house price: Incorporating a geographically weighted regression approach within the Belfast housing market, *Journal of Financial Management of Property and Construction* 17, 1, 49–72. DOI: 10.1108/13664381211211046.
- Sawilów, E. (2010). Problematyka określania wartości nieruchomości metodą analizy statystycznej rynku. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 18, 1, 21–32.
- Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* B13, 238–241.
- Swamy, P. (1971). *Statistical inference in random coefficient models*. Berlin: Springer.
- Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 234–240.
- Wang, K. & Wolverton, M.L. (2002). *Real Estate Valuation Theory*. Berlin: Springer.