Spatial Statistics For Real Estate Data[1]

Marek Kulczycki, Marcin Ligas

Key words: spatial heterogeneity, spatial autocorrelation, spatial statistics, geostatistics, Geographical Information System

SUMMARY:

The paper presents spatial statistics tools in application to real estate data, including geostatistics, spatial autoregressive models and geographically weighted regression. All approaches, mentioned above, have different principles but complement each other.

Classic statistical methods often fail while having at hand autocorrelated or heteroscedastic data which are natural for real estate. For a long time, spatial autocorrelation or spatial heterogeneity were not taken into account. Last 10 years brought a great interest in employing spatial statistics methods to data spatial in nature such as real estate data what is partially caused by wildly developing GIS software.

The content of the paper includes geostatistical methods for localizing real estate submarkets (kriging interpolation) homogenous in respect of price, direct modeling of variance – covariance matrix later used in GLS estimation. The application of GWR (Geographically Weighted Regression) for spatial heterogeneity modeling and utilization of spatial autoregressive models for real estate data can be also found.

These quite new techniques in authors' opinion, give new opportunities in a field of real estate valuation both by localizing real estate submarkets, their analysis and finally appraisal process. It can be a good tool in mass appraisal (finding taxation zones), in prediction of results of different kind of activities connected with changing spatial planning and also as a supporting tool for decision making process concerning localization of investments.

1. INTRODUCTION

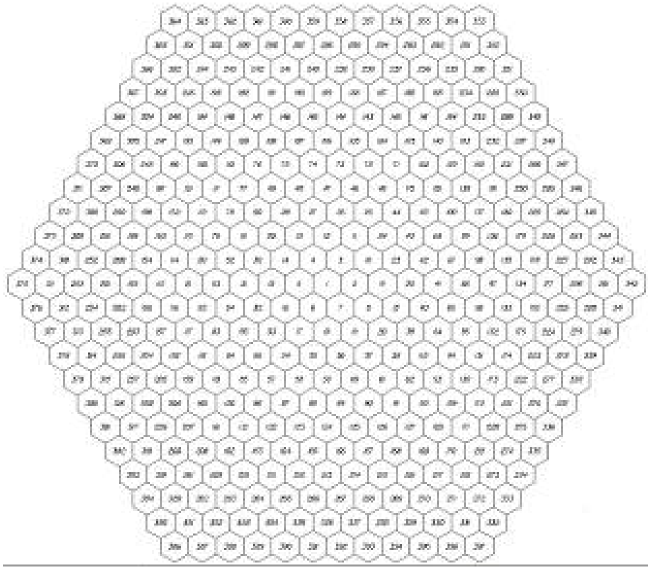
In the era of quickly developing GIS technologies tools of spatial statistics and econometrics gain value. Hitherto existing explorative techniques applied to analysis of phenomena and processes arising in geographical space become insufficient and inefficient. Increasing role in such analysis starts playing spatial location of the phenomenon or process, both absolute and relative. Application of GIS technologies stops limiting only to examining geometrical features of objects. Most often, complicated analysis on the attributes of objects stored in constantly growing databases are carried out.

[1] The paper was developed within the statutory researches of the Terrain Information Department, University of Science and Technology AGH, Krakow, Poland

No doubt, the data concerning real estate market have a spatial character. Every single real estate has its own fixed location in geographical space and a set of characteristics (attributes) which describe it. Thus, application of spatial statistics methods like: kriging, geographically weighted regression and spatial models; to real estate analysis seems to be fully justified and adequate to data we have at hand.

2. STUDY AREA "HONEYCOMB"

In the purpose of research the "honeycomb" was overlaid onto a map of a local real estate market with identical 397 fields with a radius of 50 m. Each of honeycomb's field was described by the coordinates of the centroid and a dominant unit price and dominant attributes for the properties within the field. Real estates of the analyzed market were described by means of four attributes, two of them were directly connected with spatial location and remaining two attributes takes values independent from location.



3. GEOSTATISTICAL APPROACH

The beginning of new statistical method for describing variability – geostatistics – falls into sixties of past century along with publishing works by George Matheron. Next years brought fast development of theoretical aspects and practical applications of this new method. At present, it is used in such fields like: geology, hydrology, meteorology, oceanography, geography, geodesy, photogrammetry, forestry and many others and also increasing interest in applying these methods to real estate valuation is observable.

Statistical analysis of variability of distinguished characteristic by using classical methods (mean value, variance, coefficient of variability and so on) do not include spatial configuration of the feature. Direct incorporating spatial location of a feature made up a novelty of this method.

In a real estate market researches, knowledge on spatial autocorrelation (at least intuitive) seems to be common, it has its expression in shaped up local markets – markets with similar values of real estate prices. Traditional statistical methods widely used in analysis of real estate markets assume independence of observation in geographical space. Assumption of spatial independence destroys correct inference concerning analyzed market and behavior of its participants. Knowledge on spatial dependence imply employment of adequate methods to data we have. Analyzing spatial dependence on real estate markets – geostatistics – cannot be overlooked.

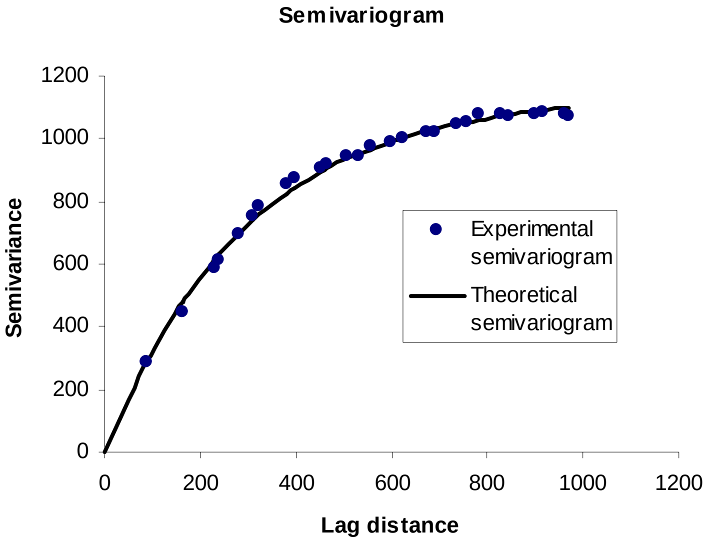
The fundamental function used in geostatistics is a semivariance, given by the formula:

$$\hat{\gamma}(h)=\frac{1}{2\cdot N(h)}\sum_{i=1}^{N(h)}\left[z(s_i+h)-z(s_i)\right]^2$$

where:
$z(s_i)$, $z(s_{i+h})$ – are the values at point being calculated and a value at a point distance h away
N(h) – the number of pairs for the distance h

The semivariance describes relation between the average differentiation of values of distinguished feature observed in given points and the distance between these points. The structure of variability in synthetic form is described by semivariogram function which is a plot of a semivariance vs. distance. For such constructed empirical semivariogram the theoretical semivariogram must be estimated, in other words we need mathematical model describing relations between semivariance and a distance (best fit). The theoretical model of semivariogram function gives opportunity of determining the range of spatial autocorrelation (A – range), contribution of random term and a non – random term for an arbitrary distance between points h.

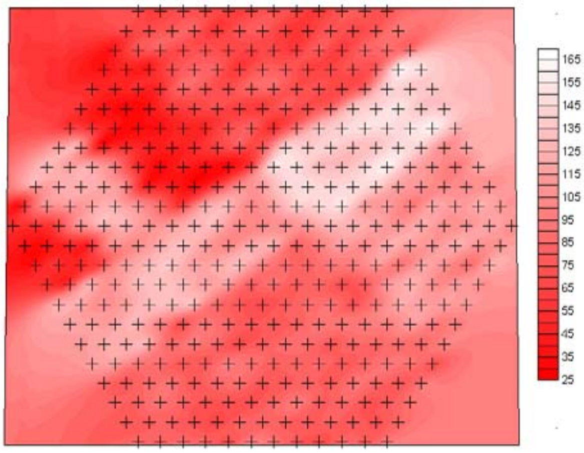
Below, there is a plot of empirical semivariogram and a fitted theoretical semivariogram (exponential model) for the distinguished characteristic – property price for the spatial configuration of “HONEYCOMB” market. It can be clearly observed spatial autocorrelation existing on this market decreasing along with a increasing distance between real estates and fading after reaching certain range of influence called effective range.



Having at hand a knowledge on functional form of theoretical semivariogram, the spatial interpolation technique called kriging can be applied. In the content of this paper, kriging method was applied to searching local markets (homogenous in respect of price) and on its basis preparing map of spatial distribution of prices in geographical space; figures below: A – kriging method; B – actual distribution of prices in space. This kind of figuration may successful serves as a tool supporting entire appraisal process and can be applied to simulation analysis of phenomena occurring on real estate markets with given a priori boundary conditions.

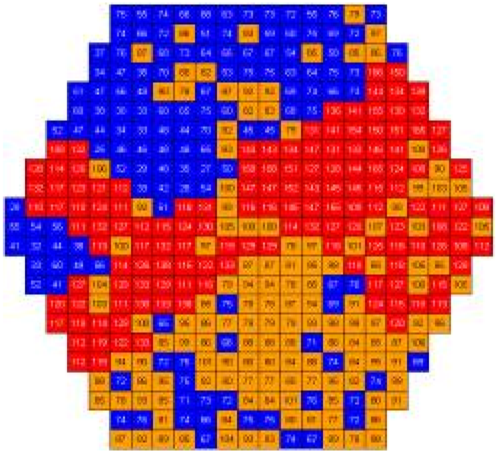
Semivariogram ⇒ Kriging procedure ⇒ Map

A)



Result of spatial interpolation via kriging method (A)

B)



Theoretical spatial distribution of prices (B)

In works by T. H. Thibodeau, direct modeling of covariance matrix (between locations $s_i = (x_i, y_i)$ and $s_j = (x_j, y_j)$) and further application of Estimated Generalized Least Squares in appraisal model estimation can be found. This procedure may be written in a following way:

$$\boxed{\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow \hat{\boldsymbol{\varepsilon}}}$$

$$\Downarrow$$

Semivariogram for ($\hat{\boldsymbol{\varepsilon}}$)

$$\gamma\left(s_i - s_j\right) = \gamma\left(h\right) = \frac{1}{2 \cdot N\left(h\right)} \sum_{i=1}^{N(h)} \left(\hat{\varepsilon}\left(s_i\right) - \hat{\varepsilon}\left(s_j\right)\right)^2$$

Under the condition of second - order stationarity (spatially constant mean and variance)

$$\gamma\left(h\right) = C\left(0\right) - C\left(h\right) \Rightarrow C\left(h\right) = C\left(0\right) - \gamma\left(h\right)$$

$$\Downarrow$$

$$\boxed{\boldsymbol{\Omega} = \mathbf{C}\left(h\right) \Rightarrow \mathbf{W} = \boldsymbol{\Omega}^{-1} = \left[\mathbf{C}\left(h\right)\right]^{-1}}$$

$$\Downarrow$$

EGLS estimation

$$\hat{\boldsymbol{\beta}}_{EGLS} = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \mathbf{W} \mathbf{y}$$

4. GEOGRAPHICALLY WEIGHTED REGRESSION APPROACH

Application of Geographically Weighted Regression (GWR) to real estate analysis enables direct including spatial heterogeneity (non – stationarity) of analyzed phenomenon as is spatial distribution of prices. In contradiction to classical regression techniques the result of GWR is construction local models corresponding to a particular location rather than one global model for entire market. In other words, every single location obtains its own set of regression parameters.

GWR model can be written as follows:

$$y_i = \beta_o\left(x_i, y_i\right) + \sum_{j=1}^{k} \beta_j\left(x_i, y_i\right) \cdot x_{ij} + \varepsilon_i,\ i = 1, \ldots, n$$

where:
x_i, y_i – the coordinates of i – th location
$\beta_j(x_i, y_i)$ – the j – th regression parameter in point (for property) i – th
ε_i – error term
x_{ij} – j – th characteristic of property i – th

Local parameters $\beta_j(x_i, y_i)$ of GWR model are estimated by using weighted least squares method with weighting matrix depending on locations. Weighting matrices are diagonal matrices for which elements $w_{(i)j}$ are functions of distance between locations i – th and j – th. From the assumption, locations closer to (i) have greater influence, greater weight, than these ones further away. For such defined model, for each location (x_i, y_i) we get set of regression coefficients in form:

$$\hat{\boldsymbol{\beta}}(x_i, y_i) = \left(\mathbf{X}^T \mathbf{W}_{(i)} \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \mathbf{W}_{(i)} \mathbf{y}$$

For each location (x_i, y_i) we obtain also fitted values of y according to a formula:

$$\hat{y}_i = \mathbf{x}_i \cdot \boldsymbol{\beta}(x_i, y_i)$$

and a set of residuals at all locations (x_i, y_i):

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i \cdot \boldsymbol{\beta}(x_i, y_i)$$

On a basis of such determined parameters and other quantities describing the model, the standard statistical measures like: coefficient of determination, residual variance and so on and also appropriate statistical test can be constructed.

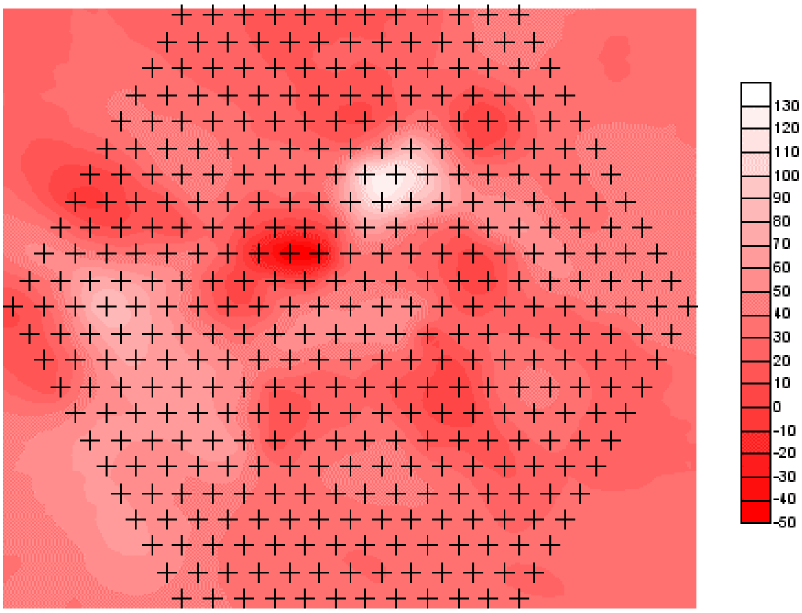
Below, there is a short summary of results obtained form application GWR to the set of market information from “HONEYCOMB”. The regression parameters are not constant over the entire study area, they change along with location and the neighborhood of actually considered location expressed by weight matrix. Statistical level of significance of GWR parameters also varies over the space, in one locations significant components of price (hedonic prices of particular attributes) lose its significance in others, what gives evidence, that not in all regions of analyzed market the same attributes influence the price in the same way. Thus, GWR can be a very good tool for analyzing spatial heterogeneity of real estate markets.

Table 4.1: The results form estimation global OLS model and GWR local models for the „HONEYCOMB” market.

	β_0	β_1	β_2	β_3	β_4	R^2	$\tilde{R}^2$	σ^2
OLS	57.59	16.69	-0.15	5.80	0.30	0.11	0.10	835.21
GWR_{MIN}	-76.195	-52.598	-1.5491	-24.514	-54.963	R^2 =0.9200		
GWR_{MAX}	106.77	124.25	2.3036	26.639	51.757	$\tilde{R}^2$ = 0.9192		

OLS – Ordinary Least Squares, GWR – Geographically Weighted Regression, R^2 – coefficient of determination, σ^2 – residual variance,

Plot 4.1: Relation between regression coefficient β_1 and its location in geographical space.



5. MEASURES OF SPATIAL AUTOCORRELATION AND SPATIAL MODELS

From the theoretical point of view spatial autocorrelation seems to be a common phenomenon on real estate markets, although, it is not very often included in classical models constructed in purpose of describing the behaving of particular market. This phenomenon can have its explanation in the behavior of market participants (buyers, sellers) who in decision making process take into consideration prices of nearby properties and also from the fact of sharing by the properties almost the same localization and what follows sharing almost the same accessibility, neighborhood and environmental characteristics.

Spatial statistics and spatial econometrics possess tools detecting these kind of dependence. On account of limited capacity of this study we constrain to present one of many possible.

One of commonly used measure of spatial correlation is Moran's I statistic, brought to life along with the article by P. A. P. Moran "The interpretation of statistical maps" in 1948. Although, over a half of a century has passed it is still a standard in analysis of spatial autocorrelation. Moran's I statistic expresses as follows:

$$I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(y_i - \hat{y}\right)\cdot\left(y_j - \hat{y}\right)}{\sum_{i=1}^{N}\left(y_i - \hat{y}\right)}$$

where:

y_i – the distinguished characteristic in location i – th

y_j – the distinguished characteristic in location j – th

$\hat{y}$ – the mean value of distinguished characteristic

N – the number of locations

w_{ij} – element w_{ij} of standardized spatial weight matrix (the sum of elements in each rows equals to unity). The literature of the subject offers considerable number of matrix of spatial structure (distance based, delaunay triangulation, nearest neighbors)

Spatial autocorrelation coefficient Moran’s I is a measure of clasterization and reveals to how extent high/low values of distinguished characteristic are surrounded by other high/low values of the characteristics.

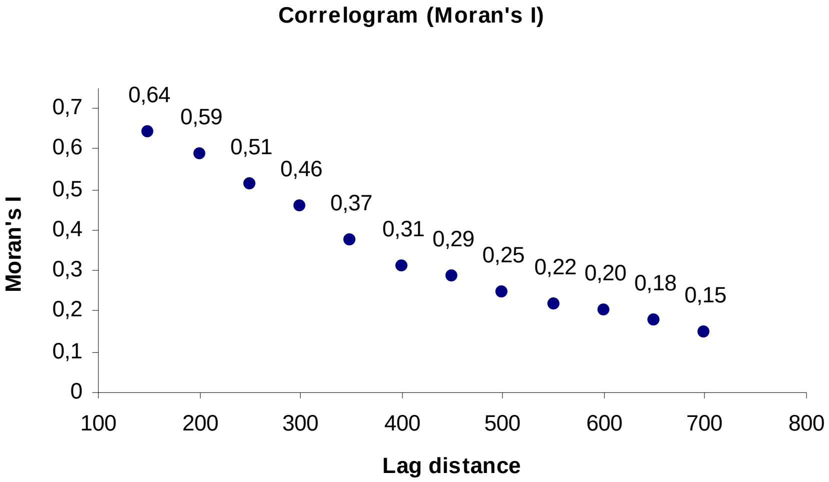
On the basis of Moran’s I statistics, the significance tests of spatial autocorrelation may be constructed (with different assumptions) for which the null hypothesis Ho is lack of significant spatial autocorrelation (spatial structure has no meaning in analysis, values do not appear to form any spatial pattern) vs. alternative hypothesis H_A that such correlation exists.

The table below presents values of Moran’s I statistics based on spatial structure matrix constructed on the basis of simple distance based criterion and the significance level of the statistics, for different distance thresholds.

Table 5.1: Values of Moran’s I statistics for different threshold distance

Weights	**Moran’s I**	**p - value**
W_150	0.6425	0.001
W_200	0.5870	0.001
W_250	0.5129	0.001
W_300	0.4589	0.001
W_350	0.3742	0.001
W_400	0.3107	0.001
W_450	0.2856	0.001
W_500	0.2473	0.001
W_550	0.2165	0.001
W_600	0.2007	0.001
W_650	0.1779	0.001
W_700	0.1487	0.001

Moran’s I correlation coefficient may also be used to graphical presentation of changes in spatial autocorrelation along with the distance in form of correlogram (the figure below, on the basis of data from table 5.1)



In the monograph (Spatial autocorrelation 1973) Cliff and Ord proposed also the significance test of spatial autocorrelation for the residual vector from the regression model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. The test is based also on Moran's I statistics, which in case of residual vector can be written as follows:

$$I = \frac{\hat{\boldsymbol{\varepsilon}}^T \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}$$

With the assumption of genuineness of the H_0 hypothesis about lack of significant spatial correlation and error term $\boldsymbol{\varepsilon}$ follows normal distribution than the distribution of Moran's I statistics may be approximated by normal distribution. Statistical inference on spatial autocorrelation significance is based on standardized version of Moran's I, with known theoretical values of the distribution parameters E(I) and V(I), expressed as:

$$E(I) = \frac{tr(\mathbf{MW})}{N-k}$$

$$V(I) = \frac{tr\left(\mathbf{MWMW}^T\right) + tr(\mathbf{MW})^2 + [tr(\mathbf{MW})]^2}{(N-k)(N-k+2)} - E(I)^2$$

$$\mathbf{M} = \left(\mathbf{I} - \mathbf{X} \cdot \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T\right)$$

where:

$\hat{\boldsymbol{\varepsilon}}$ – the residual vector from the model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$
W – the matrix of spatial structure
X – the matrix of independent variables
I – the unity matrix
N – the number of observations

k – the number of estimated parameters
tr(•) – the operator of matrix trace

In case of occurrence of spatial correlation in set of market information conventional regression models give biased and ineffective estimators, it results from the fact that the assumption of independence of observation is not fulfilled. The mean which considers and describes such a correlation in set of market information is application the spatial autoregressive models which are generalization of conventional models with reference to spatial problems.

Generally, spatial autoregressive models can be described by means of the following equations:

Type A:

$$y_i = f\left(y_j\right) + \mathbf{X}_i \cdot \boldsymbol{\beta} + \varepsilon_i$$

Type B:

$$y_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + u_i,\ u_i = f\left(u_j\right) + \varepsilon_i$$

As an example of model of type A, we have:
Spatial Autoregressive Model (spatially lagged dependent variable)

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

where:
y – the vector of dependent variable
W – the matrix of spatial structure
X – the matrix of independent variables
ε – the residual vector ε ~ N(0,σ2I)
β – the vector of regression coefficients
ρ – the autoregressive parameter

As an example of model of type B, we have:
Spatial Error Model (Error term with a spatial structure)

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$$
$$\mathbf{u} = \lambda \cdot \mathbf{W} \cdot \mathbf{u} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

where:
y – the vector of dependent variable
W – the matrix of spatial structure
X – the matrix of independent variables

u – the vector of error term with a spatial structure u ~ $N(0,\sigma2\Omega)$
ε – the pure residual vector ε ~ $N(0,\sigma2I)$
β – the vector of regression coefficients
λ – the autocorrelation coefficient

Below, there are mentioned some reasons of autocorrelation of error term (based on Welfe 2003):

- The nature of some social, economic processes
- Psychology of decision making process, the actions from the close surroundings have its influence
- Incorrect analytical form of the model
- Faulty dynamic structure of the model, lack of lagged variables
- Omission of important independent variable in the model specification

There is a question, which specification of the model should be chosen?, whether the model of type A is the appropriate model or perhaps the model of type B. Then, appropriate tests on model specification come to help, based on Lagrange multiplier test: $LM_{(lag)}$, Robust $LM_{(lag)}$, $LM_{(error)}$, Robust $LM_{(error)}$.

For the real estate database from "HONEYCOMB" market the spatial error model (SEM) appeared to be more appropriate model (residuals from the OLS model highly correlated). Below, there are results from the estimation conventional OLS model and two SEM models with two different spatial weight matrices (distance based criterion and delaunay criterion). By using spatial models we get better fit to the empirical data (table below) measured by means: coefficient of determination, residual variance and maximized value of likelihood function.

Table 5.2: The results form estimation global OLS model and SEM models for the „HONEYCOMB" market.

	βo	β_1	β_2	β_3	β_4	R^2	$\tilde{R}^2$	σ^2
OLS	57.59	16.69	-0.15	5.80	0.30	0.11	0.10	835.21
$SEM_{W\ 150}$	46.85	27.56	-0.28	4.60	0.88	0.74	0.74	240.42
λ	0.908		Llike		-1689.36			
SEM_{DEL}	51.03	24.50	-0.25	4.93	0.60	0.75	0.75	229.93
λ	0.877		Llike		-1548.49			

OLS – Ordinary Least Squares, SEM – Spatial Error Model, λ – coefficient of autocorrelation, Llike – Log Likelihood, R2 – coefficient of determination, σ2 – residual variance, subscript "w_150" – spatial weight matrix with threshold distance of 150 m., subscript "DEL" – spatial weight matrix based on delaunay triangulation

6. CONCLUSIONS

The tools of spatial statistics including geostatistics bring new explorative opportunities on real estate markets. Using these methods, we obtain a brighter image of processes and changes appearing on real estate markets. Suitable software adjusted to needs of real estate market analysis would be inestimable to dynamic, changing in time and space, analysis of these markets in spatial and generic categories. Depending on needs and character of researches, the methods mentioned in the content of this article may be successfully applied and discover these properties of market which were invisible using classical methods.

In authors' opinion, the statistical methods presented in brief, give new opportunities in a field of real estate valuation both by localizing real estate submarkets, their analysis and finally appraisal process. It can be a good tool in mass appraisal (finding taxation zones), in prediction of results of different kind of activities connected with changing spatial planning and also as a supporting tool for decision making process concerning localization of investments.

From the statistical point of view, a usage of spatial statistics method gives us more accurate estimators enabling more precise inference what means in practice that we have more explicit insight in mechanisms and processes occurring on real estates market then previously.

REFERENCES

1. Anselin L., 2004, Advances in spatial econometrics., Berlin, Springer.
2. Anselin L.,1988, Spatial Econometrics: Methods and Models (Studies in Operational Regional Science). Kluwer Academic Publishers.
3. Bavaud F., 1998, Models for spatial weights: a systematic look. Geographical analysis 30. 153 – 171.
4. Fotheringham A. S. Brunsdon C. Charlton M., 2002, Geographically Weighted Regression – the Analysis of Spatially Varying Relationships., Chichester, Wiley
5. Greene W. H., 2003, Econometric analysis. Prentice Hall.
6. Isaaks. E. H. and R. M. Srivastava., 1989, An Introduction to Applied Geostatistics. Oxford Univ. Press. New York
7. LeSage J. P., 1999, Spatial Econometrics. The Web Book of Regional Science. Regional Research Institute. West Virginia University. Morgantown.
8. Pace R. K. Barry R. Sirmans C. F., 1998, Spatial statistics and real estate. Journal of Real Estate Finance and Economics 17. 5 – 13.
9. Welfe A.,2003, Ekonometria, PWE, Warszawa
10. Zeliaś A.(praca zbiorowa),1991, Ekonometria przestrzenna. PWE. Warszawa.

BIOGRAFICAL NOTES

Name: Marek Kulczycki
Employment: University of Science and Technology AGH, Krakow, Poland
Degree: Doctor of Technical Sciences
Membership of Professional Bodies:
Polish Real Estate Scientific Society
Present Position: Lecturer

Name: Marcin Ligas
Employment: University of Science and Technology AGH, Krakow, Poland
Degree: Doctor of Technical Sciences
Membership of Professional Bodies:
Polish Real Estate Scientific Society
Present Position: Lecturer

CONTACTS

PhD **Marcin Ligas**
University of Science and Technology
Faculty of Mining Surveying and Environmental Engineering
Terrain Information Department
Al. A. Mickiewicza 30
30 – 059 Krakow
POLAND
Tel. + 48 12 617 44 80
Fax + 48 12 617 22 77
Email: marcin.ligas@agh.edu.pl

PhD **Marek Kulczycki**
University of Science and Technology
Faculty of Mining Surveying and Environmental Engineering
Terrain Information Department
Al. A. Mickiewicza 30
30 – 059 Krakow
POLAND
Tel. + 48 12 617 44 80
Fax + 48 12 617 22 77
Email: marek.kulczycki@agh.edu.pl