ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Jan Kowalik**

KRIGING – A METHOD OF STATISTICAL INTERPOLATION OF SPATIAL DATA

Abstract. In analyses of spatial phenomena it can sometimes be impossible or very expensive to obtain the value (realization) of a studied phenomenon at all its locations because of practical constraints. Then in order to estimate the values of the variables at these locations one can resort to a geostatistical method of data estimation (interpolation) called kriging. Kriging is a basic method of spatial data estimation used in geostatistics that interpolates unknown values of the regionalized (spatial) variable from its known values at other locations. The following analysis is set out to present the basic assumptions of geostatistics, theoretical grounds of kriging, its types and its applications in various areas of life.

Key words: geostatistics, regionalized variable, kriging, analysis of variograms.

1. INTRODUCTION

Spatial continuity is typical for many natural, sociological and economic phenomena. However in reality it is impossible to obtain an exhaustive number of the realization of a studied phenomenon at every desired point in the space. It is mainly because of financial and natural constraints, e.g. the landscape makes it often impossible or very expensive to conduct measurements at a given location. Thus in order to determine the spatial distribution of a studied phenomenon one can take advantage of geostatistical methods. Geostatistics is a branch of statistics that uses spatial continuity of phenomena and adapts methods of classical regression in order to use this continuity to determine (estimate, interpolate) spatial distributions of studied phenomena. Geostatistical theory is based on the following

* M.Sc., Chair of Econometrics and Statistics, Technical University in Częstochowa.

assumption: beside a point of the determined value of a certain variable there are points of similar values. So it can be said that the realizations of a studied phenomenon that are separated from one another are correlated.

The first scientist to pay attention to the importance of spatial continuity in the estimation of the distribution of phenomena was Kriege in the middle of the 20th century (1951) and he used the spatial continuity in the mining of the gold deposits in the Republic of South Africa. The theory put forward by D. G. Kriege was developed by a French mathematician, a would-be professor G. Matheron (1962). Since then geostatistics has become quite a popular field of knowledge boasting many various applications ranging from its earliest implementations in mining industry to geology, pedology, environment protection and economic issues.

The basic aim of this analysis is to present the kriging method – a basic method of spatial data estimation used in geostatistics, as well as its types and applications in various areas of life.

2. RULES OF GEOSTATISTICS

The most characteristic aspect of geostatistics is the notion of regionalized variables, which have properties intermediate between random variables and completely deterministic variables. The regionalized variable is a variable with two aspects: a local random erratic aspect and a structured aspect, which reflects the complexity of this phenomenon.

Geostatistics is based on the notion of the random function according to which the collections of the values of the parameter $z(x)$ at all locations x are regarded as individual realizations from the collection of the space of the dependent random variables $Z(x)$.

The analysis of spatial data with the use of geostatistical methods requires the knowledge of the first two statistical moments attributed to the random function for a given phenomenon, namely:

– the first-order statistical moment (mean)

$$E[Z(x)] = m(x) \tag{1}$$

– the second-order statistical moments:

a) the variance

$$D^2[Z(x)] = E[Z(x) - m(x)^2] \tag{2}$$

b) the covariance which is a function of the location x_1 and x_2

$$C(x_1, x_2) = \mathrm{E}[(Z(x_1) - m(x_1)) \cdot (Z(x_2) - m(x_2))] = \\ = \mathrm{E}[Z(x_1)] \cdot \mathrm{E}[Z(x_2)] - m(x_1) \cdot m(x_2) \quad (3)$$

c) the semivariogram defined as a half of the variances of the difference of random variables at two different locations x_1 and x_2

$$\gamma(x_1, x_2) = \frac{1}{2} \mathrm{D}^2[Z(x_1) - Z(x_2)] \quad (4)$$

In order to use various methods of geostatistical analysis a given phenomenon or process has to be stationary, i.e. it must not change its properties with the change of the beginning of the time scale or the spatial scale. The stationarity of spatial data is determined by means of the so-called hypotheses of stationarity. There are usually two main types of stationarity (Meul, Von Meirvenne 2003): stationarity in a narrower sense (first-order stationarity) and stationarity in a wider sense (second-order stationarity or weak stationarity).

The random function is stationary in a narrower sense when all the moments of its distribution remain invariable in relation to the displacement vector. However in reality the assumptions of the stationarity of the random function of a given variable are hardly fulfilled.

So the most frequently used hypothesis of stationarity is the second-order stationarity restricted only to the first two statistical moments of the random function. It can be said that the random function $Z(x)$ is weakly stationary when there is mathematical expectation $\mathrm{E}[Z(x)]$ which does not depend on the location of the point x

$$\mathrm{E}[Z(x)] = m(x) \quad \forall x \quad (5)$$

and when there is a covariance for each pair of random variables $[Z(x), Z(x + h)]$ which depends only on the separation vector h

$$C(h) = \mathrm{E}[Z(x + h) \cdot Z(x)] - m^2 \quad \forall x \quad (6)$$

The hypothesis of the second-order stationarity can be restricted when one assumes the stationarity of the first two moments from the random function increments. In such a case we define the stationarity as intrinsic stationarity. The hypothesis of intrinsic stationarity assumes that there is a variance and a mean of the random function increments $[Z(x + h) - Z(x)]$ and they do not depend on the location vector x

$$E[Z(x+h) - Z(x)] = 0 \quad \forall x \tag{7}$$

$$D^2[Z(x+h) - Z(x)] = E[(Z(x+h) - Z(x))^2] = 2\gamma(h) \tag{8}$$

where $\gamma(h)$ is the semivariance.

Depending on the type of the stationarity of the random function of a studied phenomenon various methods of methodology of geostatistical analysis are used.

3. THE THEORY OF KRIGING

Kriging is a basic method of spatial data estimation used in geostatistics. Kriging in a method of creating optimal unbiased estimations of regionalized variables at unsample locations which uses the hypothesis of stationarity and the structural properties of the covariance as well as the initial collection of data. In the literature on this subject the estimator of kriging is defined as the best linear unbiased estimator (BLUE). It is the best method because the estimation error is minimized. It is linear because the estimation Z^* is performed at an unknown location from the weighted sum Z_i of the available data (Borga, Vizzaccaro 1997)

$$Z^* = \sum_{i=1}^{n} \omega_i Z_i \tag{9}$$

where:

ω_i – weights determined for each measurement point,

Z_i – the values of the regionalized variable in measurement points,

n – number of dispersed measurement points in the collection.

The kriging weights sum to unity to ensure the unbiasedness of the estimator and it is written in the following way:

$$\sum_{i=1}^{n} \omega_i = 1 \tag{10}$$

Kriging is based on the variogram function that is also known as semivariogram or semivariance. Semivariogram is a function of the structure of the regionalized variables that presents a spatial or time behavior of this variable in a studied collection of data.

A sample semivariance is defined as a half of the variance of the difference of random variables at the separation vector h distance

$$\gamma(h) = \frac{1}{2} D^2[Z(x) - Z(x+h)] \tag{11}$$

The estimation of the variogram boils down to the calculation of the empirical variogram of the studied regionalized variable followed by finding a suitable theoretical variogram to match its course.

The empirical variogram describes spatial correlations of the random sample. It is a curve (vector) that is formed from the graphic representation of the dependence of semivariance on the distance h between the measurement points. The experimental semivariogram for the distance h is determined from the equation (Ploner, Dutter 2000)

$$\gamma(h) = \frac{\sum_{i=1}^{N(h)} [Z(x) - Z(x+h)]^2}{2N(h)} \tag{12}$$

where:

$N(h)$ – the number of pairs of points at the h distance.

The theoretical variogram describes spatial correlations for the source population and it is a simple mathematical function that models the trend in the empirical variogram. Among commonly used theoretical semivariograms we can distinguish (Ploner, Dutter 2000) and (Zawadzki 2002) the following ones: the spherical model, the nugget model, the linear model, the exponential model and the Gaussian model. These models are referred to as basic models. The equations of the models are as follows:

a) the nugget model

$$\gamma(h) = \begin{cases} 0, & \text{if } h = 0, \\ 1, & \text{if } h \neq 0, \end{cases}$$

b) the spherical model

$$\gamma(h) = \begin{cases} c_0 + c_1\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right), & \text{if } |h| \leqslant a, \\ c_0 + c_1, & \text{if } |h| > a, \end{cases}$$

c) the exponential model

$$\gamma(h) = c_0 + c_1\left(1 - e^{-\frac{h}{a}}\right),$$

d) the Gaussian model

$$\gamma(h) = c_0 + c_1\left(1 - e^{-\frac{h^2}{a^2}}\right),$$

e) the linear model

$$\gamma(h) = c_0 + bh,$$

where in all models:

c_0 – the nugget effect,
a – the range,
c_1 – the sill,
b – the ratio of the sill c_1 to the range a.

The above-mentioned elements, the nugget effect, the sill and the range are the three characteristic parameters of the semivariogram as shown at the Fig. 1.

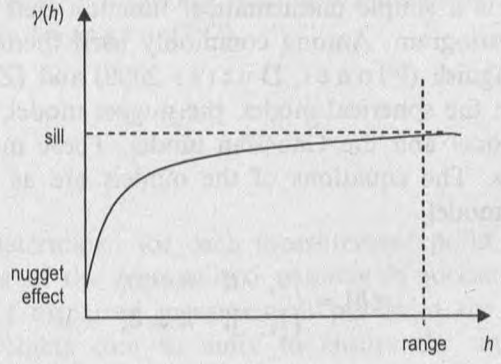

Fig. 1. The most important characteristics of the semivariogram

Source: own elaboration.

The sill is the highest value of the variogram at which an increment in the function is no longer observed. The sill should be approximately equal to the variance from the sample.

The nugget effect is a constant that in the variogram model represents a certain variation of data at a scale smaller than the sampling range. It can also be caused by the sample error (Francois-Bogarcon 2004). This quantity is viewed in the semivariogram as an absolute term.

The range is a scalar quantity that controls the degree of correlation between the points of data that are usually represented as the distance. It is a distance range from zero to point of reaching by the semivariogram 95% of the constant value.

After the theoretical and the empirical models of the semivariogram have been constructed they can be used to determine the kriging weights ω_i. For example to perform an interpolation at point P on the basis of the data from the three surrounding points P_1, P_2 and P_3 we use the following kriging equation:

$$Z_p^* = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3 \tag{13}$$

Using the semivariogram and the condition that the sum of kriging weights should be equal to one we can construct the following system of equations to calculate the weights ω_i

$$\begin{aligned}
&\omega_1\gamma(h_{11}) + \omega_2\gamma(h_{12}) + \omega_3\gamma(h_{13}) + \lambda = \gamma(h_{1p})\\
&\omega_1\gamma(h_{21}) + \omega_2\gamma(h_{22}) + \omega_3\gamma(h_{23}) + \lambda = \gamma(h_{2p})\\
&\omega_1\gamma(h_{31}) + \omega_2\gamma(h_{32}) + \omega_3\gamma(h_{33}) + \lambda = \gamma(h_{3p})\\
&\omega_1 + \omega_2 + \omega_3 + 0 = 1
\end{aligned} \tag{14}$$

where:

$\gamma(h_{ij})$ – is the semivariance of the distance h between the control points i and j,

$\gamma(h_{ip})$ – is the semivariance of the distance h between the control point i and the interpolated point p,

λ – absolute term of the equation.

The value of a studied phenomenon at point P is obtained by solving the system of equations (14) in relation to the weights ω_i and by putting the values of weights to the equation (13).

An important characteristic of kriging is that the variogram can be used to determine the estimation error at every point of the interpolation. The variance can be calculated using the following formula:

$$S_e^2 = \omega_1\gamma(h_{1p}) + \omega_2\gamma(h_{2p}) + \omega_3\gamma(h_{3p}) + \lambda \tag{15}$$

4. TYPES OF KRIGING

Taking into account various degrees of stationarity of spatial data used in geostatistical analysis we can distinguish two basic kriging techniques: ordinary kriging and universal kriging. These two kriging techniques can be used as point kriging (it estimates the value of a studied phenomenon at a given point) or as a block kriging (it provides the mean value of a studied quantity in a certain 2D or 3D area).

Ordinary kriging is one of basic kriging methods. At the unsample location x_0 the value of the regionalized variable is estimated as (Rehman, Ghori 2000)

$$Z^*(x_0) = \sum_{i=1}^{n} \omega_i Z(x_i) \tag{16}$$

where:

$Z^*(x_0)$ – the estimated value of the random variable Z at the unsample location x_0 and ω_i are n weights determined for the observed points $Z(x_i)$.

The random function $Z(x)$ can be decomposed onto the trend component and onto the remainder component $R(x)$ (Meul, Van Meirvenne 2003)

$$Z(x) = m(x) + R(x) \tag{17}$$

Ordinary kriging assumes the stationarity of the mean and takes into account that fact that $m(x)$ is a constant but unknown value. It is also assumed that the collection of data has a constant variance. The remainder component $R(x)$ is modeled as a stationary random function with the mean equal to zero, and according to the assumption of the intrinsic stationarity its spatial dependence is defined by the semivariance as

$$\gamma_R(h) = \frac{1}{2} \mathrm{E}[(Z(x+h) - R(x))^2] \tag{18}$$

where:

$R(x)$ – is the remainder component,

$Z(x+h)$ – random variable at the separation vector h distances.

Ordinary kriging is distinguished by a high reliability of obtained estimations and it is recommended for most collections of data.

Universal kriging is a more complex method of kriging as it is a two-stage procedure. Universal kriging assumes that $m(x)$ is not stationary but it changes gently in the local neighborhood representing thus a local trend. The trend component $m(x)$ is modeled as a biased sum of

the known $f_l(x)$ and of the unknown factors a_l, $l = 0, ..., L$ (Meul, Van Meirvenne 2003)

$$m(x) = \sum_{l=0}^{L} a_l f_l(x) \tag{19}$$

In reality the semivariance of the remainders (h) given in the equation (18) is calculated before the trend $m(x)$ is modeled. As the values of the attribute $z(x)$ are the only available data, the semivariance of the remainders is calculated by choosing observation pairs that are not or are only slightly influenced by the trend.

A specific type of kriging is ordinary cokriging – a multidimensional extension of ordinary kriging. It is a kriging method that for the estimation of the unknown value of the regionalized variable at location x_0 uses the information from sample points both of the main variable $z(x_i)$ and the secondary variable $z(y_j)$.

There is of course a certain statistical correlation between the main and the secondary variable. Cokriging estimator is written as

$$Z^* = \sum_{i=1}^{n} \omega_i z_i + \sum_{i=1}^{m} \lambda_j y_j \tag{20}$$

where:

ω_i and λ_j are the weight for the main and the secondary variable respectively.

The basic assumption of ordinary cokriging is the local stationarity of the main and the secondary variable at a certain neighborhood of the point x_0 in which the interpolation is made.

5. EXAMPLES OF APPLICATIONS OF KRIGING

Nowadays kriging methods apart from their original applications in mining industry (hard coal, copper and crude oil) are also used in such fields as: environment protection, pedology, agriculture, hydrology, fishing, forestry, meteorology and economy. The main reason for using kriging methods in so many fields is that the costs connected with the examinations of the distribution of spatial phenomena at a certain location can be substantially limited. On the basis of the information obtained at a few sampling locations kriging methods let us interpolate the value of the examined phenomenon at other locations without the necessity to conduct expensive examinations.

An example of kriging application in fields connected with environment protection was the implementation of this method to estimate the distribution of the radioactive radiation in Byelorussia following the Chernobyl nuclear power station disaster. Kriging is also used in examinations of the solar radiation (Rehman, Ghori 2000). The implementation of kriging methods in determining the distribution of the solar radiation helped to reduce the number of stations that gathered data on this subject.

Kriging methods are also widely used in hydrology, e.g. (Borga, Vizzaccaro 1997) to determine the spatial distribution of precipitations at a given location. Kriging is also used in examinations aimed at determining the distribution of the height of the water surface of water reservoirs and the state of the underground waters, which is of great importance in the management of water resources.

We can also use kriging methods in agriculture e.g. to determine the humidity of the soil in a certain area (Usowicz 1999) or to determine the extent to which the crops should be irrigated (Sousa, Pereira 1999).

Kriging methods are also used in economy. The Spanish scientists J. Mira and M. J. Sanchez (2004) present in their study the concept of implementing kriging methods in the iterative procedure of detecting atypical observations in economic time series. They suggest using the geostatistical kriging method to approximate the sample distribution taking into account the detection of atypical observations in a time series.

6. CONCLUSIONS

From the presented examples of the implementation of kriging methods in various fields of knowledge one can conclude that geostatistics is a branch of statistics which is developing quite rapidly. Its popularity results from the fact that the methods of geostatistical analysis let one obtain satisfactory results while estimating the distribution of spatial phenomena. Implementation of kriging method leads to a considerable reduction in costs of the examinations connected with obtaining information concerning the behavior of a given phenomenon in a defined area. Apart from these economic advantages, kriging has also pure statistical advantages compared with other methods used in the analyses of spatial data. The most important of them is the fact that it provides an estimation error in every point of interpolation. It is also described as the best unbiased linear estimator.

Despite a dynamic development and a wide range of applications of geostatistical methods around the world, it should be said that in Poland this methodology is not used frequently enough and it should be popularized.

REFERENCES

Borga M., Vizzaccaro A. (1997), *On the interpolation of hydrologic variables: formal Equivalence of multiquadratic surface fitting and kriging*, "Jurnal of Hydrology", **195**, 160–171.

Francois-Bongarcon D. (2004), *Theory of sampling and geostatistics: An intimate link*, "Chemometrica and Intelligent Laboratory Systems", **74**, 143–148.

Krige D. G. (1951), *A statistical approach to some mine valuations and allied problems at the Witwatersrand, Masrer's thesis*, University Witatersrand, South Africa.

Matheron G. (1962–1963), *Traitē de Gēostatistique Appliquēe*, Technip, Paris.

Meul M., Van Meirvenne M. (2003), *Kriging soil texture under different types of nonstationarity*, "Geoderma", **112**, 217–233.

Mira, J., Sanchez, M. J. (2004), *Prediction of deterministic functions: An application of a Gaussian kriging model to a time series outlier problem*, "Computational Statistics & Data Analysis", **44**, 477–491.

Ploner A., Dutter R. (2000), *New directions in geostatistics*, "Journal of Statistical Planning and Inference", **91**, 499–509.

Rehman S., Ghori S. (2000), *Spatial estimation of global solar radiation using geostatistics*, "Renewable Energy", **21**, 583–605.

Sousa V., Pereira L. S. (1999), *Regional Analysis of irrigation water requirements using kriging Application to potato crop at Trás – os – Montes*, "Agricultural Water Management", **40**, 221–223.

Usowicz B. (1999), *Implementation of geostatistical analysis and the theory of fractals in the study of the dynamics of the soil humidity in cultivable areas*, "Acta Agrophysica", **22**, 229–243.

Zawadzki J. (2002), *Implementation of geostatistical methods in the analysis of spatial data*, "Wiadomości Statystyczne", **12**, 23–37.

Jan Kowalik

KRIGING – METODA STATYSTYCZNEJ INTERPOLACJI DANYCH PRZESTRZENNYCH

W analizach zjawisk przestrzennych można spotkać się z sytuacją, iż z powodu praktycznych ograniczeń niemożliwe lub bardzo kosztowne jest uzyskanie wartości (realizacji) badanego zjawiska we wszystkich położeniach. W takim przupadku, w celu określenia wartości zmiennych w tych punktach badania, zastosowanie znajduje geostatystyczna metoda estymacji (interpolacji) danych zwana krigingiem. Kriging jest podstawową metodą estymacji danych przestrzennych wykorzystywaną w geostatystyce, która interpoluje nieznane wartości zmiennej zregionalizowanej (przestrzennej) w oparciu o jej znane wartości w innych położeniach. W niniejszym opracowaniu zaprezentowano podstawowe założenia dotyczące geostatystyki oraz przedstawiono podstawy teoretyczne metody krigingu, jego rodzaje oraz przykłady aplikacji w różnych dziedzinach wiedzy.