MATHEMATICAL MODELLING AND ANALYSIS
VOLUME 9 NUMBER 1, 2004, PAGES 39–50
 ISSN 1392-6292

SPATIAL ANALYSIS AND PREDICTION OF CURONIAN LAGOON DATA WITH GSTAT

R. GARŠKA and I. KRŪMINIENĖ

Klaipėda University faculty of Natural and Mathematical sciences

H. Manto 84, Klaipėda, Lithuania

E-mail: rolandas.garska@ku.lt; ingrida@ik.ku.lt

Received October 14 2003; revised December 28 2003

Abstract. The typical goal of geostatistical analysis is to interpolate values of variable under consideration at unobserved locations using data on observed locations because it is not feasible to gather all data of the observations in the study area. The second goal is to know how they represent the study area on the basis of the sample points. Kriging is one of geostatistical methods for spatial interpolation. This method relies on the spatial correlation reflected in the available data and so represents a global view of all the data as well as the nearest neighbor influence. Before spatial prediction using kriging can be executed, the semivariogram has to be computed and modelled.

The objective of our work is to create maps of the Curonian lagoon using kriging and cokriging methods. Our spatial data consist of observations on sounding and bed sediments of different Curonian lagoon locations. For computation and simulation of semivariograms, as well as for application kriging and cokriging methods and visualization of results on maps *Gstat* and *PCRaster* are used.

Key words: variogram, semivariogram, cross semivariogram, kriging, cokriging

1. Introduction

This paper discusses the use of traditional kriging techniques when when we map variables from data that are collected in Curonian lagoon. For most applications kriging is usually associated with exact interpolation, that is, the kriging predictions change smoothly in space until they get to a location where data have been collected, at this point there is a "jump" in the prediction to the exact value that was measured. This also leads to discontinuity in the standard prediction errors, that "jump" to zero at the measured locations. In the research presented in the paper we have studied the Curonian lagoon and have created maps in all area by using data of a sediments in 213 locations and data of depth in 263 locations. In the previous publication [2] the precision of the results obtained by two methods, i.e. kriging and cokriging, were compared by using cross-validation method. The results have showed that precision

of predicted values is better when cokriging method is used. In addition the present paper presents maps of the predicted values in the whole Curonian lagoon, where the prediction is based on measurement data that are mentioned above.

Geostatistics is the name associated with a class of specialized statistical techniques used to analyze and estimate values of variables which are distributed – and physically correlated – in space and /or time. The analysis of such a correlation is usually called a "structural analysis" or "variogram modeling". From a structural analysis, predictions of the value of the variable can be made at unsampled locations using "kriging" or "stochastic simulation". This approach is most useful when the processes responsible for generating the measured variable are unknown or too poorly constrained to permit construction of a quantitative process model to make spatial interpolations or predictions. The overall sequence of steps in a typical geostatistical study include: exploratory data analysis (to explain the spatial character of the variable), structural analysis (to determine the spatial correlation or continuity of the data) and estimates (kriging or simulations to predict values at unsampled locations).

Kriging prediction consists of three steps:

- Estimation of the semivariogram or covariance;
- Choice of a model among the family of valid semivariograms or covariances;
- Estimation of the semivariogram or covariance by fitting the valid model to the empirical semivariogram or covariance and use in one of the forms of kriging (e.g. ordinary kriging, simple kriging, universal kriging, etc.).

In many environmental researches the data of more than one observation (measurements of more than one variable) are often obtained. If those variables are correlated with one another and the cross covariance functions are known or can be estimated then cokriging method can be used.

More about geostatistical analysis can be found in the book of Cressie "Statistics for spatial data" [1]. Krivoruchko has applied a kriging method to radio cesium soil contamination data, collected in Belarus after the Chernobyl accident (see, e.g. the web *http://www.esri.com/software/arcgis/arcgisxtensions /geostatistical/research_papers.html*). Lophaven has computed the spatial distribution of Dissolved Inorganic Nitrogen (DIN) and Dissolved Inorganic Phosphorus (DIP) by three different variants of kriging, i.e. ordinary kriging, universal kriging and cokriging [3]. In the next section the main terms, processes and formulas which are used in geostatistical analysis are described.

The results of the study are presented in Section 3 and conclusions are given in Section 4.

2. Spatial Data Analysis

General Spatial Model (see [1]) is described in geostatistics as $\{Z(s) : s \in D\}$, where

- $s = (x, y)$ denotes the coordinates of the sample site. Here (x, y) may be Euclidean coordinates (e.g., UTM coordinates), or latitude and longitude. More generally, we may have $s = (x, y, z)$.

- $Z(s)$ denotes the variable of interest at the location s. Note that this is written as a function of the location s.
- D denotes the set of spatial locations at which data may be obtained.

For geostatistical data, the set of all multidimensional distributions of k-tuples

$$\big(Z(s_1), Z(s_2), \ldots, Z(s_k)\big)$$

for all values of k all configurations of the points $s_1, s_2, \ldots, s_k$ constitutes a stochastic process $\{Z(s) : s{\in}D\}$. The stochastic variable $Z(s)$ has mean value

$$E[Z(s)] = \mu(s), \quad s \in D.$$

We also assume that the variable of $Z(s)$ exists for all $s \in D$.

The process Z is said to be *Gaussian* if, for any $k \geq 1$ and locations $s_1, s_2, \ldots$, s_k, the vector $\big(Z(s_1), Z(s_2), \ldots, Z(s_k)\big)$ has a multivariate normal distribution.

The process Z is said to be *strictly stationarity* if the joint distribution of $\big(Z(s_1), Z(s_2), \ldots, Z(s_k)\big)$ is the same as that of $\big(Z(s_1{+}h), Z(s_2{+}h), \ldots, Z(s_k{+}h)\big)$ for any k spatial points $s_1, s_2, \ldots, s_k$ and any $h \in R^d$, provided only that all of $s_1, s_2, \ldots, s_k, s_1 + h, s_2 + h, \ldots, s_k + h$ lie within the domain D.

The process Z is said to be $second\text{-}order\ stationarity$ (also called $weakly\ stationarity$) if μ(s) = μ (i.e., the mean value is the same for all s) and

$$\mathrm{Cov}\big(Z(s_1), Z(s_2)\big) = C(s_1 - s_2), \ \text{ for all } s_1{\in}D,\ s_2{\in}D,$$

where $C(s)$ is the covariance function of an observation at location $s{\in}D$.

It can immediately be seen that with all variances assumed finite, a strictly stationary process is also second-order stationary. The converse is in general false, but a $Gaussian$ process which is second-order stationary is also strictly stationary (see [4], 35 p.). Intrinsic stationarity is a weaker property than second-order stationarity. The variogram of intrinsic random function is written as

$$2\gamma(h) = Var[(Z(s_1) - Z(s_2))]\,.$$

The function $2\gamma(\cdot)$ is called the *variogram (variance)* and $\gamma(\cdot)$ the *semivariogram (semivariance)*. If the semivariogram (covariance) depends only on distance between locations the process is called *isotropic*. The variogram is the variance of the difference between $Z(s_1)$ and $Z(s_2)$. If two units are close together, their difference will typically be small, as would the variance of the difference. As units get apart, their differences get larger and usually the variance of the difference gets larger.

If second-order stationarity is assumed, the relationship between the function semivariogram and the covariance is given as

$$\gamma(h) = C(0) - C(h)\,. \tag{2.1}$$

Note that $C(0) = \sigma^2$, the variance of the random function when $h = 0$. Equation (2.1) shows that if the covariance is known, the variogram is also known. In practice the variogram (or/and semivariogram) is used more often then the covariance, because the variogram, unlike the covariance, does not require the knowledge of the mean value. Also semivariogram is less sensitive to any unidentified trend.

2.1. Estimation of the semivariogram and cross semivariogram

Determination of spatial variability is often based on a semivariogram. The sample estimator of the semivariogram, which is based on the method-of-moments, is given by

$$\hat{\gamma}(h) = \frac{1}{2N(|h|)} \sum_{(s_k,s_l)\in N(|h|)} [Z(s_k) - Z(s_l)]^2,$$

where $N(|h|)$ denotes all pairs (s_k, s_l) for which $|s_k - s_l| = |h|$ ([4], 38 p.). The spatial variability between two correlated random variables is described by the cross semivariogram. When the intrinsic hypothesis is assumed, it is defined as

$$\gamma_{12}(h) = \gamma_{21}(h) = \frac{1}{2} E\big[\big(Z_1(s_k) - Z_1(s_l)\big)\big(Z_2(s_k) - Z_2(s_l)\big)\big],$$

where $Z_1(s)$ and $Z_2(s)$ denote two different variables. An estimator of the cross semivariogram is defined as

$$\hat{\gamma}_{12}(h) = \frac{1}{2N(|h|)} \sum_{(s_k,s_l)\in N(|h|)} \big(Z_1(s_k) - Z_1(s_l)\big)\big(Z_2(s_k) - Z_2(s_l)\big),$$

where $N(|h|)$ is the number of pairs of observations within distance $|h|$. Usually $\hat{\gamma}_{ij}(h)$ is called the experimental or sample cross semivariogram (see [3]).

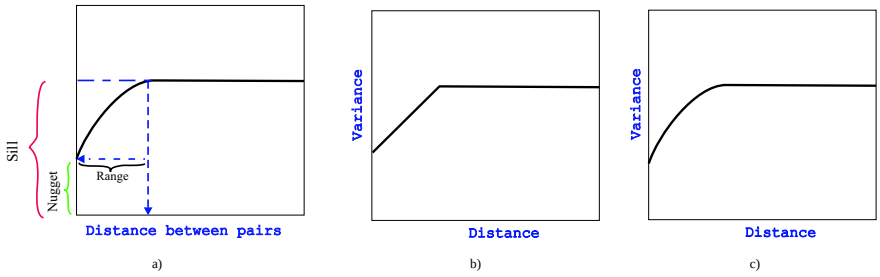


Figure 1. Variograms: a) idealized form of variogram function; b) linear variogram; c) spherical variogram.

2.2. Modelling the semivariogram and the cross semivariogram

Modelling of semivariogram and cross semivariogram is done in the same way. The estimated semivariogram (cross semivariogram) is fitted with a model, and the best models are used in the kriging estimation. Several methods have been proposed for fitting semivariogram models. One relatively simple method that appears to perform well is the *Weighted Least Squares*. Figure 1a show representation of general variogram.

The *range* is the distance beyond which observations are uncorrelated or at least approximately uncorrelated. On the semivariogram, the range is the point on the x – axis where the curve reaches a plateau. *Sill* is the value of semivariogram where observations are uncorrelated or nearly uncorrelated. On the semivariogram shown, the sill is the height of the curve at the plateau.

The *nugget variance* or *nugget effect* is the resulting discontinuity of the semivariogram at the origin, the difference between zero and the semivariogram at a lag distance is some greater than zero. The nugget effect is caused by measurement errors and micro-variability. A variogram model can consist of pure nugget effect.

Isotropic processes are convenient to deal with because there are a number of widely used parametric forms for $\gamma(\cdot)$. An often used semivariogram model is the linear and the spherical model with nugget effect. A reason for this is an easy interpretation of the parameters.

A linear semivariogram model (Fig.1b) in the isotropic case is defined as:

$$\gamma(h) = \begin{cases} 0, & \text{if} \quad |h| = 0, \\ C_0 + C_1 h, & \text{if} \quad 0 < |h| < R, \\ C_0 + C_1 R, & \text{if} \quad |h| > R\,. \end{cases} \tag{2.2}$$

Spherical semivariogram model (Fig.1c) is defined as:

$$\gamma(h) = \begin{cases} 0, & \text{if} \quad |h| = 0, \\ C_0 + C_1 \left[\dfrac{3}{2}\dfrac{h}{R} - \dfrac{1}{2}\left(\dfrac{h}{R}\right)^3 \right], & \text{if} \quad 0 < |h| =< R, \\ C_0 + C_1, & \text{if} \quad |h| > R\,, \end{cases} \tag{2.3}$$

where C_0 is the nugget effect, R is the range and $C_0 + C_1$ is the sill [3].

When two or more variables are correlated, the nature of spatial cross correlation between the primary variable and several secondary variables can provide valuable information for estimation and simulation of the primary variable. Cross semivariogram modelling is always done for the purpose of developing a model to be used in estimation or simulation. The models that are to be used in estimation and simulation must obey a number of stringent constraints to ensure that the matrix solutions to the kriging equations exist and are numerically stable.

Traditionally fitting of the cross semivariogram is done by eye, because it has been shown that predictions computed by kriging are reasonably insensitive to the specification of the cross semivariogram model. The best semivariogram (cross semivariogram) model can be found using the least squared criterion [3].

2.3. Kriging Concept

Kriging is a generic name adopted by the geostatisticians for a family of generalized least-squares regression algorithms that allow one to account the spatial dependence between observations, as revealed by the semivariogram, into spatial prediction. It is a procedure for spatial prediction at an unobserved location, using data at observed locations, optimized with reference to a specific error criterion.

Kriging is known to be a Best Linear Unbiased Predictor (B.L.U.P.), because it minimizes the variance error between the true value and the predictor. Linear predictor of the value $Z_1(s_0)$ of the data at the unsampled site s_0 from the data $Z(s) = Z(s_1(s)), Z(s_2(s))$ at the sampled sites s_1, s_2 is:

$$\hat{Z}_1(s_0) = \sum_{k=1}^{n} w_k Z_1(s_k),$$

where w_k is the weight for the k-th variable of observation at location s_k and n is the number of observations. The weights w_k are chosen to minimize the mean squared error

$$\mathrm{MSE} = E\big[\hat{Z}_1(s_0) - Z_1(s_0)\big]^2.$$

$\hat{Z}(s_0)$ is unbiased for $Z(s_0)$ if and only if

$$\sum_{k=1}^{n} w_k = 1\,.$$

Ordinary kriging gives the optimal predictions under the assumption that the mean value is constant (but unknown) across the whole area under study. The ordinary kriging variance for Z_1 is given by

$$\sigma^2_{ok} = \sum_{k} w_k \gamma(s_k - s_0) + m, \tag{2.4}$$

where m is a Lagrange multiplier

$$m = \big(\mathbf{1}' \sum_{\mathbf{z}}^{-\mathbf{1}} \mathbf{c_z} - \mathbf{1}\big) / \big(\mathbf{1}' \sum_{\mathbf{z}}^{-\mathbf{1}} \mathbf{1}\big)\,,$$

$\sum_z$ is the covariance matrix among the data, c_z is $Cov(z, Z(s_0))$ (see [1], p 143).

Cokriging is prediction of a primary variable using additional information from a secondary variable. This method is used in data sets containing two or more regionalized variables which are correlated with one another. Suppose that $q = 2$. The prediction of $\hat{Z}_1$ is done not only on the basis of Z_1, but also on measurements of Z_2. Cokriging involves the prediction of $Z_1(s_0)$ at an unsampled site s_0 from the data $Z(s_1), Z(s_2), \ldots, Z(s_n)$, $Z(s)^T = \big(Z_1(s), Z_2(s)\big)$ at the sampled sites $s_1, s_2, \ldots, s_n$. The linear prediction of cokriging is defined as:

$$\hat{Z}_1(s_0) = \sum_{k=1}^{n} v_1^k Z_1(s_k) + \sum_{k=1}^{n} v_2^k Z_2(s_k).$$

To obtain an unbiased estimate the following constraints are needed:

$$\sum_{k=1}^{n} v_1^k = 1, \qquad \sum_{k=1}^{m} v_2^k = 0\,.$$

Similarly as (2.4) the variance of cokriging prediction can be written as

$$\sigma^2_{cok} = \sum_{k=1}^{n} v_1^k \gamma_{k1}(s_k - s_0) + \sum_{k=1}^{n} v_2^k \gamma_{k2}(s_k - s_0) + m_1.$$

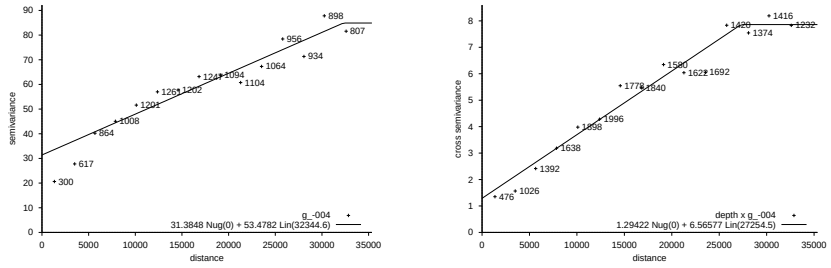
3. Results

The above procedure of variogram estimation, variogram model fitting, kriging and cokriging was applied to the Curonian lagoon data. The Curonian lagoon (also known as Kuršių marios, Kurshskij zaliv, Kurische Haff) is a large (length 95 km, width up to 48 km), shallow (mean depth of 3.8 m, the maximum depth - 5.8 m) coastal water body in the south-eastern part of the Baltic Sea. The outlet of the lagoon to the Baltic Sea, Klaipėda Strait, is artificially deepened down to 12 m.

The data have been collected in 1990 year by S. Gulbinskas. It consists of bed sediments and soundings of the Curonian Lagoon. Sediments where measured in 213 locations, depth was measured in 263 locations. Their x coordinate values are between 278199 and 333376 and y coordinate values are between 6088178 and 6172784. Sediments have been divided into 7 groups (granulometric fractions) depending on median diameter (Md) in mm: (1) more than 0.5, (2) 0.5-0.25, (3) 0.25-0.125, (4) 0.125-0.063, (5) 0.063-0.01, (6) 0.01-0.004, (7) less than 0.004.

In order to apply the above statistical methods for data analysis, and mapping we have chosen free available software *Gstat* and *PCRaster*. *Gstat* is a program for the modelling, prediction and simulation of geostatistical data in one, two or three dimensions. In *Gstat* geostatistical modelling comprises calculation of sample variograms and cross variograms (or covariograms) and fitting models to them. In this paper *Gstat* has been used for modelling semivariance of all above fractions and for simulation cross variance between depth and sediment fractions. *PCRaster* has been used for showing kriging and cokriging prediction maps.

In *Gstat* a simple variogram model is denoted $cMod(a)$ with c the vertical (variance) scaling factor, *Mod* the model type, and *a* the range (horizontal, distance scaling factor) of this simple model. Linear and spherical models defined in (2.2) and (2.3) equations, in *Gstat* are denoted by $Lin(\cdot)$ and $Sph(\cdot)$, respectively. The nugget effect is indicated by $Nug(\cdot)$.



a) b)

Figure 2. a) semivariogram of fraction (7) where Md of sediments is less than 0.004; b) cross semivariogram between fraction (7) and depth of the Curonian lagoon.

To describe results of our research we took measurements of depth and fraction (7). Figure 2a presents semivariogram of fraction (7) where Mod of sediments is less than 0.004. Equation of this semivariogram is given by

$$1.29422Nug(0) + 6.56577Lin(27254.5),$$

where *Lin* represents model type, $Nug(0) = 1.29422$, when $\frac{1}{2}\gamma(h)$ = 0, the sill is 6.56577 and the range is 27254.5. The parameters and models of all semivariogram fractions are given in Table 1.

Table 1. Types of semivariogram models of all fractions and the parameters: range, sill and nugget effect.

Fraction	Model	Range	Sill	Nugget effect
more than 0.5	Linear	14402.5	5.05205	15.6034
0.5-0.25	Linear	22615.3	167.631	290.326
0.25-0.125	Linear	11036.5	167.631	504.511
0.125-0.063	Linear	9773.75	50.812	264.483
0.063-0.01	Linear	22051.9	297.963	316.533
0.01-0.004	Linear	31426.3	52.2167	15.1969
less than 0.004	Linear	32344.6	53.4782	31.3848

Figure 2b presents cross semivariogram between fraction (7) and depth of the Curonian lagoon. Equation of this cross semivariogram is given by

$$31.3846Nug(0) + 53.4782Lin(32344.6),$$

where *Lin* represents model type, the nugget effect equals 31.3846, when $\frac{1}{2}\gamma(h)$ = 0, the sill is 53.4782 and the range is 32344.6. The parameters and models of all cross semivariogram fractions are given in Table 2.

Table 2. Types of cross semivariogram models between depths and fractions and the parameters: range, sill and nugget effect.

Fraction	Model	Range	Sill	Nugget effect
more than 0.5	Linear	14360.4	-0.888949	-0.718705
0.5-0.25	Linear	20658.5	-9.04819	-7.57507
0.25-0.125	Linear	30420.6	-20.0815	-3.64506
0.125-0.063	Linear			4.58617
0.063-0.01	Linear	25260.5	16.3492	5.60972
0.01-0.004	Linear	30887.2	6.71625	0.738019
less than 0.004	Linear	27254.5	6.56577	1.29422

Figure 3 presents the linear and spherical semivariograms of the depth. In this case the spherical semivariogram is preferred. The parameters and models of depth are given in Table 3.

Kriging is most sensitive to the behavior of the variogram near zero. In particular, it is sensitive to the presence / absence of the nugget effect. Maps of variations and predictions created using kriging method are shown in Figure 4 (here (a) prediction

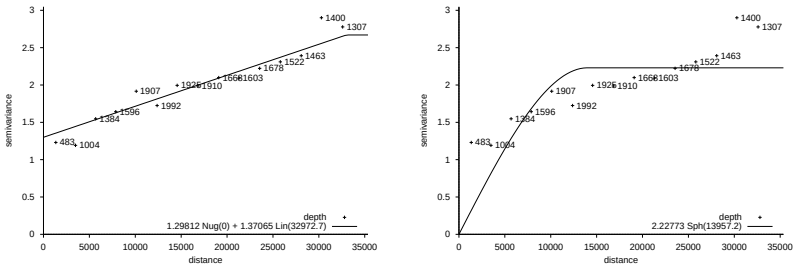


Figure 3. a) linear semivariogram of depth; b) spherical semivariogram of depth.

Table 3. Types of semivariogram models of depth and the parameters: range, sill and nugget effect.

Model	Range	Sill	Nugget effect
Linear	32972.7	1.37065	1.29812
Spherical	13957.2	2.22773	

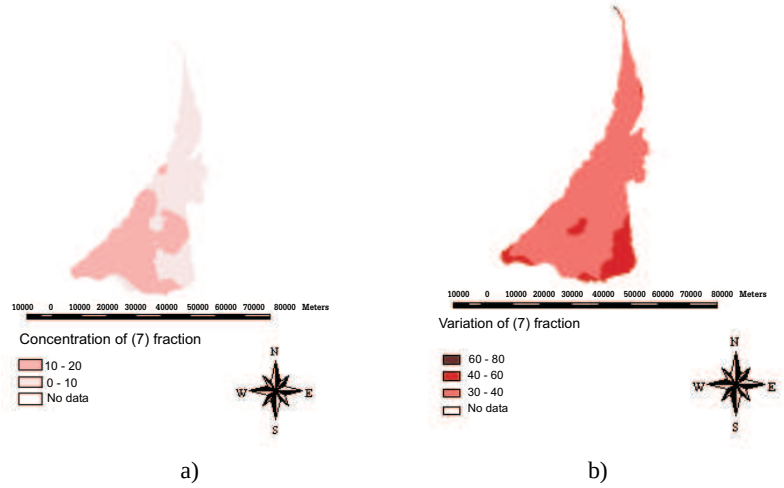
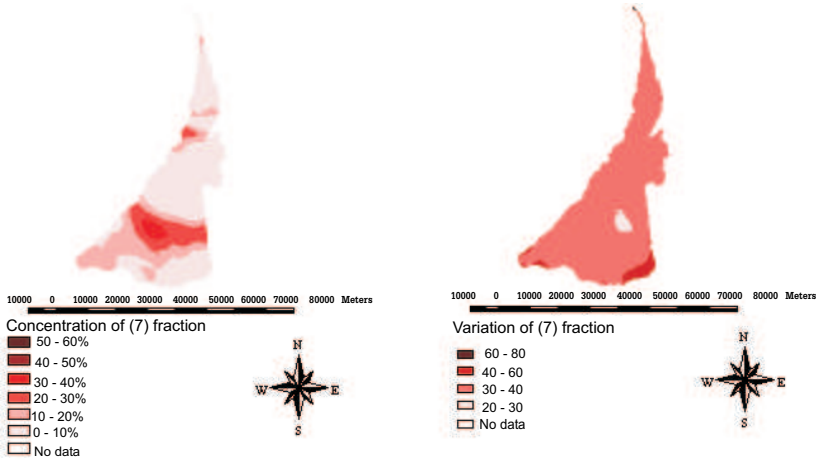


Figure 4. Kriging method: a) prediction map of fraction (7); b) variation map of fraction (7).

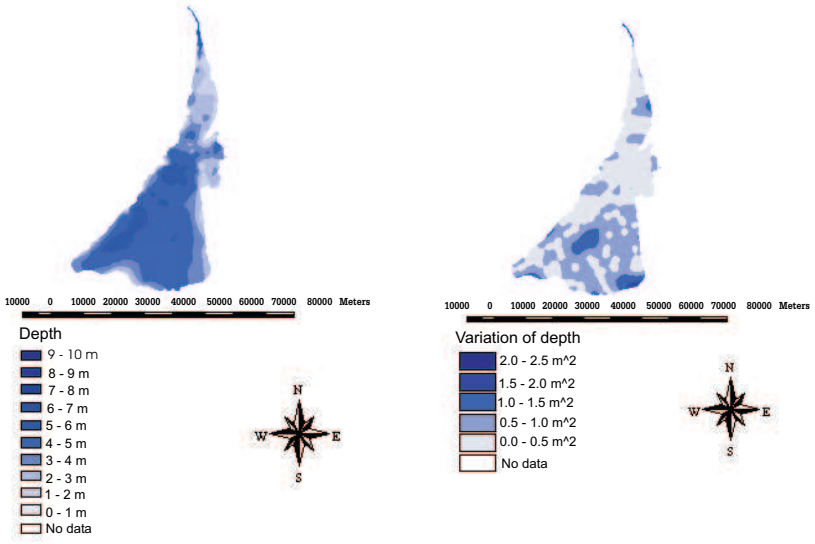
map, (b) variation map). Maps obtained using cokriging method are shown in Figure 5. Figure 6a presents prediction map of depth, while Figure 6b presents variation map of depth. These maps have been created using kriging method.

In order to check which one of the kriging and cokriging maps correspond best to true data we must first choose one point on the map, then compare these variation maps, and finally determine which map has smaller variation for the selected point. The method containing a smaller variation has created a better prediction map.



a) b)

Figure 5. Cokriging method: a) prediction map of fraction (7); b) variation map of fraction (7).



a) b)

Figure 6. Kriging method: a) prediction map of depth; b) variation map of depth.

4. Conclusions

Statistical methods for data on bed fractions percentage and soundings have been described, applied and mapped. The methods are general, but in this paper they have been applied only to measurements of the Curonian lagoon.

Semivariogram and cross semivariogram models have been made using percentage of fractions and soundings of the Curonian lagoon. Variance distribution and distribution of bed fractions percentage have been mapped using kriging and cokriging methods. Variance distribution and distribution of soundings have been mapped using kriging method.

Results demonstrate that:

- Nugget and linear models best describe semivariance and cross semivariance of percentage of ground fractions.
- Spherical model best describes semivariance and cross semivariance of soundings.
- Prediction variations of percentage of bed fractions made by kriging and cokriging methods are very similar.

Also cross semivariance show interdependence of parameters of models and depths. Prediction results of bed fractions percentage made by kriging method are very close to the mean value, while cokriging method shows that the variation of data are less close to the mean value.

References

[1] N. Cressie. *Statistics for Spatial Data*. John Wiley, New York, 1993.

[2] I. Kr¯uminien˙e, K. Dučinskas and R. Garška. Applying of kriging and cokriging methods for prediction of Curonian lagoon data. *Liet. matem. rink.*, **43**(spec. nr.), 504 – 508, 2003.

[3] Soren Nymand Lophaven. *Reconstruction of data from the aquatic environment.* LYNGBY, 2001.

[4] Richard L. Smith. *Environmental Statistics*. University of North Carolina Chapel Nill, NC 27599-3260, USA, 2001.

Apie Kuršių marių duomenų erdvinę analizę ir prognozavimą Gstat programos pagalba

I. Krūminienė, R. Garška

Šio darbo pagrindinis tikslas - Gstat bei PCRaster programų pagalba sukurti prognozuojamų duomenų ir jų dispersijų žemėlapius. Žemėlapiams sudaryti pritaikyti krigingo ir kokrigingo metodai. Krigingas yra vienas iš geostatistikos metodų, kuris atsižvelgdamas į erdvinį dviejų kintamųjų ryšį ir kaimyninių taškų reikšmes atlieka erdvinę interpoliaciją. Tuo tarpu kokrigingas atlieka pirminio kintamojo duomenų prognozę naudojant antrinių kintamųjų duomenis. Pagrindinis geostatistinės analizės tikslas yra interpoliuoti duomenis nežinomuose srities taškuose, nes dažniausiai atliekant geostatistinius tyrimus naudojami daliniai stebėjimai, kurie apima tik visumos dalį; arba nėra žinoma, ar imties duomenys pakankamai gerai atspindi visą studijuojamą sritį. Rezultatų analizė parodė, kad tikslesnė prognozė gaunama taikant kokrigingo metodą