

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Ingrida BORISENKO

**MODELLING OF SPATIAL DATA USING
SEMIVARIOGRAMS OF STATIONARY
SPATIAL PROCESSES**

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Vilnius  2009

Doctoral dissertation was prepared at Klaipėda University and Vilnius Gediminas Technical University in 2001–2009.

The dissertation is defended as an external work.

Scientific Consultant

Prof Dr Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics – 01P).

The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

Members:

Prof Dr Habil Mindaugas BLOZNELIS (Vilnius University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Vitalijus DENISOVAS (Klaipėda University, Physical Sciences, Informatics – 09P),

Assoc Prof Dr Mečislavas MEILŪNAS (Vilnius Gediminas Technical University, Physical Sciences, Mathematic – 01P),

Prof Dr Habil Henrikas PRAGARAUSKAS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

Opponents:

Prof Dr Habil Vilijandas BAGDONAVIČIUS (Vilnius University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Marijus RADAVIČIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics in the Senate Hall of Vilnius Gediminas Technical University at 1 p. m. on 6 November 2009.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail: doktor@adm.vgtu.lt

The summary of the doctoral dissertation was distributed on 5 October 2009.

A copy of the doctoral dissertation is available for review at the Libraries of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius, Lithuania).

© Ingrida Borisenko, 2009

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Ingrida BORISENKO

**ERDVINIŲ DUOMENŲ MODELIAVIMAS
NAUDOJANT STACIONARIŲ ERDVINIŲ
PROCESŲ SEMIVARIOGRAMAS**

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

Vilnius  2009

Disertacija rengta 2001–2009 metais Klaipėdos universitete ir Vilniaus Gedimino technikos universitete.

Disertacija ginama eksternu.

Mokslinis konsultantas

prof. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Nariai:

prof. habil. dr. Mindaugas BLOZNELIS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

doc. dr. Vitalijus DENISOVAS (Klaipėdos universitetas, fiziniai mokslai, informatika – 09P),

doc. dr. Mečislavas MEILŪNAS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P),

prof. habil. dr. Henrikas PRAGARAUSKAS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Oponentai:

prof. habil. dr. Vilijandas BAGDONAVIČIUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

prof. dr. Marijus RADAVIČIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2009 m. lapkričio 6 d. 13 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.

Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;

el. paštas doktor@adm.vgtu.lt

Disertacijos santrauka išsiuntinėta 2009 m. spalio 5 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 1662-M mokslo literatūros knyga.

© Ingrida Borisenko, 2009

Introduction

Topicality of the problem

Spatial statistics is one of the youngest trends in the science of statistics. First, it has been applied in mining, during the fifth decade of the last century. In fifty years after this trend of science had been discovered, the circle of the scientists involved in it has grown drastically as well as areas of application. Also, a wide range of theoretical and practical material has been issued. Nowadays, spatial statistics methods are used in: ecology, quantity geology, image processing and analysis, epidemiology, studying global climate change and even cosmology. However, in Lithuania, the methodology of spatial data analysis has been studied only from the beginning of this Millennium. Since only few scientists (Dumbrasukas, A.; Kumetaitis, A.; Kumetaitienė, A. and others) are involved, it is very important to expand this area and develop the existing methods. Also it is essential to study the spatial data modelling methods thoroughly and provide general spatial data modelling methodology.

In order to apply the methods of spatial statistics, it is necessary to know the location of data in space, which is usually expressed in geographic coordinates. Thus, one of the main distinctions of spatial statistics which makes it different from the classical is the ability to model both spatial trend and spatial autocorrelation.

One of the main objectives of spatial statistics is creating a mathematical model of spatial data, which can be used for interpolation (extrapolation) or for other purposes. To accomplish data estimation the kriging method can be used. Under certain conditions it minimizes mean square prediction error, that is to say, provides the best linear unbiased prediction. Depending on spatial data nature, the mean model of spatial process and other data characteristics, several types of kriging can distinguish.

The spatial autocorrelation among observations is described by covariance function or semivariogram. The last is often used in practical researches, because it covers the wider class of spatial processes. To describe the variation of spatial data by semivariogram, firstly, the spatial structure of observations is modelled using the graph of the empirical semivariogram. The empirical semivariogram estimator mostly is calculated by using Matheron method of moments (MoM). Then one of semivariogram parametric model is fitted. Parametric model is described by several semivariogram parameters θ . Main of them are nugget, sill and range.

The best linear unbiased predictor can not be computed unless θ is known, which is typically not. To evaluate semivariogram parameters, the least squares methods (ordinary, weighted and generalised least squares methods) are often used. As stated Lahiri *et al.* in "On asymptotic distribution and asymptotic

efficiency of least squares estimators of spatial variogram parameters“ (2002), under the general conditions the generalised least squares method (GLS) asymptotically effective by comparing with other least squares methods.

The least squares methods can be also used for the estimation of parameters of the mean model. Since in literature GLS method is presented through covariances functions, it is important to take the expression of this method through semivariogram.

Should also be noted that the estimators of semivariogram parameters by analytical method is not available. Therefore, the asymptotic analysis of estimators is live issue.

Another frequent task in the spatial statistic is the determination of the optimal sampling plan. D and A criterion of sampling design is widely used. Again, providing of expression of D and A criterion in semivariogram form is very important.

Research object

The research object of the dissertation is the methodology of modelling, prediction and results analysis of spatial data generated by spatial processes with stationary errors.

Aim of the work

The main objective of thesis is to propose the methodology for modelling of spatial data through semivariograms and prediction by kriging, when the spatial data are generated by spatial processes with stationary errors.

Tasks of the work

1. To review the additive models of spatial processes with stationary errors and describe the fundamental characteristics of spatial processes.
2. To propose the methodology of modelling and prediction of spatial data by using various parametric models of mean and stationary semivariogram.
3. To adapt the described and proposed methodology to modelling and predictions of depth and sediments data of the Curonian Lagoon.
4. To compare the spatial sampling design using various criterions.
5. To perform the analysis of statistical properties of MoM estimator.

Methodology of research

In analytical part of work when the mathematical models of spatial data and the methodology of modelling and prediction of spatial data are described, the analysis of literature and comparative analysis are used. To summarize the steps of modelling and prediction of spatial data, the techniques of conceptual modelling and summary of the concepts are applied. For the practical

realization of modelling and prediction methodology proposed in this work, experimental analysis is used.

Scientific novelty

In this work the general methodology of modelling and prediction of spatial data through semivariograms, when the spatial data are generated by spatial processes with stationary errors is proposed. The analysis of above mentioned methodology discovers the possibility of improving of some of the spatial statistical methods.

In this work the following findings are provided:

1. The semivariogram estimator MoM is provided through spacial design (Laplace) matrix.
2. The expression of generalized least squares estimator of β is presented through semivariogram.
3. The expressions of D and A criterion functions for parameters and estimators of β are given through semivariogram.
4. For increasing domain asymptotic the asymptotical normality of MoM estimator, when errors of data model are stationary and Gaussian, is proofed.
5. The expression of asymptotic variance-covariance matrix in the case of exponential isotropic semivariogram is derived.
6. The proposed methodology of modelling and prediction of spatial data generated by spatial processes with the stationary errors allows:
 - a) to describe the variation of spatial data through semivariograms by including spatial trend and anisotropy;
 - b) to predict the data by ordinary, universal, median polish kriging and cokriging;
 - c) to apply methods (SSE – sum of square of errors, cross-validation method, Akaike criterion), which verificate the selected parametric semivariogram models and kriging methods.

Practical value

The proposed methodology of modelling and prediction can be used in other scientific fields (for example in econometrics, demographics and others) to analyse a various types of spatial data. The theoretical results can be used in analysis of optimal sampling design. The practical results can be an example how to realise methodology of modelling and prediction of spatial data by functions in the various packages of open source system R.

Defended propositions

1. The proposed methodology of spatial data modelling through semivariograms and prediction by universal, ordinary, median polish kriging and cokriging.
2. The given expression of variance-covariance matrix of semivariogram estimator (MoM) through theoretical semivariogram in the case of stationary Gaussian data model.
3. The expression of generalised least squares (GLS) estimator through semivariogram for β is put by replacing variance-covariance matrix with matrix of semivariogram accordingly.
4. The given expressions of D and A criterion functions through semivariogram for GLS estimator of β .
5. The asymptotical normality of MoM semivariogram estimator, when errors of data model are stationary and Gaussian.
6. The derived expression of asymptotic variance-covariance matrix MoM semivariogram estimator for spatial sampling design of increasing domain asymptotic named LIDA, which described in Lahiri et al. (2002), in case of exponential isotropic semivariogram.

The scope of the scientific work

The scientific work consists of the general characteristic of the dissertation, 4 chapters, conclusions, list of literature, list of publications and addenda. The total scope of the dissertation – 127 pages, 60 figures, 20 tables and 4 addenda.

1. Mathematical models of spatial data

In the first chapter of the work the additive models of spatial processes and their generating spatial data are introduced. Cressie and others scientists define three main groups of spatial data: geostatistical, lattice and point patterns. The one of fundamental attributes of spatial data is spatial autocorrelation: observations closer together tend to be more alike than observation further apart. The standard methodology in spatial statistics is essentially based on the assumptions of stationary (strictly stationarity, weak stationarity, intrinsically stationarity) and isotropy (see section 1.2).

In this work the additive spatial model $\{ \mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D, \text{ where } D \subset \mathfrak{R}^d \}$ of spatial observation at location $\mathbf{s}$ is analyzed:

$$\mathbf{Z}(\mathbf{s}) = \beta^T f(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad (1)$$

where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_Q)^T$ – vector of parameters, $f(\mathbf{s}) = (1, f_1(\mathbf{s}), \dots, f_Q(\mathbf{s}))^T$ – vector of explanatory variables and $\mathcal{E}(\mathbf{s})$ is zero-mean error finite variance stationary process.

The spatial dependence structure of $\mathcal{E}(\mathbf{s})$ could be expressed by covariance function $C(\mathbf{h})$ or by semivariogram:

$$\gamma(\mathbf{h}) = \frac{\text{var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))}{2}. \quad (2)$$

It is obvious that $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$, where $C(\mathbf{0})$ is variance σ^2 . The semivariogram used more often as it covers the wider class of spatial processes. Since the class of intrinsic stationary process contains the class of second-order stationary process, the semivariogram of a second-order stationary process can be constructed from the covariance function. If the process is intrinsic but not second order stationarity, the covariance function can not be used to describe spatial correlation between observations. In this work the assumptions of weak stationarity or intrinsic stationarity is mostly used.

The semivariogram and its parameters, some semivariogram estimators, the most common parametric models of semivariogram and advices on their practical usage are described in section 1.4. Noted, that in applications the MoM estimator is mostly used. The semivariogram estimator MoM for stationary spatial data model is calculated by such formula:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2, \quad (3)$$

where $|N(\mathbf{h})|$ are number of pairs in the population $N(\mathbf{h})$, $\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}$. The semivariogram estimator MoM may be written in such way:

$$\hat{\gamma}(\mathbf{h}) = \frac{\mathbf{Z}^T \mathbf{L}(\mathbf{h}) \mathbf{Z}}{2|N(\mathbf{h})|}, \quad (4)$$

where $\mathbf{L}(\mathbf{h})$ in graph theory is known as Laplace matrix. In the spatial statistics this matrix called the spatial design matrix.

Depending on properties of spatial process, the empirical semivariograms are divided into isotropic and anisotropic. The spatial variability between two correlated physical magnitudes is described by the cross semivariogram.

2. Methods of spatial data statistical modelling and estimation

Prediction or estimation of spatial data at unobserved location is particular important for geostatistical data. Complete sampling of a surface is impossible and the user may be interested in predicting the amount $\mathbf{Z}$ at arbitrary location $\mathbf{s}$.

The understanding of geostatistical assumptions and methods allows choosing the best way of estimation. It looks too confusing for the representatives of other scientific areas the sequence of stages of spatial data statistical modelling and estimation.

Thus, in the second chapter the stages of spatial data modelling by semivariograms and estimation by kriging are described and the execution plan is proposed. The chapter is divided into six parts. In the section 2.1 the classification of estimation methods of spatial data and the common stages of geostatistical modelling and estimation are described. In the section 2.2 the first stage of preliminary spatial data analysis and tools for it are reviewed.

In the section 2.3 the modelling by semivariogram, identification of the trend and anisotropy and types of anisotropy are analyzed. The elimination of anisotropies is considered in the section 2.3.1.

The empirical semivariogram displays half the average difference in Z values of two data points as a function of the separation distance between the points. In order to calculate the empirical semivariogram, it is needed to specify a lag distance and a maximum number of lags. The empirical semivariogram provides the estimates only in finite set of lags. In order to obtain estimates at arbitrary lag the empirical semivariogram must be smoothed – replaced by some parametric form. The fitting of parametric models to the empirical semivariogram can be done by several methods: least squares estimation, maximum likelihood and restricted maximum likelihood. The least squares estimation methods are reviewed in details in the section 2.3.3.

The section 2.4 is designated for spatial data prediction method – kriging and their types. The emphasis is put on: simple, ordinary, block (ordinary), universal, median polishing krings and cokriging.

Further, in the section 2.5 the validation methods, such as cross-validation methods and Akaike criteria, of the selected semivariogram model and/or kriging, are analyzed.

In the next section 2.6 the influence of semivariogram parameters and models to estimation (prediction), the advantages and disadvantages of kriging by comparing with other prediction methods are reviewed.

The analysis of all stages of geostatistical modelling and prediction showed that the statistical analysis of spatial data allows to verify the data attributes, logicity of topological information; to analyze data by classical methods of statistics (to select the data distributions, to detect the outliers, to analyse trend and etc.) (Fig. 1).

To detect the anisotropy, the semivariogram map and directional semivariograms can be created.

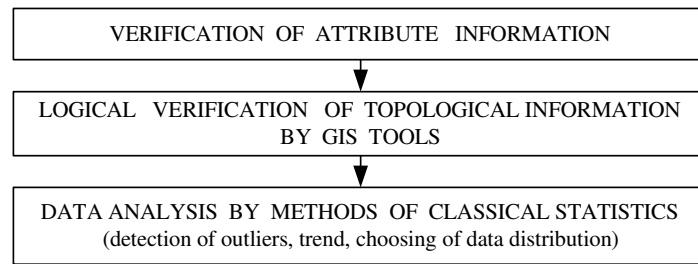


Fig. 1. The statistical analysis of spatial data

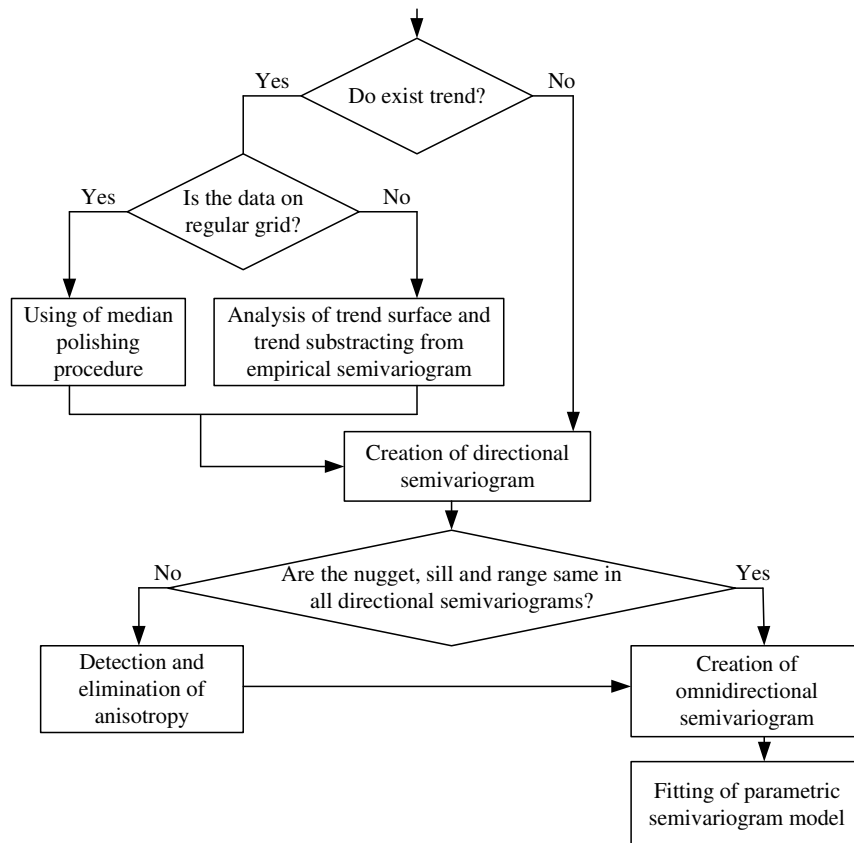


Fig. 2. The block scheme of semivariogram modelling stages

The last stage in semivariogram modelling after elimination of trend and anisotropy is the fitting of parametric models to the empirical semivariogram. The block scheme of semivariogram modelling is shown in Fig. 2.

The analysis of reading materials shows that a popular approach for estimating semivariogram parameters are the methods of least squares semivariogram model fitting and are based on the least squares principle. Those are: ordinary least squares (OLS), generalized least squares (GLS) and weighted least squares (WLS) methods. Under very general conditions, the generalized least squares estimators of semivariogram parameters are asymptotically efficient among all least squares procedures. The possible situations when the described kriging methods (section 2.4) can be used to predict the spatial data are shown in Fig. 3.

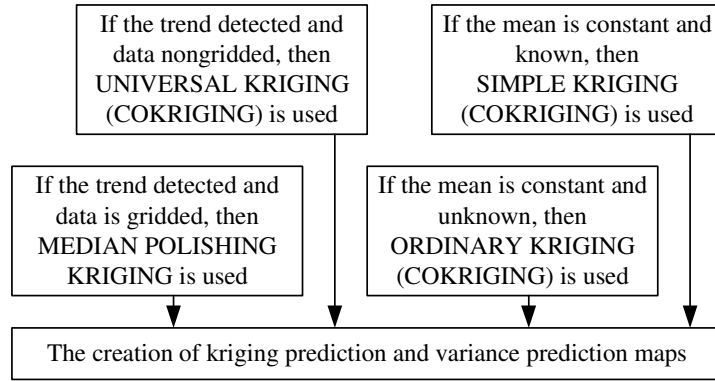


Fig. 3. Using of kriging methods under different conditions

These and other prediction/interpolation procedures were used and compared in the chapter 4 of dissertation.

3. Methods of estimation of semivariogram parameters

In third chapter the new theoretical results are described. Theoretical results investigated for spatial data, which generated by stationary Gaussian spatial processes or by Gaussian process with stationary errors. In this work the term of stationarity is used in both situations.

To estimate the mean parameters of spatial processes and semivariogram parameters the least squares (LS) and maximum likelihood (ML) methods are most popular. In this work main focus is on least squares methods. In the case of estimation of mean parameters, the optimal is GLS estimator. Its expressions

usually available during covariance function. Thus, in the section 3.1 the expression of GLS estimator $\hat{\beta}$ is given by:

$$\hat{\beta}_{GLS}(\theta) = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{Z}_n = (\mathbf{X}^T \Gamma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Gamma^{-1} \mathbf{Z}, \quad (5)$$

where Σ is variance-covariance matrix and Γ – matrix of semivariograms, $\mathbf{Z}_n$ vector of observations, $\mathbf{X}$ – matrix of explanatory variables. Then the D

criterion function for the $\hat{\beta}_{GLS}$ estimator have the form:

$$\Psi_D(\hat{\beta}_{GLS}) = (-1)^q (1 - \mathbf{C}(\theta) \mathbf{G}_{11}) / \det(\mathbf{X}^T \Gamma^{-1} \mathbf{X}), \quad (6)$$

where $\mathbf{G}_{11}$ is element of matrix $\mathbf{X}^T \Gamma^{-1} \mathbf{X}$ in position (1,1).

The A criterion function for the the $\hat{\beta}_{GLS}$ estimator have the form:

$$\Psi_A(\hat{\beta}_{GLS}) = -tr(\mathbf{X}^T \Gamma^{-1} \mathbf{X})^{-1} + \mathbf{C}(\theta). \quad (7)$$

In the section 3.2 the known asymptotical properties of semivariogram estimators MoM methods are described. The case of stationary Gaussian spatial data model, for increasing domain asymptotic type called LIDA (described in Lahiri et al. (2002)), is considered. Under some regularity conditions, the asymptotic normality of semivariogram estimator MoM is proved. The analytical expression of the element (i, j) of asymptotic variance-covariance matrix $\sum_{\gamma}(\theta)$ for isotropic exponential theoretical semivariogram model is derived (section 3.2.).

4. The application of proposed methodology of the data modelling and prediction

In the fourth chapter the proposed methodology of spatial data modelling through the semivariogram and the prediction by kriging is applied to the two data arrays of the Curonian Lagoon in such way:

- 1) The depth data of Curonian Lagoon (263 observations), bottom sediments (213 observations) measured in 1990 year and the depth data (2990 observations) measured in 2005 year are analyzed by preliminary spatial data analysis methods: the Gaussian distribution of data is examined, the histograms and box plot are created, the main descriptive statistics are calculated.
- 2) The empirical semivariogram is calculated; the trend and anisotropy are detected; the linear, spherical, exponential and Gaussian

semivariogram model are fitted to the empirical semivariogram. The suitable semivariogram model by the smallest error sum of squares (SSE) is selected.

- 3) To predict data in entire area of the Curonian Lagoon, when using the data of 1990, the ordinary, universal krigings and ordinary cokriging are used. In the case of data of 2005's, median polish kriging and universal kriging are used. The kriging, which performs the suitable prediction, by the cross-validation method is selected.
- 4) The comparative analysis of two deterministic methods and four krigings methods by terms of SSE has been performed. The prediction results of depth data of 1990 against measured data of 2005'.

Also, in order to identify the optimal design of points in regular grid and apply D and A criterions described in section 3.1, the sampling grid 2x2 with center (0,0) was chosen. The experiment with 5 points was executed. During the experiment one of point with coordinates (0,0) is stable and positions of others points is variable.

General conclusions

After performing the general methodology of modelling and prediction of spatial data generated by spatial processes with stationary errors the following conclusions were made:

1. The expression of GLS estimator for β , when the vector of semivariogram parameters θ is known, was given. In the final expression the variance-covariance matrix is being replaced by matrix of semivariogram accordingly.
2. In the case of nonconstant mean, the D and A optimal criterions are expressed for GLS estimator through the semivariogram. It's determined, that in the case of constant mean model these expressions are equal.
3. The experiment, which was used to determine the optimal sampling design in grid 2x2, showed that the greatest sum of distances ensure the optimality of D and A criterions.
4. The proposed methodology of spatial data modelling through the semivariogram and the prediction by kriging was applied to two data arrays of the Curonian Lagoon measured in different time. To describe the spatial variation of depth data the linear, spherical, exponential and Gaussian semivariogram models have been used. The spherical semivariogram model, in

many cases, describes the spatial variation of Curonian Lagoon depth data in the best way.

5. During the spatial statistical modeling of Curonian Lagoon depth data, was determined that the depth data can be characterized by trend and anisotropy. Most likely the reason of such nonstationarity is the streams of influent rivers.

6. The ordinary kriging, universal kriging, ordinary cokriging, inverse distance interpolation and trend surface interpolation methods were used for prediction/interpolation of depth data of 1990. The comparative analysis of that prediction/interpolation methods and depth data of 2005 by SSE showed that the minimum SSE value is obtained when depth data is predicted by kriging methods including trend and anisotropy.

7. After conducting the prediction of depth data of 2005 by median polishing kriging and universal kriging methods it was determined, that the results of median polish kriging is closer to the data average then the results of universal kriging. The reason for it is the robustness of median polish kriging method.

List of Published Works on the Topic of the Dissertation **In the reviewed scientific periodical publications**

Krūminienė (Borisenko), I.; Dučinskas, K.; Garška, R. 2003. Applying of kriging and cokriging methods for prediction of Curonian Lagoon data, *Lietuvos matematikos rinkinys, special issue 43*: 504–508.

Krūminienė (Borisenko), I.; Garška, R. 2004. Spatial analysis and prediction of Curonian Lagoon data with Gstat, *Mathematical Modelling and Analysis*, 9 (1): 39–50.

Krūminienė (Borisenko), I.; Dučinskas, K. 2004. Generalized product-sum semivariogram model of the data of the center of marine research, *Lietuvos matematikos rinkinys, special issue 44*: 736–741.

Krūminienė (Borisenko), I. 2006. Analysis of anisotropic semivariogram models for prediction of the Curonian Lagoon data. *Mathematical Modelling and Analysis* 11 (1): 73–86.

Krūminienė (Borisenko), I.; Dučinskas, K. 2004. Comparison ordinary and universal kriging of Curonian Lagoon by cross-validation method, *Technologijos mokslo darbai Vakarų Lietuvoje IV*: 185–191.

Krūminienė (Borisenko), I. 2004. Using universal block kriging to prediction of the Curonian Lagoon depth data with gstat package of R system, *Vadyba*, 1(4): 46–50.

In the other editions

Krūminienė (Borisenko), I. 2004. Using Gstat package of R program for spatial data prediction, in *Proceedings of the 9-th MA and PhD Conference „Informacinė visuomenė ir universitetinės studijos“*, 213–219.

Dučinskas, K.; Borisenko, I. 2009. Optimal sampling design for the GLS estimators of the mean parameters for the stationary spatial random process, in *Proceedings of XIIIth International Conference “Applied Stochastic Models and Data Analysis”*: 23–36.

About the author

Ingrida Borisenko was born in Kaunas, on 5th of May 1976.

In the 1995 she started studying in the Faculty of Natural Sciences of Klaipėda University. In 1999 she developed and defended work “The adaptation of Aided-design system for Windows platform”. In this way, she has acquired bachelor's degree of Applied Mathematics science (Computer science specialization). Ingrida Borisenko (at the time Krūminienė) started the MBA's in the Faculty of Natural Sciences and Mathematics of Klaipėda University in 1999. In 2001 she defended the master thesis of numeric analysis and systems studies in mathematics science by title “The Creation of Hydraulic Model of the Curonian Lagoon”. The author was PhD student of the Klaipėda University in 2001–2007. Currently works as the Lecturer in Statistics and Computer Science Departments at Klaipėda University.

ERDVINIŲ DUOMENŲ MODELIAVIMAS NAUDOJANT STACIONARIŲ ERDVINIŲ PROCESŲ SEMIVARIOGRAMAS

Mokslo problemos aktualumas

Erdvinė statistika – viena jauniausių statistikos mokslo krypčių, praėjusio šimtmečio penktame dešimtmetyje pradėta taikyti kalnakasyboje. Per pusšimtį metų nuo šios statistikos atšakos atsiradimo, išaugo ją nagrinėjančių mokslininkų ratas bei taikymo sričių spektras, buvo išleista nemažai teorinio ir praktinio pobūdžio metodinių priemonių. Šiuo metu erdvinės statistikos metodai taikomi ekologijoje, kiekybinėje geologijoje, vaizdų apdorojime ir analizėje, epidemiologiniuose, klimato pokyčių ir net Visatos tyrimuose. Lietuvoje gi erdvinį duomenų tyrimo metodologija pradėta nagrinėti tik šio šimtmečio pradžioje, šio mokslo taikymai vykdyti tik kelių mokslininkų (Dumbrasas, A.; Kumetaitis, A.; Kumetaitienė, A. ir kt.). Tad labai svarbu plėtoti šią sritį, tobulinti esamus metodus, išnagrinėti erdvinį duomenų

modeliavimo metodus ir pateikti bendrinę erdvinį duomenų modeliavimo metodiką.

Norint taikyti erdvinės statistikos metodus, būtina žinoti duomenų išsidėstymą erdvėje, kuris dažniausiai aprašomas geografinėmis koordinatėmis. Tad, vienas esminių bruožų, išskiriančių erdvinę statistiką nuo klasikinės, yra tas, kad ji modeliuoja ir erdvinį trendą, ir erdvinę koreliaciją.

Viena pagrindinių erdvinės statistikos užduočių yra erdvinį duomenų matematinio modelio sudarymas, kuris gali būti naudojamas interpoliacijai (ekstrapoliacijai) ar kitiems tikslams įgyvendinti. Duomenų prognozei atlikti gali būti taikomas kriginio metodas, kuris prie tam tikrų sąlygų minimizuoja vidutinę kvadratinę prognozės paklaidą – pateikia geriausią tiesinę nepaslinktą prognozę. Priklausomai nuo erdvinį duomenų pobūdžio, erdvinio proceso vidurkio modelio ir kitų duomenų savybių kriginiai skirstomi į kelis tipus.

Erdvinį duomenų sklaidą aprašoma kovariacine funkcija arba semivariograma. Pastaroji apima platesnę erdvinį procesų klasę, tad yra dažniau naudojama praktiniuose skaičiavimuose, ypač juos vykdant programiniais paketais. Turimų duomenų pagrindu sudaroma empirinė semivariograma (dažniausiai naudojant momentų metodu gaunamą Matheron'o įvertinį (MoM)), kuriai parenkamas ir priglodinamas vienas iš semivariogramos parametrinių modelių. Dažniausiai naudojami semivariogramų modeliai aprašomi keliais semivariogramos parametrais, kurių pagrindiniai yra grynuolis, slenkstis ir slenkščio atstumas.

Semivariogramų parametrų įvertinimui dažnai taikomi mažiausių kvadratų metodai (svertinis, paprastas bei apibendrintas mažiausių kvadratų metodai). Kaip nurodyta Lahiri ir kitų autorių straipsnyje "On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters" (2002), prie bendrų sąlygų, lyginant su kitais mažiausių kvadratų vertinimo metodais, apibendrintas mažiausių kvadratų metodas (AMKM) vertinimo metodas yra asimptotiškai efektyvus.

Mažiausių kvadratų metodai gali būti taikomi ir vidurkio modelio parametrų įvertinimui. Kadangi AMKM metodas literatūroje pateikiamas per kovariacines funkcijas, svarbu užrašyti šio įvertinio išraišką per semivariogramas.

Taip pat reiktų atkreipti dėmesį, kad semivariogramų parametrų įvertiniai analitiniu būdu nėra gaunami. Todėl aktuali šių įvertinių asimptotinė analizė.

Kitas dažnai erdvinėje statistikoje pasitaikantis uždavinys, yra optimalaus imčių plano nustatymas. Plačiau taikomi du eksperimentų planų optimalumo kriterijai: kriterijus D, kriterijus A. Vėlgi, kadangi praktiniuose skaičiavimuose dažniau naudojamos semivariogramos, aktualu šiuos planus suformuluoti semivariogramų pavidalu.

Tyrimų objektas

Disertacijoje autorė tiria erdvinių duomenų generuotų erdviniais procesais su stacionariomis klaidomis modeliavimo, prognozavimo ir gautų rezultatų tyrimo metodiką.

Darbo tikslas

Pagrindinis darbo tikslas yra apibendrinti ir realizuoti vieningą erdvinių duomenų generuojamų erdviniais procesais su stacionariomis klaidomis modeliavimo metodiką, pagrįstą semivariogramomis.

Darbo uždaviniai

1. Apžvelgti adityvius erdvinių procesų modelius su stacionariomis klaidomis, aprašyti pagrindines erdvinių procesų charakteristikas.
2. Pasiūlyti apibendrintą erdvinių duomenų modeliavimo metodiką, naudojant įvairius parametrinius vidurkių ir stacionarių semivariogramų modelius.
3. Aprašytą bei pasiūlytą metodiką pritaikyti Kuršių marių gylio ir dugno nuosėdų duomenims modeliuoti ir prognozuoti.
4. Atlikti erdvinių imčių planų palyginimus, naudojant įvairius optimalumo kriterijus.
5. Atlikti MoM įvertinio statistinių savybių analizę.

Tyrimų metodika

Analitinėje darbo dalyje aprašant erdvinių duomenų matematinius modelius bei erdvinių duomenų modeliavimo ir prognozavimo metodiką naudojami literatūros ir lyginamosios analizės metodai. Erdvinių duomenų modeliavimo etapams apibendrinti naudojami sąvokų apibendrinimo ir koncepcinio modeliavimo metodai. Darbe apibendrintų modeliavimo ir prognozavimo metodų praktinei realizacijai naudojami eksperimentinės analizės metodai.

Mokslinis naujumas

Šioje disertacijoje pateikiama vieninga erdvinių duomenų su stacionariomis klaidomis modeliavimo metodika, pagrįsta semivariogramomis. Analizuojant erdvinių duomenų su stacionariomis klaidomis modeliavimo metodiką, pastebėtos galimybės tobulinti kai kuriuos erdvinės statistikos metodus.

Šiame darbe pateikti nauji rezultatai:

1. MoM įvertinys pateiktas per erdvinio plano (Laplaso) matricą.
2. Užrašyta β parametru apibendrinto mažiausių kvadratų metodo įvertinio išraiška per semivariogramas.

3. β parametrų įvertiniui išvestos D ir A kriterijų funkcijų išraiškos per semivariogramas.
4. Stacionaraus Gauso erdvinio proceso atvejui įrodytas MoM semivariogramos įvertinio asimptotinis normališkumas, naudojant didėjančios srities asimptotikos schemą.
5. Izotropinės eksponentinės semivariogramos atvejui išvesta minėto įvertinio asimptotinės dispersijų-kovariacijų matricos išraiška.
6. Pateikta apibendrinta erdvinių duomenų generuojamų erdviniais procesais su stacionariomis klaidomis modeliavimo ir prognozavimo metodika, leidžianti:
 - a) duomenų sklaidą aprašant semivariogramomis atsižvelgti į erdvinį trendą bei anizotropiją;
 - b) duomenis prognozuoti ordinariu, universaliu, medianinio išlyginimo krigingais ar kokrigingu;
 - c) taikyti kelis geriausio modelio ar prognozės rezultatų patikros metodus (klaidų kvadratų sumos, kryžminės patikros metodus, Akaike kriterijų).

Praktinė vertė

Darbe pateikta vieninga erdvinių duomenų su stacionariomis klaidomis modeliavimo metodika gali būti naudojama įvairaus pobūdžio (pvz.: ekonometrinių, demografinių ar kt.) erdvinių duomenų analizėje. Teoriniai tyrimų rezultatai gali būti naudojami ieškant optimalių imčių planų. Praktiniai tyrimų rezultatai gali būti pavyzdžiu kaip realizuoti apibendrintą erdvinių duomenų metodiką atviro kodo sistemos R įrankiais.

Ginamieji teiginiai

1. Pasiūlyta vieninga erdvinių duomenų modeliavimo per semivariogramas ir prognozavimo ordinariu, universaliu, medianinio išlyginimo krigingais ir kokrigingu metodika.
2. Stacionaraus Gauso duomenų modelio atvejui pateikta MoM semivariogramų įvertinio dispersijų-kovariacijų matricos išraiška per teorines semivariogramas.
3. Apibendrinto mažiausių kvadratų įvertinio išraiška per semivariogramas β parametrams užrašoma dispersijų-kovariacijų matricas atitinkamai pakeičiant semivariogramų matricomis.
4. β parametrų apibendrinto mažiausių kvadratų įvertinio D ir A kriterijų funkcijų išraiškos išvestos per semivariogramas.
5. MoM semivariogramos įvertinio normališkumas atvejui, kai duomenų modelio klaidos yra stacionarios ir Gausinės.

6. Išvesta MoM semivariogramos įvertinio asimptotinės dispersijų-kovariacijų matricos išraiška Lahiri ir kt. (2002) apibrėžtam didėjančios erdvės asimptotikos erdviniam imčių planui izotropinės eksponentinės semivariogramos atveju.

Darbo struktūra

Darbą sudaro santrumpų ir žymėjimų sąrašas, įvadas, keturi skyriai, išvados, literatūros sąrašas, publikacijų ir konferencijų pranešimų disertacijos tematika sąrašas bei priedai. Bendra disertacijos apimtis – 127 puslapiai, 60 iliustracijų, 20 lentelių ir 4 priedai.

Pirmame skyriuje analizuojami erdvinių duomenų tipai, skaliariniai erdviniai procesai, pagrindinės jų charakteristikos. Nurodomi erdvinių duomenų vidurkio modeliai. Taip pat analizuojamas vienas pagrindinių ir dažnai praktikoje naudojamų sklaidos rodiklių – semivariograma – bei jos parametrai, nurodomi keli semivariogramos įvertiniai, pateikiami populiariausi semivariogramų parametriniai modeliai bei jų naudojimo rekomendacijos.

Antrame disertacijos skyriuje siekiant pasiūlyti apibendrintą erdvinių duomenų modeliavimo ir prognozavimo metodiką pateikiama erdvinių duomenų prognozavimo metodų klasifikacija, nurodomi geostatistinių duomenų modeliavimo ir prognozavimo etapai. Šie etapai vėliau aprašomi detaliau. Ypatingas dėmesys skiriamas semivariogramų modeliavimo etapui, duomenų trendo ir anizotropijos identifikavimui, trumpai apžvelgiamas semivariogramos modelių parametrų vertinimas mažiausių kvadratų metodais. Taip pat aprašomi keli praktiniuose skaičiavimuose naudoti erdvinių duomenų prognozavimo metodai (krigingai), geriausio semivariogramos modelio ar kringingo patikros metodai (kryžminės patikros metodas, Akaike kriterijus). Kiekvienas modeliavimo ir prognozavimo etapą aprašantis skyrelis užbaigiamas etapą vaizduojančia schema. Skyriaus pabaigoje pateikiama semivariogramos parametrų bei modelių įtakos prognozei apžvalga, kringingo privalumų ir trūkumų lyginant su kitais prognozės metodais analizė.

Trečias skyrius skirtas teorinėms darbo naujovėms aprašyti. Čia pateikiama β parametrų apibendrinto mažiausių kvadratų įvertinio išraiška bei D ir A kriterijų funkcijos per semivariogramas. Vėliau analizuojamos MoM semivariogramos įvertinio statistinės savybės. Tuomet naudojant didėjančios erdvės asimptotikos schemą Gauso duomenų modelio atveju įrodomas MoM įvertinio asimptotinis normališkumas. Taip pat izotropinės eksponentinės semivariogramos atvejui pateikiama minėto įvertinio asimptotinės dispersijų-kovariacijų matricos išraiška.

Ketvirtas skyrius skirtas disertacijoje siūlomos erdvinių duomenų modeliavimo ir prognozavimo metodikos taikymui. Dviem – 1990 ir 2005 metų

– duomenų masyvams taikomi statistinės erdvinės duomenų analizės, modeliavimo semivariogramomis bei prognozavimo ordinariu, universaliu, medianinio išlyginimo kriginu bei kokrigingu etapai. Atlikta lyginamoji kelių deterministinių interpoliavimo procedūrų bei prognozės kriginu analizė, kurios metu kelios prognozuotų (interpoliuotų) 1990 metų gylio duomenų reikšmės lyginamos su 2005 metais išmatuotais gylio duomenimis. Taip pat siekiant nustatyti optimalų taškų išsidėstymą taisyklingoje 2×2 gardelėje, kurios centras (0,0), vykdomas eksperimentas su 5 taškais. Skaičiavimams atlikti naudotos įvairios nemokamos R sistemos paketų funkcijos.

Bendrosios išvados

Darbe apibendrinta ir realizuota vieninga erdvinių duomenų su stacionariomis klaidomis modeliavimo metodika. Gautus darbo rezultatus apibendriname ir pateikiame išvadas:

1. Išvesta nepastovaus vidurkio su laisvu nariu parametro $\beta^{\wedge}$ AMKM įvertinio išraiška per semivariogramas, kai semivariogramos parametrų vektorius θ yra žinomas. Pastabėta, kad galutinėje išraiškoje dispersijų-kovariacijų matricą atitinkamai keičia semivariogramos matrica.
2. Nepastovaus vidurkio su laisvu nariu atvejui išvestos D ir A kriterijų AMKM įvertinio išraiškos per semivariogramas. Išsiaiškinta, kad pastovaus vidurkio atveju šios išraiškos yra lygios.
3. Eksperimentas, kurio metu buvo siekiama nustatyti optimalų taškų išsidėstymą taisyklingoje 2×2 gardelėje, įrodė, kad D ir A kriterijų optimalumą garantuoja didžiausia atstumų tarp konfigūracijos taškų suma.
4. Taikant darbe siūlomą vieningą erdvinių duomenų su stacionariomis klaidomis modeliavimo metodiką atliktas skirtingais laiko periodais (1990 ir 2005 metais) matuotų Kuršių marių gylio duomenų erdvinis statistinis modeliavimas. Gylio duomenų erdvinei variacijai aprašyti taikyti tiesinis, sferinis, eksponentinis ir/arba Gauso semivariogramų modeliai. Nustatyta, kad Kuršių marių gylio duomenų erdvinę sklaidą prie įvairių stacionarumo sąlygų daugeliu atveju tiksliausiai aprašo sferinis semivariogramos modelis.
5. Kuršių marių gylio erdvinio statistinio modeliavimo metu, išsiaiškinta, kad gylio duomenims būdingas trendas bei anizotropija. Tikėtina, kad šios nestacionarumo savybės tokio pobūdžio erdviniam procese atsiranda dėl į Kuršių marias įtekančių upių srovių.

6. 1990 metais matuotų gylio duomenų prognozei/interpoliacijai atlikti naudoti ordinarus krigingas, universalus krigingas, ordinarus kokrigingas, atvirkštinio atstumo metodas, trendo paviršiaus interpoliacijos metodas. Šių prognozavimo/interpoliavimo procedūrų rezultatų ir 2005 metais išmatuotų duomenų lyginamoji PKS pagrindu analizė parodė, kad mažiausia PKS reikšmė gaunama duomenis prognozuojant krigingu, kai atsižvelgiama į duomenų tendą ir anizotropiją.

7. Atlikus 2005 metų gylio duomenų prognozę medianinio išlyginimo krigingu bei universaliu krigingu bei palyginus prognozių paklaidas, nustatyta, kad medianinio išlyginimo prognozės rezultatai labiau artimi duomenų vidurkiui dėl prognozės metodo robusiškumo.

Trumpos žinios apie autorių

Ingrida Borisenko gimė 1976 m. gegužės 5 d. Kaune.

Studijas Klaipėdos universitete Gamtos mokslų fakultete pradėjo 1995 metais. 1999 m. parengė ir apgynė diplominį darbą tema „Automatizuoto projektavimo sistemos adaptavimas Windows terpei“. Tokiu būdu ji įgijo taikomosios matematikos mokslo (informatikos specializacijos) bakalauro laipsnį. 1999 m. Ingrida Borisenko (tuo metu Krūminienė) įstojo į magistrantūrą Klaipėdos universitete Gamtos ir matematikos mokslų fakultete. 2001 m. ji apgynė skaitinės analizės ir sistemų studijų matematikos mokslo magistro baigiamąjį darbą tema „Hidraulinio Kuršių marių modelio sudarymas“. 2001–2007 m. autorė studijavo matematikos mokslo krypties doktorantūroje Klaipėdos universitete. Šiuo metu dirba lektore Klaipėdos universiteto Informatikos katedroje.

Ingrida Borisenko

MODELLING OF SPATIAL DATA USING SEMIVARIOGRAMS OF
STATIONARY SPATIAL PROCESSES

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Ingrida Borisenko

ERDVINIŲ DUOMENŲ MODELIAVIMAS NAUDOJANT STACIONARIŲ
ERDVINIŲ PROCESŲ SEMIVARIOGRAMAS

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

2009 09 29. 1,5 sp. l. Tiražas 70 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, LT-10223 Vilnius
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius