An integrated application for geostatistical analysis of sea outfall discharges based on R software

Nuno Abreu
INESC Porto
Campus da FEUP, Rua Dr. Roberto Frias, 378
4200-465 Porto, Portugal
Email: nabreu@inescporto.pt

Patrícia Ramos
INESC Porto
Campus da FEUP, Rua Dr. Roberto Frias, 378
4200-465 Porto, Portugal
ISCAP - IPP
Rua Jaime Lopes Amorim, s/n
4465-004 S. Mamede de Infesta, Portugal
Email: pramos@inescporto.pt

***Abstract*— This paper describes an integrated application that performs a geostatistical analysis of data acquired by an AUV in monitoring missions to sewage outfalls. This comes as an effort for automating the procedures of a monitoring campaign from data acquisition to data processing. This application is based on the *R* statistical software and uses the Gstat package for the geostatistical prediction. *R* is a console based application that uses software packages developed by the community. The application interfaces with R guiding the user through several steps that perform the geostatistical analysis. It was not our intention to cover all geostatistical procedures but only the ones that are needed for the data processing concerned. The major advantage of this application is that the user does not need to be familiar with methods and data structures associated with the base software, allowing the processing and analysis to be more simple, fast and efficient which is particularly important for routine monitoring. This software application also enables us to give a quicker response in case of contamination to near-by beaches.**

I. Introduction

Several ocean outfall plume monitoring approaches have been used to understand the physical, chemical and biological processes associated with coastal treated sewage discharges, which are significant sources of contaminants to coastal ocean ecosystems. Autonomous Underwater Vehicles (AUVs) already proved to be very appropriate for high-resolution surveys of small areas such as outfall plumes [1]. Some of the advantages of these solutions include: easier operation management, low cost per deployment, good spatial coverage and capability of feature-based or adaptive sampling. Increasing environmental awareness and more rigorous environmental standards are pushing for the obtainance of more reliable model predictions.

The effluent's dispersion process, although highly dynamic, tends to a natural variability mode when the plume stops rising and the intensity of turbulent fluctuations approaches to zero. It is likely that after this point the pollutant substances are spatially correlated. Therefore geostatistics seams to be a suitable technique to model the spatial distribution of the outfall plume, not only because it predicts the value of the variable of interest at unsampled locations but specially because it gives the uncertainty associated with the prediction [2][3]. This is one of the advantages of geostatistics over traditional methods for assessing pollution.

A geostatistical analysis involves several steps that can be implemented using *R* statistical software [4]. *R* is a console based application that uses software packages developed by the community. In order to automate the geostatistical procedure, used in processing data acquired by an AUV in a monitoring mission to a sewage outfall, we develop an integrated application that interfaces with *R* and guides the user through the several steps needed [5]. In the next section we present the application developed describing with detail these steps. The details about geostatistical methodology can be found in literature [6][7][8][9][10][11][12]. The Gstat package [13] was used for the geostatistical prediction namely for computing variograms, to do variogram cloud diagnostics and modeling, to perform cross-validation, to do block kriging and mapping.

In the third section we present the most important details about the development and implementation of the software application. We end this paper presenting the major conclusions and suggestions for future work.

II. The geostatistical application

The steps that are performed in a geostatistical analysis of data obtained by an AUV in the vicinity of the wastewater discharge are:

- Load data
- Exploratory data analysis
- Trend analysis
- Variogram construction
- Model adjustment
- Cross-validation
- Kriging

Each one of these steps is described in the next subsections and illustrated with the salinity data set obtained in a monitoring campaign to a Portuguese outfall using an AUV. Details of this campaign can be found in [14].

A. Load data

First the user must load the text file with the dataset. The file is loaded with success if the data are arranged correctly.

978-1-4244-4333-8/10/$25.00 ©2010 IEEE

The position coordinates x and y must be in the first two columns followed by the set of variables available. The names of the variables must be in the first row (followed by x and y) and the size of all columns must be equal to guarantee data consistency.

In this step the user has also to specify if validation is to be performed. (Details of validation procedure are given below.) If the user chooses to perform validation, the size of the data set for validation (in percentage) needs to be specified (25% is the default percentage). If the user does not want to perform validation the geostatistical analysis will be done using the full dataset. Fig. 1 illustrates this first step.

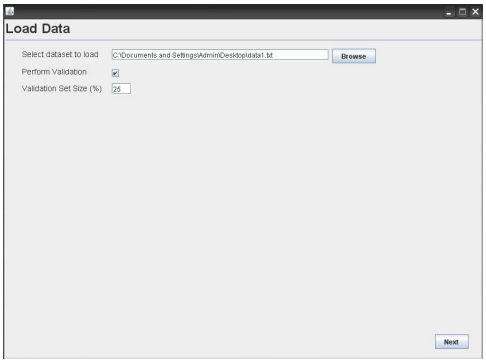

Fig. 1. Load data screen of the geostatistical application.

B. Exploratory data analysis

In order to get an elementary knowledge of the data set an exploratory analysis should be conducted. This analysis defines how we must study the dataset, what we look for and how should we interpret it. The goal of the application in this step is to present an approach that:

- maximizes what is learned from the data;
- gives a graphical representation of the data;
- uncovers the underlying structure;
- derives important information about the variables;
- detects outliers and anomalies;
- tests underlying assumptions.

To achieve this goal a preliminary exploration of the selected variable is done presenting a table with a summary statistics (minimum, maximum, mean, standard deviation, variance, skewness and size of dataset). Additionally an histogram with an adjusted normal probability curve and an overview of the data set are also displayed. In this step the user can also experiment some typical transformations of the data (logarithmic, square root and normal score). Fig. 2 illustrates the exploratory analysis of salinity data obtained at 2 m depth by the AUV in the vicinity of the outfall discharge.

C. Trend analysis

The purpose of trend analysis is to look for a global trend, i.e. a pattern in the data. A trend in the data has to be taken into account when applying the kriging method. If that is the case only the detrended (residual) data will be kriged

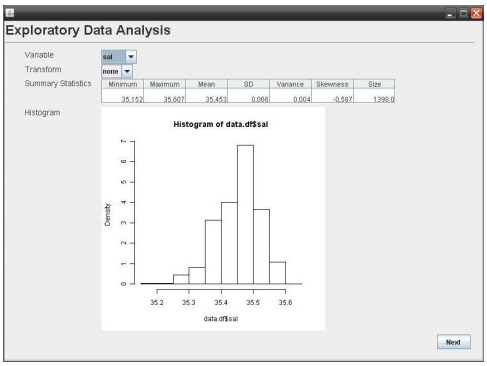

Fig. 2. Exploratory data analysis screen.

using universal kriging. When a trend of the phenomena under investigation, expressed as a function of position coordinates, is already known by the user, it can be specified in the trend analysis screen of the application. At the end of this step a plot of the relationship between the variable of interest and the trend function is displayed. Fig. 3 shows a plot of the relationship between the values of salinity and the corresponding distances between the position of the measurement and the middle point of the outfall diffuser.

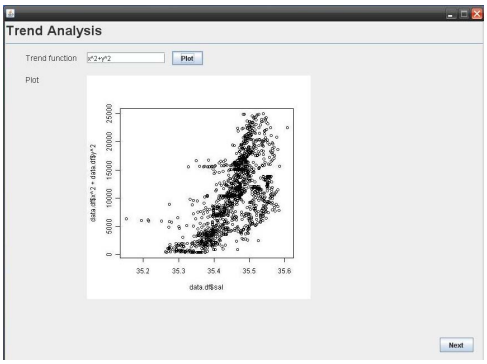

Fig. 3. Trend analysis screen.

D. Variogram construction

The variogram construction is a crucial step in geostatistical analysis. The variogram can be interpreted as a measure of spatial autocorrelation of the variable. It models the process dependence structure and has great impact on the prediction of the spatial random process. Spatial autocorrelation is a statistical relation that defines the variability of a variable with itself through space. When a variable is distributed in space, taking values according to its spatial location, it is known as a regionalized variable. Unlike random variables, regionalized variables exhibit spatial continuity. Despite that, the complexity of the underlying process is such that it cannot be described by any deterministic function. The spatially-correlated component of the regionalized variable is modeled by the variogram.

The estimator of the variogram usually used, known as Matheron's method-of-moments (MME) estimator, is defined

as [12][15]:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left[z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\right]^2 \quad (1)$$

where $z(\mathbf{x}_i)$ is the value of the variable of interest at location $\mathbf{x}_i$, $N(\mathbf{h})$ is the number of pairs of points separated by a distance and direction $\mathbf{h}$ (known as lag), and the quantity $\gamma(\mathbf{h})$ is known as the semivariance at lag $\mathbf{h}$.

Cressie and Hawkins [16] developed an estimator of the variogram that should be robust to the presence of outliers and enhance the variogram spatial continuity, having also the advantage of not spreading the effect of outliers in computing the maps. This estimator (CRE) is defined as follows [16]:

$$\gamma(\mathbf{h}) = \frac{1}{2} \times \frac{\left\{ \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left|z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\right|^{1/2} \right\}^4}{0.457 + \frac{0.494}{N(\mathbf{h})} + \frac{0.045}{[N(\mathbf{h})]^2}} \quad (2)$$

In order to create the variogram the user needs to specify the cutoff and lag width. (The cutoff is the maximum distance up to which point pairs should be considered.) Two different estimators can be chosen in the application: Matheron's method of moments estimator or Cressie's robust estimator. When the "Plot" button of the variogram construction screen is clicked, a plot of the variogram is displayed. Fig. 4 shows the variogram of salinity data calculated using the MME estimator with a cutoff of 120 m and lag width of 2 m.

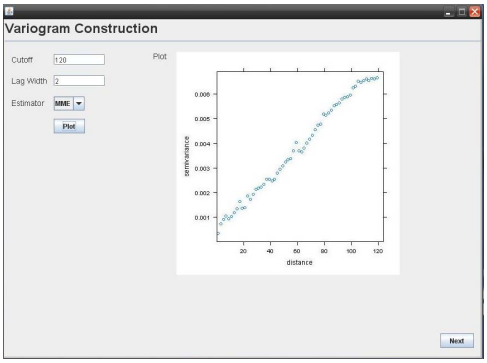

Fig. 4. Variogram construction screen.

E. Model adjustment

Once the variogram has been calculated, a function (called the variogram model) has to be fit to it. The most commonly used variogram models are divided into two types: those that reach a sill (called transition models) and those that don't. (Fig. 5 shows the parameters of the variogram model. The maximum value reached by the variogram is called the sill. The distance at which the sill is reached is called the range. The vertical jump from zero at the origin to the value of semivariance at extremely small separation distances is called the nugget effect.) Variogram models of the second type are usually used when there is a trend in the data values. The most common transition models are the spherical model, the exponential model and the Gaussian model. The linear model increases linearly with the magnitude of $\mathbf{h}$. The choice between the three transition models usually depends on the behavior of the variogram near the origin. If the variogram shows a parabolic behavior near the origin, the Gaussian model will usually provide the best fit. If the variogram has a linear behavior near the origin, either the spherical or exponential model is preferable. When the fitted straight line to the first few points of the variogram intersects the sill at about one fifth of the range, an exponential model will usually provide the best fit. If it intersects the sill at about two thirds of the range, then a spherical model will likely fit better [8]. When there is an abundance of empirical data there is the interest of being able to better characterize the variability of the variogram for short distances. The behavior of the variogram close to the origin is related to the continuity and differentiability of the random field and its local smoothness. The transition model of Matern is very flexible for characterizing this smoothness allowing to model random fields that can be non-differentiable or differentiated an infinite number of times [17][18][19]. The smoothness of the random field is controlled by the shape parameter of the model ν. For $\nu = 0.5$, the Matern model is the exponential model and the Gaussian model is a limiting case of the Matern model when ν tends to infinity. Once the variogram model is chosen, modeling the variogram becomes an exercise of curve fitting in which the parameters of the model are specified [8][12].

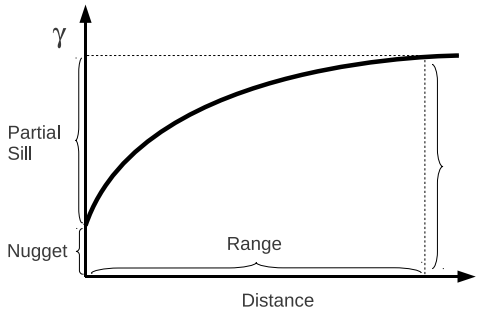

Fig. 5. Variogram model parameters.

In the model adjustment screen the user needs to specify the nugget, sill, range and the model to be adjusted. A plot of the variogram is displayed to help the user to do this specification. When the "Plot" button is clicked the application performs the model adjustment and displays the parameters of the adjusted model. Fig. 6 shows the variogram model for the salinity data set. A Matern model with $\nu = 0.5$ was adjusted.

F. Cross-validation

Cross-validation is a procedure used to compare the performance of several competing models [12]. It starts by splitting the data set into two sets: a modelling set and a validation set. Then the modelling set is used for variogram modelling and kriging on the locations of the validation set. Finally

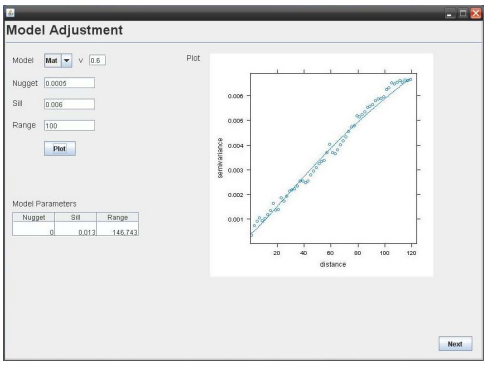

Fig. 6. Model adjustment screen.

the measurements of the validation set are compared to their predictions [5][20]. If the average of the cross-validation errors (or Mean Error, ME) is close to 0,

$$\text{ME} = \frac{1}{m}\sum_{i=1}^{m}[z(x_i) - \hat{z}(x_i)] \tag{3}$$

we may say that apparently the estimates are unbiased ($z(x_i)$ and $\hat{z}(x_i)$ are, respectively, the measurement and estimate at point x_i and m is the number of measurements of the validation set). A significant negative (positive) mean error can represent systematic overestimation (underestimation). The magnitude of the Root Mean Squared Error (RMSE) is particularly interesting for comparing different models [11][12]:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[z(x_i) - \hat{z}(x_i)]^2} \tag{4}$$

The RMSE value should be as small as possible indicating that estimates are close to measurements. The kriging standard deviation represents the error predicted by the estimation method. Dividing the cross-validation error by the corresponding kriging standard deviation allows to compare the magnitudes of both actual and predicted error [11][12]. Therefore, the average of the standardized squared cross-validation errors (or Mean Standardized Squared Error, MSSE)

$$\text{MSSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{[z(x_i) - \hat{z}(x_i)]^2}{\sigma^2_{R(x_i)}} \tag{5}$$

should be about one, indicating that the model is accurate. A scatterplot of true versus predicted values provides additional evidence on how well an estimation method has performed. We typically want that the set of points comes as close as possible to the line $y = x$, a 45-degree line passing through the origin on the scatterplot. The coefficient of determination R^2 is a good index for summarizing how close the points on the scatterplot come to falling on the 45-degree line passing through the origin [8]. R^2 should be close to one.

Fig. 7 shows the cross-validation screen. If the validation procedure was selected by the user in the "Load data screen", a table appears in the screen with descriptive information about the analysis that was previously performed. This descriptive information includes the following performance measures:

- Coefficient of determination (R^2)
- Mean Error (ME)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

In the cross-validation screen the user can choose to experiment other kriging methods, jumping to the previous panels. If a new analysis is conducted, a new row is added to the table with its descriptive information. The kriging method that is highlighted in the table when the "Krige" button is clicked is the one used for mapping in the next step.

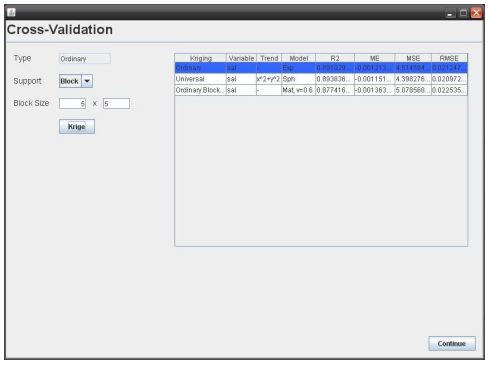

Fig. 7. Cross-validation screen.

G. Kriging

Most geostatistical estimation procedures use stationary random function models. A random function is a set of random variables that have some spatial locations and whose dependence on each other is specified by some probabilistic mechanism. A random function is *stationary* if all the random variables have the same probability distribution and if any pair of random variables has a joint probability distribution that depends only on the separation between the two points and not on their locations. If the random function is stationary, then the expected value and the variance can be used to summarize the univariate behavior of the set of random variables. The parameter that is commonly used to summarize the bivariate behavior of a stationary random function is its covariance function, its correlogram, and its variogram [8][11].

The Ordinary kriging method is often associated with the acronym BLUE which stands for "Best Linear Unbiased Estimator": "Linear"because its estimates are weighted linear combinations of the available data; "Unbiased"since it tries to have the mean error equal to 0; and "Best"because it aims at minimizing the variance of the errors.

For any point at which we want to estimate the unknown value, our model is a stationary random function that consists of n random variables, one for the value at each of the n sample locations and one for the unknown value at the point we are trying to estimate $Z(\mathbf{x}_0)$. Each of these random variables has the same probability law; at all locations, the expected value of the random variable is m and the variance is σ^2. Every value

in this model is seen as an outcome (or realization) of the random variable. Our estimate is also a random variable since it is a weighted linear combination of the random variables at the n sampled locations ([6][7][8][9][10][11][12]):

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{n} w_i \cdot Z(\mathbf{x}_i) \tag{6}$$

The estimation error is defined as the difference between the random variable modeling the true value and the estimate:

$$\varepsilon(\mathbf{x}_0) = Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0) \tag{7}$$

The estimation error is also a random variable. Its expected value, often referred to as the bias, is

$$E\left[\varepsilon(\mathbf{x}_0)\right] = m\left(1 - \sum_{i=1}^{n} w_i\right) \tag{8}$$

Setting this expected value to 0, to ensure an unbiasedness estimate results in:

$$\sum_{i=1}^{n} w_i = 1 \tag{9}$$

This is known as the condition of unbiasedness [8][9][11].

A consideration in many environmental applications has been that ordinary kriging usually yield large prediction errors [5]. This is due to the larger variability in the observations. When predicting averages over larger areas, i.e. within blocks, much of the variability averages out and consequently block mean values have lower prediction errors. If the blocks are not too large the spatial patterns do not disappear. An equivalent procedure, that can be computationally more expensive than block kriging, is to obtain the block estimate by averaging the N kriged point estimates within the block [7][8].

The user is presented with a screen for performing the selected kriging procedure. A combobox is available to specify whether point or block kriging is going to be conducted. If block kriging is selected, the size has to be specified. Changes to the maps scale can also be specified. All the information and plots are displayed, including prediction and variance maps.

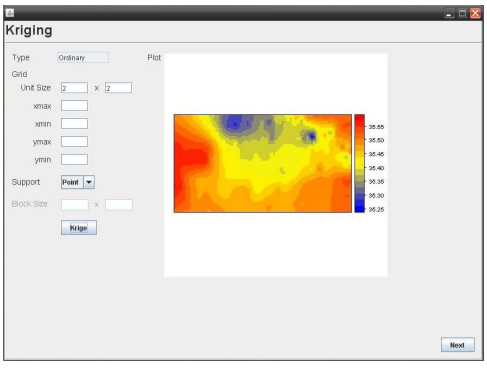

Fig. 8. Kriging screen.

III. Discussion

It is clear that geostatistical interpolation methods are a valid way to analyse the collected data. Many software packages exist for data interpolation and two were particularly interesting for us. It would be either ArcGIS with its Geostatistical Analyst extension or R with the Gstat package.

SRI ArcGIS is the most widely used commercial GIS. The Geostatistical Analyst extension provides a range of global and local interpolation methods including IDW, trend surface, local polynomial, kriging and cokriging. Like other commercial software it fails to give scientific support for the procedures, but it provides an easy to use graphical interface.

R is a free software environment for statistical computing and graphics which is available for most computing platforms. It is used extensively in teaching, in student projects and by researchers. It is completely operated through command line. This is an disadvantage for most people because today most programs have point-and-click interfaces and they are used to it. The main advantage is being open source, providing full access to its internal structure and algorithms. R is developing fast as a research tool in statistical computing, and the results of research and development at universities and laboratories have been often made freely available in the form of software packages.

Gstat is a package (included in R) used for multivariable geostatistical modelling, prediction, simulation and several visualisation functions. It handles a variety of geostatistical analysis and interpolation methods and related features, including change of support (block kriging), simple/ordinary/universal (co)kriging, fast local neighbourhood selection, trend modelling, simulation of large spatially correlated random fields, indicator kriging and simulation, and variogram and cross variogram modelling.

Our choice was R. It was straightforward since it is developed (including packages) by researchers to researchers, gives us the possibility to better understand the used procedures, and obviously, its free. There are also extensive documentation about executing spatial data analysis with R and Gstat.

Not having a graphical interface was a big disadvantage for us, for productivity reasons. So we decided to build one ourselves.

We wanted to embed a R session inside our application. The easier and most efficient way to achieve this was by using JRI. It is a Java/R Interface, which allows to run R inside Java applications as a single thread. It loads R dynamic library into Java and provides a Java API to R functionality. It supports both simple calls to R functions and a full running R engine.

When implementing our application we need to consider how to call R functions and how to pass arguments from Java to R and get the returned results back into Java. JRI enables us to call R functions passing an R expression as a string to the R evaluator. Then this evaluator waits for the response, which is an R object (typically a list). When this is returned to the Java method that called it, standard mechanisms for converting R objects to Java objects are applied.

So we used this approach to build an application that guides the user through a sequential series of panels, each representing a phase of the geostatistical analysis procedure. It is not intended to be an interface to R and gstat since we only need to use a specific set of the available methods. From our perspective it is clear that it is valuable in the sense that it saves time during a analysis but also eliminates learning process that is required for becoming an advanced R/gstat user.

IV. Conclusions

The described application is designed to guide the user through a geostatistical procedure. It uses the R software and gstat package in order to perform geostatistical computations.

It is wizard-based, which may help non-programming users to employ sophisticated algorithms in their analysis.

It allows amateur and novice users in this field to do the required analysis fast and effectively without the trouble of getting acquainted with R, while still maintaining the advantages of using open source software.

With the advancements in AUV technology, other methods may be preferable to analyse the collected data thus making the development of this software continuous.

Acknowledgment

This work was supported by the project WWECO - Environmental Assessment and Modeling of Wastewater Discharges using Autonomous Underwater Vehicles Bio-optical Observations - funded by FCT under Programa I&DT (ref. PTDC/MAR/74059/2006).

References

[1] P. Ramos, M. Monego and S. Carvalho, *Spatial Distribution of a Sewage Outfall Plume Observed with an AUV*, Proceedings of the MTS/IEEE International Conference Oceans 2008, September 15-18, Quebec, Canada, 2008.

[2] P. Ramos and N. Abreu, *Spatial Analysis of Sea Outfall Discharges*, Water Research Conference 2010, April 11-14, Lisbon, Portugal, 2010.

[3] M. Monego, P. Ramos and M.V. Neves, *Geostatistical Mapping of Outfall Plume Dispersion Data Gathered with an Autonomous Underwater Vehicle*, Proceedings of GeoENV 2008 - 7th International Conference on Geostatistics for Environmental Applications, September 8-10, Southampton, UK, 2008.

[4] R Development Core Team, *R: a language and environment for statistical computing and graphics*, http://r-project.org, R Foundation for statistical computing, 2009.

[5] Bivand, R. S. Pebesma and E. J. Gómez-Rubio, V., *Applied spatial data analysis with R*, Series: Use R, XIV, 378 p., Softcover, ISBN: 97 -0-387-78170-9, 2008.

[6] N. Cressie, *Statistics for spatial data*, A Wiley Interscience Publication, New York, 900p, 1993.

[7] P. Goovaerts, *Geostatistics for natural resources evaluation*, Applied Geostatistics Series, ISBN13: 9780195115383, ISBN10: 0195115384, Oxford University Press 496p, 1997.

[8] E. H. Isaaks and R. M. Srivastava, *Applied geostatistics*, New York Oxford, Oxford University Press, ISBN 0-19-505012-6-ISBN 0-19-505013-4 (pbk.) 561p, 1989.

[9] P. Kitanidis, *Introduction to geostatistics: applications in hydrogeology*, New York (USA), Cambridge University Press, 249p, 1997.

[10] M. L. Stein, *Interpolation of spatial data: some theory for kriging*, Springer, New York, 1999.

[11] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*, Berlin, Springer, 291p, 2003.

[12] R. Webster and M. A. Oliver, *Geostatistics for environmental scientists*, 2nd Edition, John Wiley & Sons, Ltd, ISBN-13: 978-0-470-02858-2(HB), 2007.

[13] E. J. Pebesma and C. G. Wesseling, *GSTAT: a program for geostatistical modelling, prediction and simulation*. Comput. Geosci. 24, 1, 1998.

[14] P. Ramos and M. V. Neves, *Environmental Impact Assessment and Management of Sewage Outfall Discharges using AUVs*, Underwater Vehicles, A. Inzartsev ed., ISBN 978-953-7619-49-7, Prentice Hall, In-Tech, Austria, 2009.

[15] G. Matheron, *Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature*, Paris, France: Masson, 305p, 1965.

[16] N. Cressie and D. M. Hawkins, *Robust estimation of the variogram, I*, Journal Int. Assoc. Math. Geol., Vol 12, No. 2, p. 115-125, 1980.

[17] J. P. Chilès and P. Delfiner, *Geostatistics: modeling spatial uncertainty*, Wiley, New York, 695pp, 1999.

[18] B. Minasny and A. B. McBratney, *The matern function as a general model for soil variograms*, Geoderma, Vol. 128(3-4), pp.192-207, 2005.

[19] E. Pardo-Iguzquiza and M. Chica-Olmo, *Geostatistics with the Matern semivariogram model: A library of computer programs for inference, kriging and simulation*, Computers & Geosciences, Vol. 34(9), 1073-79, 2008.

[20] M. Voltz, R. Webster, *A comparison of kriging, cubic splines and classification for predicting soil properties from sample information*, Journal of Soil Science 41, 473-490, 1990.