Spatial analysis of sea outfall discharges using block kriging

Patrícia Ramos [1], **Nuno Abreu** [2]

Keywords

Ordinary block kriging, robust estimation, sewage outfall discharges, plume dispersion maps, environmental risk assessment

Introduction

Several monitoring approaches have been used to understand the physical, chemical and biological processes associated with coastal treated sewage discharges. However, these efforts are not improving knowledge of effluent plume/coastal ocean processes interaction.

Autonomous Underwater Vehicles (AUVs) already demonstrated to be very appropriate for high-resolution surveys of small features such as outfall plumes. Some of the advantages of these platforms include: easier field logistics, low cost per deployment, good spatial coverage, sampling over repeated sections and capability of feature-based or adaptive sampling. Demands for more reliable model predictions, and predictions of quantities that have received little attention in the past are now increasing. These are driven by increasing environmental awareness, more stringent environmental standards, and application of diffusion theory in new areas.

Although very chaotic due to turbulent diffusion, the effluent's dispersion process 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. In this case, geostatistics appears to be an appropriate technique to model the spatial distribution of the effluent. In fact, geostatistics has been used with success to analyze and characterize the spatial variability of soil properties, to obtain information for assessing water and wind resources, to design sampling strategies for monitoring estuarine sediments, to study the thickness of effluent-affected sediment in the vicinity of wastewater discharges, to obtain information about the spatial distribution of sewage pollution in coastal sediments, among others. As well as giving the estimated values, geostatistics provides a measure of the accuracy of the estimate in the form of the kriging variance. This is one of the advantages of geostatistics over traditional methods of assessing pollution.

In this study we use ordinary block kriging to model and map the spatial distribution of temperature measurements gathered with an AUV in a monitoring campaign to a sea outfall, aiming to distinguish the effluent plume from the receiving waters and characterize its spatial variability in the vicinity of the discharge. For the purpose of the analysis, the temperature measurements were divided into a modeling set and a validation set. Models were then developed using data from the modeling set, and tested on the remaining validation set to select the preferred modelling method on the basis of the R^2 and RMSE values. The geostatistical analysis was carried out with the R statistical software environment ([14]). The GSTAT package ([10,11]) was used for computing variograms, to do variogram cloud diagnostics and modeling, to perform cross-validation, to do block kriging and mapping.

Study area

Ordinary kriging

The data set used in this analysis was obtained in a monitoring campaign to *S. Jacinto* outfall,

[1] Patrícia Ramos, Mrs., PhD/Assistant Professor, Instituto de Engenharia de Sistemas e Computadores do Porto, Campus da FEUP, Rua Dr. Roberto Frias 378, 4200-465 Porto, Portugal - +351.22.2094399 – pramos@inescporto.pt. Instituto Superior de Contabilidade e Administração do Porto, Rua Jaime Lopes Amorim s/n, 4465-004 S. Mamede Infesta, Portugal - +351.22.9050000 - patricia@iscap.ipp.pt.

[2] Nuno Abreu, Mr, MSc/Researcher, Instituto de Engenharia de Sistemas e Computadores do Porto, Campus da FEUP, Rua Dr. Roberto Frias 378, 4200-465 Porto, Portugal - +351.22.2094399 - nabreu@inescporto.pt

located off the Portuguese west coast near Aveiro region, using the AUV of Underwater Systems and Technology Laboratory of University of Porto. The outfall diffuser is 98.2 m long, has a true bearing direction of 290º and was discharging at about 15 m depth. A rectangular area of aproximatly 200×100 m^2 starting 20 m downstream from the middle point of the outfall diffuser was covered. As planned, the vehicle performed 6 horizontal trajectories at 2, 4, 6, 8, 10 and 12 m depth. In each horizontal section the vehicle described 6 parallel transects, perpendicular to the current direction (in average 176.1º over the whole water column), of 200 m long spaced of 20 m. While navigating at a constant velocity of approximately 2 knots (about 1m/s), CTD (conductivity, temperature, depth) data were collected and recorded at a rate of 2.4 Hz. Consecutive measurements at horizontal sections were therefore distanced of about 0.4 m. In this study we analyze the temperature data of the horizontal section at 2 m depth, where the effluent plume was found established and dispersing horizontally. Figure 1 shows the temperature measurements obtained by the AUV while it was navigating at that depth.

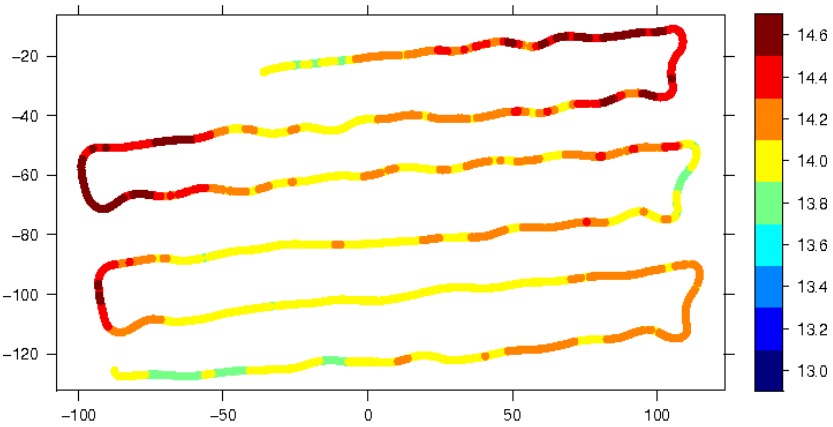

Figure 1 - Plan view of temperature measurements (ºC) at 2 m depth.

Geostatistical analysis

Ordinary kriging

The variables of interest in most earth science data sets are typically the end result of a large number of processes whose complex interactions we usually are not able to describe quantitatively. So we are forced to admit that there is some uncertainty about how the phenomenon behaves between the sample locations. The random function models take into account this uncertainty and give us tools for estimating values at unknown locations once we have made some assumptions about the statistical characteristics of the phenomenon ([5]).

The most widely used 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 ([5,6]). The complete definition of the probabilistic generating mechanism of a random function is usually difficult even in one dimension. Fortunately, for many of the problems we typically encounter, we do not need to know the probabilistic generating mechanism. We usually adopt a stationary random function as our model and specify only its covariance or variogram ([5,18]).

Ordinary kriging method is often referred 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. Let us then see how the concept of a random function model can be used to decide how to weight the nearby samples so that our estimates are unbiased.

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, $Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2), \ldots, Z(\boldsymbol{x}_n)$, and one for the unknown value at the point we are trying to estimate $Z(\boldsymbol{x}_0)$. 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 ([3,4,5,6,15,18,19]):

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

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

$$\varepsilon(\boldsymbol{x}_0) = Z(\boldsymbol{x}_0) - \hat{Z}(\boldsymbol{x}_0). \tag{2}$$

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

$$E\left[\varepsilon(\boldsymbol{x}_0)\right] = m\left(1 - \sum_{i=1}^{n} w_i\right). \tag{3}$$

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

$$\sum_{i=1}^{n} w_i = 1. \tag{4}$$

This is known as the condition of unbiasedness ([5,6,18]). The expression of the variance of the modeled error is

$$\mathrm{var}\left[\varepsilon(\boldsymbol{x}_0)\right] = \sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \mathrm{var}\left[Z(\boldsymbol{x}_0)\right] - 2\,\mathrm{cov}\left[\hat{Z}(\boldsymbol{x}_0), Z(\boldsymbol{x}_0)\right] + \mathrm{var}\left[\hat{Z}(\boldsymbol{x}_0)\right]. \tag{5}$$

Since we have already assumed that all of the random variables have the same variance σ^2, then $\mathrm{var}\left[Z(\boldsymbol{x}_0)\right] = \sigma^2$. The second term in Eq. 5 can be written as

$$-2\,\mathrm{cov}\left[\hat{Z}(\boldsymbol{x}_0), Z(\boldsymbol{x}_0)\right] = -2\sum_{i=1}^{n} w_i \cdot \mathrm{cov}\left[Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_0)\right] = -2\sum_{i=1}^{n} w_i \cdot C_{i0}. \tag{6}$$

The third term of Eq. 5, the variance of $\hat{Z}(\boldsymbol{x}_0)$ can be expressed as

$$\mathrm{var}\left[\hat{Z}(\boldsymbol{x}_0)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i \cdot w_j \cdot \mathrm{cov}\left[Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, C_{ij}. \tag{7}$$

Then, the expression of the error variance $\sigma^2_{\varepsilon(\boldsymbol{x}_0)}$ comes in the following way:

$$\sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \sigma^2 - 2\sum_{i=1}^{n} w_i\, C_{i0} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, C_{ij}. \tag{8}$$

The Eq. 8 expresses the error variance as function of the n weights, once chosen the random model function parameters, namely the variance σ^2 and all the covariances C_{ij}. The minimization of $\sigma^2_{\varepsilon(\boldsymbol{x}_0)}$ is constrained by the unbiasedness condition imposed earlier, which can be solved using the method of Lagrange multipliers. We start by introducing a new parameter μ, called the Lagrange multiplier, in Eq. 8 in the following way:

$$\sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \sigma^2 - 2\sum_{i=1}^{n} w_i\, C_{i0} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, C_{ij} + \underbrace{2\,\mu\left(\sum_{i=1}^{n} w_i - 1\right)}_{=0}. \tag{9}$$

Then we minimize $\sigma^2_{\varepsilon(\boldsymbol{x}_0)}$ by calculating the $n+1$ partial first derivatives of Eq. 9 with respect to the n weights and the Lagrange multiplier, and setting each one to 0, which produces the following system of equations:

$$\frac{\partial\left(\sigma^2_{\varepsilon(\boldsymbol{x}_0)}\right)}{\partial\left(w_1\right)} = -2\,C_{10} + 2\sum_{j=1}^{n} w_j\, C_{1j} + 2\mu = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} w_j\, C_{1j} + \mu = C_{10}$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\frac{\partial\left(\sigma^2_{\varepsilon(\boldsymbol{x}_0)}\right)}{\partial\left(w_i\right)} = -2\,C_{i0} + 2\sum_{j=1}^{n} w_j\, C_{ij} + 2\mu = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} w_j\, C_{ij} + \mu = C_{i0}$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\frac{\partial\left(\sigma^2_{\varepsilon(\boldsymbol{x}_0)}\right)}{\partial\left(w_n\right)} = -2\,C_{n0} + 2\sum_{j=1}^{n} w_j\, C_{nj} + 2\mu = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} w_j\, C_{nj} + \mu = C_{n0}$$

$$\frac{\partial\left(\sigma^2_{\varepsilon(\boldsymbol{x}_0)}\right)}{\partial\left(\mu\right)} = 2\left(\sum_{i=1}^{n} w_i - 1\right) = 0 \qquad \Rightarrow \qquad \sum_{i=1}^{n} w_i = 1$$

which can also be written in a compact way as

$$\sum_{j=1}^{n} w_j\, C_{ij} + \mu = C_{i0}, \quad \forall i = 1, 2, \ldots, n; \qquad \sum_{i=1}^{n} w_i = 1\,. \tag{10}$$

This system of equations, often referred to as the *ordinary kriging system*, can be written in matrix notation as

$$\underset{}{\boldsymbol{C}} \cdot \boldsymbol{w} = \boldsymbol{D}$$

$$\underbrace{\begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}}_{(n+1)\times(n+1)} \cdot \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix}}_{(n+1)\times 1} = \underbrace{\begin{bmatrix} C_{10} \\ \vdots \\ C_{n0} \\ 1 \end{bmatrix}}_{(n+1)\times 1}. \tag{11}$$

The set of weights and the Lagrange multiplier that will produce an unbiased estimate of $Z(\boldsymbol{x}_0)$ with the minimum error variance are then given by

$$\boldsymbol{w} = \boldsymbol{C}^{-1} \cdot \boldsymbol{D}\,. \tag{12}$$

The value of $\sigma^2_{\varepsilon(\boldsymbol{x}_0)}$ can be obtained in a quicker way using an alternative expression to Eq. 8. Multiplying each of the n equations given in Eq. 10 by w_i and summing these n equations leads to the following:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, C_{ij} = \sum_{i=1}^{n} w_i\, C_{i0} - \mu\,.$$

Substituting this into Eq. 8 the minimized error variance comes as follows:

$$\sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \sigma^2 - \left(\sum_{i=1}^{n} w_i\, C_{i0} + \mu\right). \tag{13}$$

The Eq. 13 expressed in terms of matrices is

$$\sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \sigma^2 - \boldsymbol{w}^T \cdot \boldsymbol{D}\,. \tag{14}$$

The minimized error variance is usually called the *ordinary kriging variance*.

Block kriging

A consideration in many environmental applications has been that ordinary kriging usually exhibits large prediction errors ([12]). 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. The *block kriging* system is similar to the point kriging system given by 11 ([5]). The matrix $\boldsymbol{C}$ is the same since it is independent of the location at which the block estimate is required. The covariances for the vector $\boldsymbol{D}$ are point-to-block covariances. Supposing that the mean value over a block V is approximated by the arithmetic average of the N point variables contained within that block ([4,5]), i.e.

$$Z_V \approx \frac{1}{N}\sum_{j=1}^{N} Z(\boldsymbol{x}_j), \tag{15}$$

the point-to-block covariances required for vector $\boldsymbol{D}$ are

$$\bar{C}_{iV} = \mathrm{cov}\left[Z(\boldsymbol{x}_i), Z_V\right] = \frac{1}{N}\sum_{j=1}^{N} C_{ij}, \quad \forall i = 1,2,\ldots,n. \tag{16}$$

The block kriging variance is

$$\sigma_V^2 = \bar{C}_{VV} - \left(\sum_{i=1}^{n} w_i \bar{C}_{iV} + \mu\right), \tag{17}$$

where $\bar{C}_{VV}$ is the average covariance between pairs of points within V:

$$\bar{C}_{VV} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}. \tag{18}$$

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 ([4,5]).

Spatial continuity

Spatial continuity exists in most earth science data sets. When we look at a contour map, or anything similar, the values do not appear to be randomly located, but rather, low values tend to be near other low values and high values tend to be near other high values. I.e. two measurements close to each other are most likely to have similar values than two measurements far apart ([5]). H-scatterplots can be used to display the spatial continuity of a data set. The x-coordinate of a point corresponds to value of the variable of interest at a particular location and the y-coordinate to the value of the variable of interest a distance and direction $\boldsymbol{h}$ away.

The shape of the cloud of points on an h-scatterplot tells us how continuous the data values are over a certain distance in a particular direction. If the data values at locations separated by $\boldsymbol{h}$ are very similar then the pairs will plot close to the line $x = y$, a 45-degree line passing through the origin. As the data values become less similar, the cloud of points on the h-scatterplot becomes fatter and more diffuse. The correlation coefficient ρ, the covariance C_{xy}, and the moment of inertia MI, can be used to summarize quantitatively the fatness of the cloud of points on an h-scatterplot ([5]):

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sigma_x\,\sigma_y} = \frac{C_{xy}}{\sigma_x\,\sigma_y} \qquad \mathrm{MI} = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2, \tag{19}$$

where, the number of data is n; $x_1,\ldots,x_n$ are the data values for the x-axis, m_x is their mean, and σ_x is their standard deviation; $y_1,\ldots,y_n$ are the data values for the y-axis, m_y is their mean,

and σ_y is their standard deviation. As the cloud of points gets fatter, we expect the correlation coefficient and the covariance to decrease and we expect the moment of inertia MI to increase ([5]).

To compute the set of weights and the Lagrange multiplier, that will produce each estimate and the resulting minimized error variance, we need to know the covariances of $\boldsymbol{C}$ and $\boldsymbol{D}$ matrices. As we said before, since our random function is stationary, all pairs of random variables separated by a distance and direction $\boldsymbol{h}$ (known as lag) have the same joint probability distribution. The covariance function, $C(\boldsymbol{h})$ is the covariance between random variables separated by a lag $\boldsymbol{h}$ ([5,6,18]):

$$C(\boldsymbol{h}) = \text{cov}\left[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\right] = E\left[Z(\boldsymbol{x})\, Z(\boldsymbol{x}+\boldsymbol{h})\right] - E\left[Z(\boldsymbol{x})\right] E\left[Z(\boldsymbol{x}+\boldsymbol{h})\right]. \tag{20}$$

For a stationary random function, the covariance function $C(\boldsymbol{h})$ is:

$$C(\boldsymbol{h}) = E\left[Z(\boldsymbol{x})\, Z(\boldsymbol{x}+\boldsymbol{h})\right] - \left\{E\left[Z(\boldsymbol{x})\right]\right\}^2. \tag{21}$$

The covariance between random variables at identical locations is the variance of the random function:

$$C(\boldsymbol{0}) = E\left[\left\{Z(\boldsymbol{x})\right\}^2\right] - \left\{E\left[Z(\boldsymbol{x})\right]\right\}^2 = \text{var}\left[Z(\boldsymbol{x})\right] = \sigma^2. \tag{22}$$

The correlogram is the correlation coefficient between random variables separated by a lag $\boldsymbol{h}$:

$$\rho(\boldsymbol{h}) = \frac{\text{cov}\left[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\right]}{\sqrt{Z(\boldsymbol{x})\, Z(\boldsymbol{x}+\boldsymbol{h})}} = \frac{C(\boldsymbol{h})}{C(\boldsymbol{0})}. \tag{23}$$

The correlation coefficient between random variables at identical locations is one, i.e. $\rho(\boldsymbol{0}) = 1$. The semivariogram, or simply variogram, is half the expected squared difference between random variables separated by a lag $\boldsymbol{h}$:

$$\gamma(\boldsymbol{h}) = \frac{1}{2} E\left[\left\{Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right\}^2\right] = \frac{1}{2}\text{var}\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right]. \tag{24}$$

The quantity $\gamma(\boldsymbol{h})$ is known as the semivariance at lag $\boldsymbol{h}$. The "semi" refers to the fact that it is half of a variance. The variogram between random variables at identical locations is zero, i.e. $\gamma(\boldsymbol{0}) = 0$. Using Eqs. 21, 22 and 24, we can relate the variogram with the covariance function as:

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}) = \sigma^2 - C(\boldsymbol{h}). \tag{25}$$

For the random functions usually used, the pairs of widely separated random variables are independent on one another. Therefore, the covariance function and the correlogram eventually reach zero while the variogram eventually reaches a maximum value, usually called the sill. (The distance at which the sill is reached is called the range.) The sill value of the variogram is also the variance of the random function, σ^2. The vertical jump from zero at the origin to the value of semivariance at extremely small separation distances is called the nugget effect. The graph of the variogram is a mirror image of the graph of the covariance function about the x-axis ([5,6,18]). Using Eqs. 23 and 25 we can relate the variogram with the correlogram as:

$$\gamma(\boldsymbol{h}) = \sigma^2\left[1 - \rho(\boldsymbol{h})\right]. \tag{26}$$

Using Eqs. 10 and 25 we can write the ordinary kriging system in terms of the variogram:

$$\sum_{j=1}^{n} w_j\, \gamma_{ij} - \mu = \gamma_{i0}, \quad \forall i = 1, 2, \ldots, n; \qquad \sum_{i=1}^{n} w_i = 1, \tag{27}$$

and the ordinary kriging variance is given by:

$$\sigma^2_{\varepsilon(\boldsymbol{x}_0)} = \sum_{i=1}^{n} w_i\, \gamma_{i0} - \mu. \tag{28}$$

The ordinary kriging system and the kriging variance can also be written in terms of the

correlogram using a similar approach. In practice, the pattern of spatial continuity chosen for the random function is usually taken from the spatial continuity evident in the sample data set. Geostatisticians usually define the spatial continuity of the sample data set through the variogram (known as sample variogram or experimental variogram) and solve the ordinary kriging system using covariance ([5]). The estimator of the variogram usually used, known as Matheron's method-of-moments estimator (MME) is ([7,19])

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \left[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h}) \right]^2 . \tag{29}$$

where $Z(\boldsymbol{x}_i)$ is the value of the variable of interest at location $\boldsymbol{x}_i$ and $N(\boldsymbol{h})$ is the number of pairs of points separated by the particular lag vector $\boldsymbol{h}$. Notice that the variogram is the relationship between the moment of inertia of an h-scatterplot and $\boldsymbol{h}$. Cressie and Hawkins ([2]) 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 ([2]):

$$\gamma(\boldsymbol{h}) = \frac{1}{2} \times \frac{\left\{ \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \left| Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h}) \right|^{1/2} \right\}^4}{0.457 + \frac{0.494}{N(\boldsymbol{h})} + \frac{0.045}{\left[N(\boldsymbol{h}) \right]^2}} . \tag{30}$$

Once the sample variogram has been calculated, a function (called the variogram model) has to be fit to it. First because the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ may need semivariance values for lags that are not available from the sample data. And second, because the use of the sample variogram does not guarantee the existence and uniqueness of solution to the ordinary kriging system. To be guaranteed one and only one solution, we must ensure that the kriging matrices are positive definite ([5]). Usually we guarantee that by fitting the sample variogram with functions that are known to be positive definite. The most commonly used variogram models are divided into two types: those that reach a sill (called transition models) and those that not. Variogram models of the second type are usually used when there is a trend or drift 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 $\boldsymbol{h}$. The choice between the three transition models usually depends on the behavior of the sample variogram near the origin. If the sample variogram shows a parabolic behavior near the origin, the Gaussian model will usually provide the best fit. If the sample 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 sample 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 ([5]). When there is an abundance of empirical data, as is the case here, there is the interest of being able to better characterize the variability of the sample variogram for short distances. The behavior of the sample 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 ([1,8,9]). 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 sample variogram becomes an exercise of curve fitting in which the parameters of the model are specified ([5,19]).

Cross-validation

Cross-validation is a procedure used to compare the performance of several competing models. 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 the measurements of the validation set are compared to their predictions ([12,16]). If the average of the cross-validation errors (or Mean Error, ME) is close to 0,

$$\mathrm{ME} = \frac{1}{m}\sum_{i=1}^{m}\left[Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)\right]. \quad (31)$$

we may say that apparently the estimates are unbiased ($Z(\boldsymbol{x}_i)$ and $\hat{Z}(\boldsymbol{x}_i)$ are, respectively, the measurement and estimate at point $\boldsymbol{x}_i$ 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 ([18,19]):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left[Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)\right]^2}. \quad (32)$$

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 ([18,19]). Therefore, the average of the standardized squared cross-validation errors (or Mean Standardized Squared Error, MSSE)

$$\mathrm{MSSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{\left[Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)\right]^2}{\sigma^2_{R(\boldsymbol{x}_i)}}. \quad (33)$$

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 ([5]). R^2 should be close to one.

Results

In order to get an elementary knowledge of the data set, conventional statistical analysis was conducted for the temperature measured using software R. Basic statistic characteristics of the temperature data set obtained at 2 m depth are presented in Table 1.

Table 1 - Summary statistics of temperature measured at 2 m depth.

	Temperature (2 m)
Samples	2426
Mean	14.14ºC
Median	14.10ºC
Minimum	13.64ºC
Maximum	14.63ºC
Lower Quartile	14.02ºC
Upper Quartile	14.22ºC
Standard Deviation	0.19ºC
Interquartile Range	0.21ºC
Coefficient of Skewness	0.70
Coefficient of Variation	0.01

At 2 m depth the temperature ranged from 13.64ºC to 14.63ºC. The mean value of the data set

was 14.14ºC, very close to the median value that was 14.10ºC. The coefficient of skewness is not very high (0.70) indicating that the distribution is only slightly asymmetric. A distribution is considered highly positively/negatively skewed when the coefficient of skewness is higher/lower than 1/-1 ([19]). The very low value of the coefficient of variation (0.01) reflects the fact that the histogram does not have a tail of high values. A coefficient of variation greater than one indicates the presence of some erratic high sample values that may have a significant impact on the final estimates ([5]).
Ordinary kriging method works better if the distribution of the data values is close to a normal distribution ([5]). Therefore, it is interesting to see how close the distribution of our data values comes to being normal. Figure 2 shows the plot of the normal distribution adjusted to the histogram of the temperature measurements. Apart from some erratic high values we may say the histograms reasonably approximate the normal distribution.

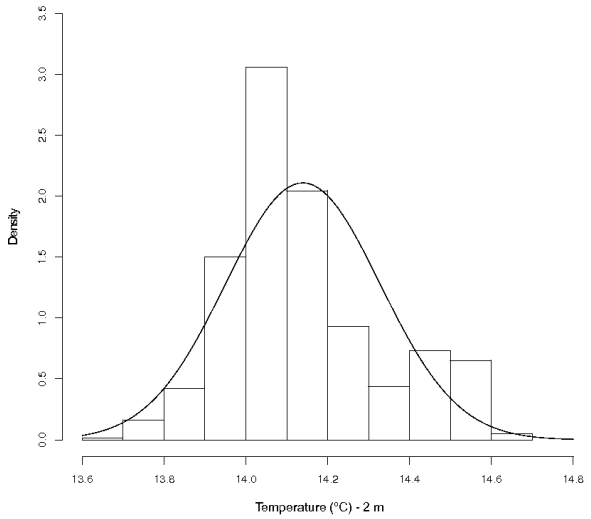

Figure 2 - Histogram of temperature measurements.

For the purpose of the analysis, the temperature measurements were divided into a modeling set (comprising 75% of the samples) and a validation set (comprising 25% of the samples). Modeling and validation sets were then compared in terms of their temperature measurements, using Student's-t test, to check that they provided unbiased sub-sets of the original data. Then, sample variograms for the modeling sets were constructed using the Matheron's method-of-moments estimator (MME) and the Cressie and Hawkins robust estimator (CRE). This robust estimator was choosen for dealing with outliers and enhance the variogram spatial continuity.
Figure 3 and Figure 4 show the omnidirectional sample variograms for temperature using, respectively, MME and CRE fitted by Matern models and by a spherical and an exponential model.
Estimation of semivariance was carried out using a lag distance of 2 m. The parameters of the fitted models are presented in Table 2. The range value is an indicator of extension where autocorrelation exists. The variograms show significant differences in range. For the same model, the autocorrelation distances are always larger for the CRE. The semivariance estimated by CRE also fluctuates less, demonstrating the enhancement of the variogram spatial continuity. All variograms have low nugget values and very low nugget/sill ratios. For all variograms the nugget/sill ratio is less than 0.05. These results indicate that local variations could captured, due to the high sampling rate, and that the variable under study has strong spatial dependence. Anisotropy was investigated by calculating directional variograms. However, no effect of anisotropy could be shown.

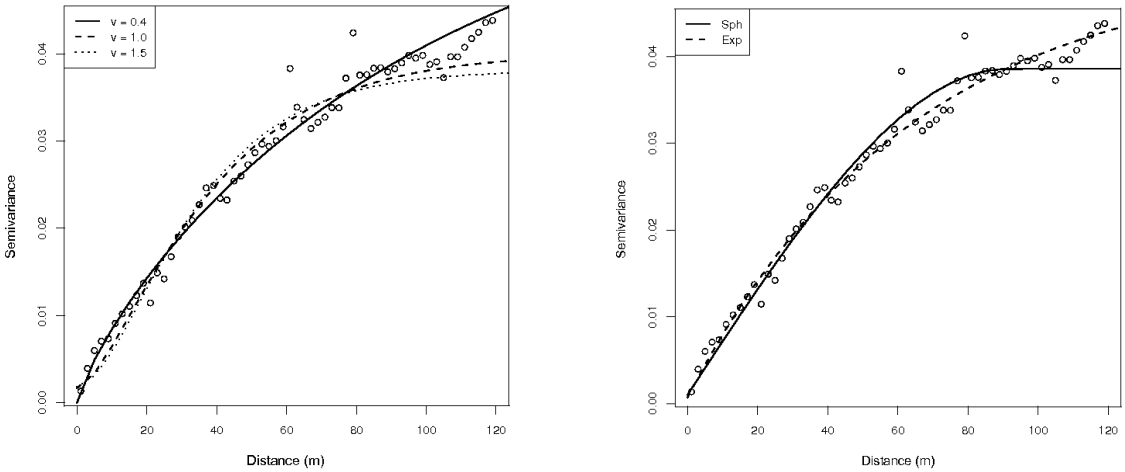

Figure 3 - Variogram for temperature using MME fitted by Matern models (left) and a spherical and an exponential model (right).

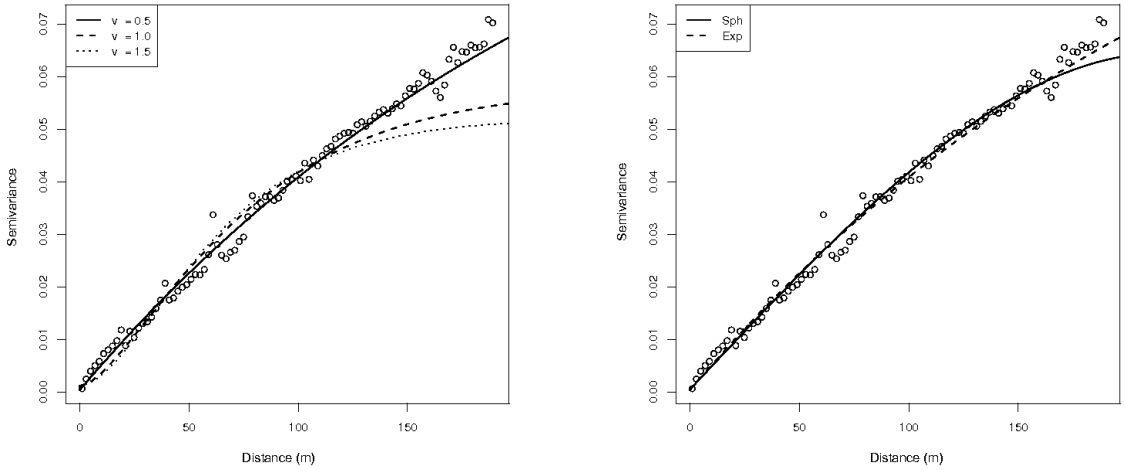

Figure 4 - Variogram for temperature using CRE fitted by Matern models (left) and a spherical and an exponential model (right).

Block kriging method was preferred since it produced smaller prediction errors and smoother maps than point kriging. Using the 75% modeling set, two-dimensional ordinary block kriging, with blocks of 10×10 m^2, was applied to estimate temperature at the locations of the 25% validation sets. The validation results are shown in Table 3. Temperature was best estimated by the Matern model $\nu = 0.5$ using the Cressie and Hawkins robust estimator. The difference in performance between the two estimators: block kriging using the Matheron's method-of-moments estimator (MBK) or block kriging using the Cressie and Hawkins robust estimator (CBK), for the several models, is not substantial.

Table 2 - Parameters of the fitted variogram models.

Variogram estimator	Model	Cutoff (m)	Nugget	Sill	$\frac{\text{Nugget}}{\text{Sill}}$	Range (m)
MME	Matern ($\nu = 0.4$)	120	0.00000	0.063	0	115.5
	Matern ($\nu = 1.0$)	120	0.00169	0.038	0.04	25.3
	Matern ($\nu = 1.5$)	120	0.00186	0.036	0.05	17.9
	Spherical	120	0.00099	0.038	0.03	91.1
	Exponential	120	0.00069	0.050	0.01	63.6
CRE	Matern ($\nu = 0.5$) Exponential	190	0.00038	0.125	0	255.6
	Matern ($\nu = 1.0$)	190	0.00116	0.057	0.02	50.7
	Matern ($\nu = 1.5$)	190	0.00129	0.051	0.03	33.4
	Spherical	190	0.00049	0.064	0.01	215.7

Table 3 - Validation results.

Method	Model	R^2	ME	RMSE	MSSE
MBK	Matern ($\nu = 0.4$)	0.9146	0.00264	0.0541	5.5
	Matern ($\nu = 1.0$)	0.9125	0.00305	0.0547	17.0
	Matern ($\nu = 1.5$)	0.9067	0.00314	0.0565	25.9
	Spherical	0.9137	0.00284	0.0544	8.5
	Exponential	0.9150	0.00279	0.0539	7.1
CBK	Matern ($\nu = 0.5$)[a] Exponential	0.9154	0.00277	0.0538	11.4
	Matern ($\nu = 1.0$)	0.9085	0.00307	0.0560	34.4
	Matern ($\nu = 1.5$)	0.8967	0.00326	0.0595	54.6
	Spherical	0.9148	0.00279	0.0540	12.2

a The preferred model.

Figure 5 shows the scatterplots of true versus estimated values for the most satisfactory model. This plot shows that observed and predicted values are highly positively correlated. The R^2 value was 0.9154 and the RMSE was 0.0538ºC (Table 3). The MSSE values are relatively high probably due to the smoothing effect of block kriging.

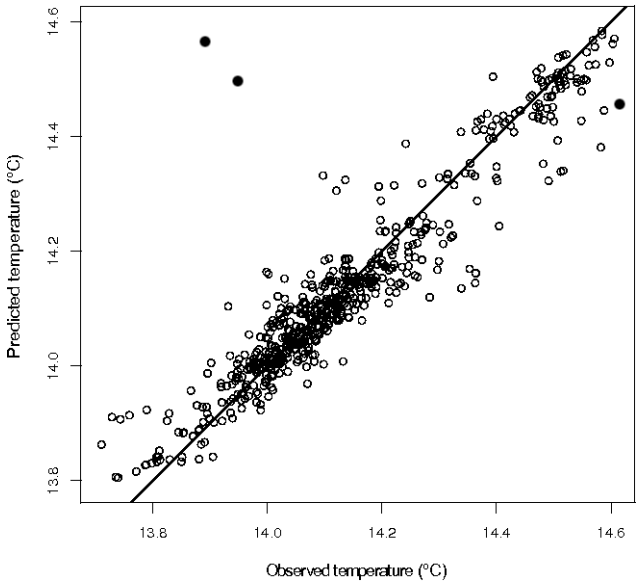

Figure 5 - Predicted versus observed temperature using the preferred model.

Figure 6 shows the block kriged map on a grid 2×2 m^2 using the preferred model. The range in temperature measurements was 13.64-14.63ºC, and the mean value was 14.14ºC. The minimum predicted temperature in the dispersion map is 13.80ºC and the maximum predicted temperature is 14.58ºC, with a mean predicted temperature of 14.15ºC.

The effluent temperature was measured at the discharge chamber nearshore using a 24h-composed sample and the average value obtained was 35ºC. This value is not so relevant since probably the effluent cools down to the soil temperature during its trip of about 3.4 km through the outfall pipe. A vertical profile obtained by the AUV at the end of the campaign measured a background temperature of about 14.5ºC at 2 m depth.

The prediction map shows clearly the spatial variation of temperature in the studied area. From this map it is possible to identify the effluent plume and its dispersion downstream in the current direction. It appears as a region of lower temperature compared to the surrounding ocean waters at the same depth. It is also possible to observe the plume edges since the wastefield width is shorter than the survey width. This indicates that the sampling strategy adopted, based on a plume prediction model runned in real time during the campaign, was successful, even for a surfacing plume which can be considered as the most complicated case in terms of natural tracer tracking. The major difference in temperature compared to the surrounding waters is about -0.7ºC. Washburn et al. ([17]) observed temperature anomalies in the plume only of the order -0.3ºC compared with the surrounding waters in the same depth range. A gradient in temperature at the effluent plume lateral edges is also visible, being the wastefield spreading almost centered in the survey area. The plume spreading direction agrees quite well with the measured currents. The plume exhibits a considerably more complex structure than the compact shape of the classical picture of the buoyant plume, but not so patchy as in previous studies ([13,17,20]), maybe due to the increase in horizontal resolution and also possibly due to the successful kriging results. The small plume-related anomaly in the local temperature of Figure 6 is evidence of the rapid mixing process. Due to the large differences in density between the rising effluent plume and ambient ocean waters, entrainment and mixing processes are vigorous, and the properties within the plume change rapidly ([17]). These results confirm then that large gradients in background temperature and small differences in temperature between the effluent plume and the ambient waters can easily obscure the signature of the plume.

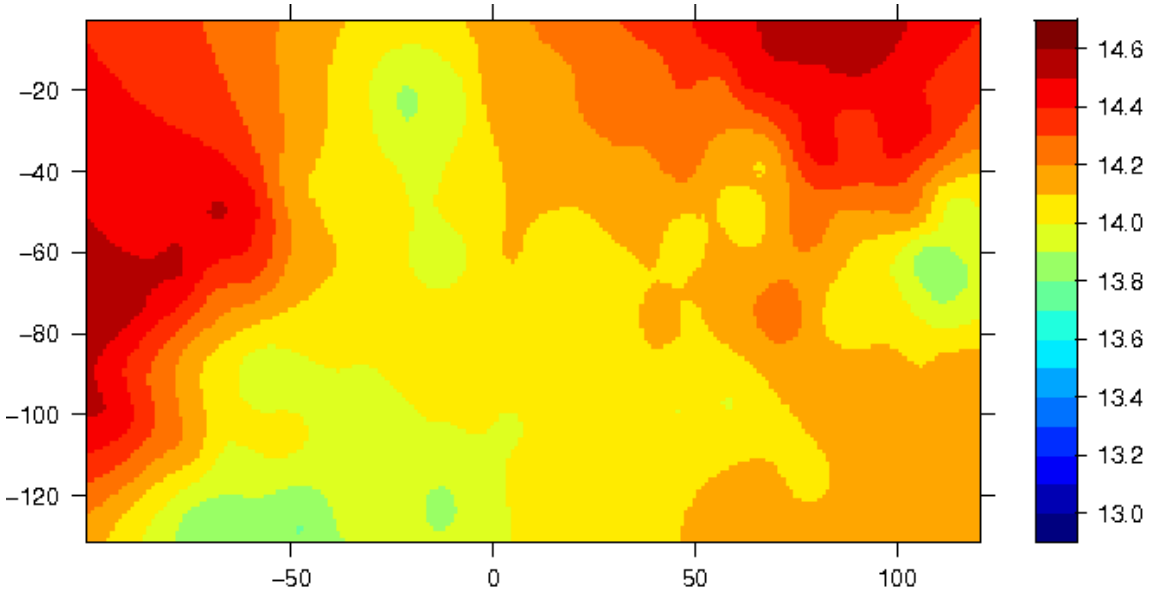

Figure 6 - Prediction map of temperature distribution at 2 meters depth.

Conclusions

Geostatistical analysis of temperature, obtained by an AUV at 2 m depth in a monitoring campaign to an ocean outfall, was able to produce a kriged map of the sewage dispersion in the field. The spatial variability of the sampled data was analyzed previously and the results indicated an approximated normal distribution of the temperature measurements, which is desirable. Then, Matheron's classical estimator and Cressie and Hawkins robust estimator were used to compute the omnidirectional variograms that were fitted to Matern models (for several shape parameters) and to spherical and exponential models. For the same model, the autocorrelation distances were always larger for the CRE and semivariance also fluctuated less when estimated by CRE, demonstrating the enhancement of the variogram spatial continuity. The performance of the several competing models was compared using a split-sample approach. The validation results using two-dimensional ordinary block kriging suggested the Matern model ($\nu = 0.5$) with semivariance estimated by CRE. The difference in performance between the two estimators was not substantial. Block kriged map of temperature at 2 m shows clearly the spatial variation of temperature in the studied area and it is possible to identify the effluent plume that appears as a region of lower temperature compared to the surrounding waters.

Acknowledgements
This work was supported by the project WWECO - Environmental Assessment and Modeling of Wastewater Discharges using Autonomous Underwater Vehicles Bio-optical Observations - funded by Fundação para a Ciência e a Tecnologia under Programa de Projectos de Investigação em todos os domínos científicos (ref. PTDC/MAR/74059/2006).

Patrícia Alexandra Gregório Ramos obtained her licenciate degree in Applied Mathematics – Field of Computer Science from Faculty of Sciences of University of Porto in 1993, Master degree in Electrical and Computer Engineering- Field of Systems from Faculty of Engineer of University of Porto (FEUP) in 1996 and PhD in Engineering Sciences from FEUP in 2005. She is Assistant Professor at the Department of Mathematics of Institute of Accountancy and Administration of Porto, Polytechnic School of Porto. Her research activity is carried out in the Group of Robotics and Intelligent Systems at INESC Porto, in the areas of monitoring of ocean outfalls using Autonomous Underwater Vehicles, assessment and management of sea outfalls, sewage plumes climatology, plumes dispersion mapping, geostatistics, environmental impact studies of wastewater discharges, etc.

Instituto de Engenharia de Sistemas e Computadores do Porto
Campus da FEUP, Rua Dr. Roberto Frias 378, 4200-465 Porto, PORTUGAL

Instituto Superior de Contabilidade e Administração do Porto
Rua Jaime Lopes Amorim s/n, 4465-004 S. Mamede Infesta, PORTUGAL

Nuno Abreu obtained his Integrated Master degree in Electrical and Computer Engineering from Faculty of Engineer of University of Porto (FEUP) in 2008. He is researcher of the Group of Robotics and Intelligent Systems at Instituto de Engenharia de Sistemas e Computadores do Porto and student of the Doctoral Programme in Electrical and Computer Engineering of FEUP. Some of his research areas include assessment and management of sea outfalls, plumes dispersion mapping, geostatistics, Java/R Interface.

Instituto de Engenharia de Sistemas e Computadores do Porto
Campus da FEUP, Rua Dr. Roberto Frias 378, 4200-465 Porto, PORTUGAL

References

[1] Chilès, J.P., Delfiner, P.: "Geostatistics: modeling spatial uncertainty", Wiley, New York, 695p, 1999.
[2] Cressie, N., Hawkins, D.M.: "Robust estimation of the variogram, I", Jour. Int. Assoc. Math. Geol., Vol. 12, No. 2, pp. 115-125, 1980.
[3] Cressie, N.: "Statistics for spatial data", A Wiley Interscience Publication, New York, 900p, 1993.
[4] Goovaerts, P.: "Geostatistics for natural resources evaluation", Applied Geostatistics Series, ISBN13: 9780195115383, ISBN10: 0195115384, Oxford University Press, 496p, 1997.
[5] Isaaks, E.H., Srivastava, R.M.: "Applied geostatistics", New York Oxford, Oxford University Press, ISBN 0-19-505012-6-ISBN 0-19-505013-4 (pbk.), 561p, 1989.
[6] Kitanidis, P.: "Introduction to geostatistics: applications in hydrogeology", New York (USA), Cambridge University Press, 249p, 1997.
[7] Matheron, G.: "Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature", Paris, France: Masson, 305p, 1965.
[8] Minasny, B., McBratney, A.B.: "The matern function as a general model for soil variograms", Geoderma, Vol. 128(3-4), pp.192-207, 2005.
[9] Pardo-Iguzquiza, E., Chica-Olmo, M.: "Geostatistics with the Matern semivariogram model: A library of computer programs for inference, kriging and simulation", Computers & Geosciences, Vol. 34(9), pp. 1073-79, 2008.
[10] Pebesma, E.J., Wesseling, C.G.: "Gstat: a program for geostatistical modelling, prediction and simulation", Computers & Geosciences, Vol. 24, No. 1, pp. 17-31, 1998.
[11] Pebesma, E.J.: "Multivariable geostatistics in S: the gstat package", Computers & Geosciences, Vol. 30, No. 1, pp. 683-691, 2004.
[12] Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V.: "Applied spatial data analysis with R", Series: Use R, XIV, Softcover, ISBN: 97 -0-387-78170-9, 378p., 2008.
[13] Petrenko, A.A., Jones, B.H., Dickey, T.D.: "Shape and initial dilution of Sand Island, Hawaii sewage plume", Journal of Hydraulic Engineering, ASCE Vol. 124, No. 6, pp. 565-571, 1998.
[14] R Development Core Team: "R: a language and environment for statistical computing and graphics", http://www.r-project.org, R Foundation for statistical computing, 2009.
[15] Stein, M.L.: "Interpolation of spatial data: some theory for kriging", Springer, New York, 1999.
[16] Voltz, M., Webster, R.: "A comparison of kriging, cubic splines and classification for predicting soil properties from sample information", Journal of Soil Science 41, 473-490, 1990.
[17] Washburn, L., Jones, B.H., Bratkovich, A., Dickey, T.D., Chen, M.-S.: "Mixing, dispersion, and resuspension in vicinity of ocean wastewater plume", Journal of Hydraulic Engineering, ASCE Vol. 118, No. 1, pp. 35-58, 1992.
[18] Wackernagel, H.: "Multivariate geostatistics: an introduction with applications", Berlin, Springer, 291p, 2003.
[19] Webster, R., Oliver, M.A.: "Geostatistics for environmental scientists", 2nd Edition, John Wiley & Sons, Ltd, ISBN-13: 978-0-470-02858-2(HB), 2007.
[20] Wu, Y., Washburn, L., Jones, B.H.: "Buoyant plume dispersion in a coastal environment: evolving plume structure and dynamics", Continental Shelf Research, Vol. 14, No. 9, pp. 1001-1023, 1994.