Environmental Modelling & Software 26 (2011) 1349–1357

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A Copula based observation network design approach

Jing Li [a,*], András Bárdossy [b], Lelys Guenni [c], Min Liu [d]

[a] School of Civil, Environmental and Mining Engineering, The University of Adelaide, Australia
[b] Institute of Hydraulic Engineering, University of Stuttgart, 70569 Stuttgart, Germany
[c] Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar, Venezuela
[d] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 5 November 2010
Received in revised form
25 April 2011
Accepted 1 May 2011
Available online 2 June 2011

Keywords:
Copulas
Geostatistics
Observation network design
Utility
Spatial dependence
Estimation uncertainty

ABSTRACT

In this paper, a method for environmental observation network design using the framework of spatial modeling with copulas is proposed. The methodology is developed to enlarge or redesign an existing monitoring network by taking the configuration which would increase the expected gain defined in a utility function. The utility function takes the estimation uncertainty, critical threshold value and gain-loss of a certain decision into account. In this approach, the studied spatial variable is considered as a random field in where variations in time is neglected and the variable of interest is static in nature. The uniqueness of this approach lies in the fact that the uncertainty estimation at the unsampled location is based on the full conditional distribution calculated as conditional copula in this study. Unlike the traditional Kriging variance which is a function of mere measurements density and spatial configuration of data points, the conditional copula account for the influence from data values. This is important specially if we are interested in purpose oriented network design (pond) as for example the detection of noncompliance with water quality standards, the detection of higher quantiles in the marginal probability distributions at ungauged locations, the presence or absence of a geophysical variable as soil contaminants, hydrocarbons, golds and so on. An application of the methodology to the groundwater quality parameters in the South-West region of Germany shows its potential.

1. Introduction

An observation network is usually required to measure specific environmental variables to evaluate compliance with regulatory standards. In this context, statistical methods can be applied to provide a reasonable degree of assurance that certain criteria is met as well as to estimate the impact of additional measurements on error reduction before the measurement are taken. Based on that one can decide where to collect additional measurement so that the objectives of *monitoring* are met in the most cost-effective way (Kitanidis, 1997).

Different design strategies are proposed in the literature. Entropy based approaches can be dated back to Lindley (1956) and Bernardo (1979). Caselton and Zidek (1984), Guttorp et al. (1993) and Zidek et al. (2000) have developed the maximum entropy design approach, where entropy is used as a measure of the uncertainty about the variable of interest. The design/sampling criterion is to maximize the amount of variability of the samples so that conditionally on the sample the unsampled population has minimum variability (Shewry and Wynn, 1987). In this context, doable optimization algorithm was developed for Gaussian process. In Keats et al. (2010), a Bayesian approach is applied to solve the inverse problem of isolating the source of an unknown contaminant emission, where Markov chain and Monte Carlo and a posterior sampling technique are used to calculate the expected information over a grid of potential detector locations. Geostatistical based approaches are also available, e.g., Cressie (1991), Mardia and Goodall (1993), Journel (1994). In these approaches, spatial dependence is described by the variogram and the optimal design of the observation network is obtained basically by minimizing the related estimation variance, which is expressed by the Kriging variance. The variogram and Kriging variance is a mere function of measurement density and spatial configuration of the data points, but is not dependent on the variable values. The latter factor cannot be neglected in many cases. For example, high concentration data points might be more dependent than the low ones and this difference propagates to the estimation uncertainties of different values. These facts have important consequences for network design, especially for extreme events detection (Chang et al., 2007).

The problem addressed above brings the genesis of the copula based approach described in this paper. As apposed to the

* Corresponding author. Tel.: +61 4917624078631.
E-mail address: jing.li@gmx.de (J. Li).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.05.001

traditional geostatistical tools, copula enables to model the dependence as a function of the variable values, thus the difference in estimation uncertainties between different quantiles can be reflected. Another advantage of copula is that it captures the pure dependence without the influence from marginal distributions which often cause a problem for traditional geostatistical analysis (Bárdossy, 2006). Recently it has become increasingly popular in environmental sciences, to name a few, Grimaldi and Serinaldi (2006), Gebremichael and Krajewski (2007), Bárdossy and Pegram (2009), AghaKouchak and Bádossy (2010), Klein et al. (2010) and Salvadori and Michele (2010). A comprehensive reference can be found in http://www.stahy.org. Readers who are interested in more details of copula theory can refer to Nelsen (1999), Genest and Favre (2007) and Salvadori et al. (2007).

The first step in the proposed approach is to consider a random field in where variations in time can be neglected and the variable of interest is static in nature. As apposed to the traditional geostatistical tools, the spatial dependence of the natural variable under interest is modeled using copulas. The uncertainties in the estimation of unsampled locations are described with the help of the conditional copula. False positive or negative decisions are penalized with a certain cost, while right decisions are favored with a certain gain. Then a utility function is defined such that the gain or loss of a certain decision is weighted by the conditional copula of the possible state of nature of the variable under interest. The candidate location which yields the highest utility for the whole estimation grid will be selected as the best place for adding a new measurement.

In the following sections, first the proposed methodology is described in details. After that a synthetic example is presented to better explain the methodology. Then a case study on redesigning part of the existing observation network of groundwater nitrate concentration in the south-western region of Germany is discussed to show the potentialities of this approach. Final remarks about this study are drawn at the end.

2. Description of the proposed methodology

We denote $Z(\mathbf{s})$ as a spatial random field of interest indexed by location $\mathbf{s}$ in a two dimensional space and $\mathbf{s}\epsilon\boldsymbol{S}$, where S is the study domain. In a purpose oriented network design, we are interested in measuring at locations where the value of the variable $Z(\mathbf{s})$ is above a certain threshold β such that:

$$\theta(s) = \begin{cases} \theta_0(s) & Z(s) < \beta \\ \theta_1(s) & Z(s) \geq \beta \end{cases} \quad (1)$$

where $\theta(\mathbf{s})$ is a dichotomic spatial random field representing the state of nature at each location $\mathbf{s}$.

From the point of view of decision theory, we propose a very simple approach by using decision theory tools. At any location $\mathbf{s}$, we can take the decision d_i of taking water or not depending on our judgment on whether the unknown location gets a value below the threshold β or not. The positive decision (taking water) is denoted as d_0, and the counterpart is denoted as d_1. We can define a utility function $U_s(\theta_i, d_i)$, depending on the state of nature $\theta_i(\mathbf{s})$ at each location and the decision to be taken. The term *utility* is a measure of the relative satisfaction from consumption of various goods and services or attainment of goals (Ingersoll, 1987). The defined utility function for our case is presented in Table 1.

The entries of the utility matrix k_{ij} can be defined in relative units. Negative k_{ij} values represent costs, while positive ones represent gains. Note that the diagonal elements k_{00} and k_{11} correspond to gains if a correct decision is taken, while the off diagonal ones correspond to losses when a wrong decision is made. It should be mentioned that the gains and costs can be very different for different cases. For instance, exceeding a threshold of a certain pollutant concentration in groundwater can have a severe adverse effect if used for drinking purpose. On the other hand, to decide not to use the water, even if in reality, the threshold is not exceeded leads, generally, to a small loss. But in the case of gold mining, the reverse holds true, i.e., the loss of deciding not to mine but actually there is gold might be higher than deciding to mine but no gold is present. The expected utility at each location $\mathbf{s}$ for a certain decision d_i can then be calculated with the help of the coefficients of utility function as:

Table 1
Utility matrix.

$U_s(\theta_i, d_i)$	θ_0	θ_1
d_0	k_{00}	k_{01}
d_1	k_{10}	k_{11}

$$E(Us|di) = k_{i0}p(\theta(s) = \theta_0) + k_{i1}p(\theta(s) = \theta_1)\; i = 0, 1 \quad (2)$$

The decision should be taken based on the estimated probability $p(\theta(s) = \theta_0)$. If it exceeds a certain limit p_l then d_0 is taken, else d_1 will be taken. That means if the probability of pollutant concentration or gold grade being below the critical threshold is big enough so that the risk of taking water is sufficiently low or it is not worth mining, the decision d_0 is chosen and vice versa. The aforementioned terms *positive decision* and *negative decision* are relative to the specific design purpose. In the project of water quality control, if $p = p(Z(\mathbf{s}) < \beta)$ is greater than a certain limit, the decision of using water will be made which means d_0 indicates a positive decision. While in the project of gold mining, in this case the decision of not mining will be taken, and hence d_0 indicates a negative decision. Since we are mainly dealing with environmental variables, in the following, we will comply with the definition for the water quality control case.

The value of the limit p_l should be specified according to the principle that the expected utility is maximized. That means, if $p(\theta(\mathbf{s}) = \theta_0) > p_l$ then the expected utility of the positive decision is greater than that of the negative decision:

$$k_{00}p + k_{01}(1-p) > k_{10}p + k_{11}(1-p) \quad (3)$$

else if $p(\theta(\mathbf{s}) = \theta_0) < p_l$ then:

$$k_{00}p + k_{01}(1-p) < k_{10}p + k_{11}(1-p) \quad (4)$$

For $p(\theta(\mathbf{s}) = \theta_0) = p_l$, the utilities of the two opposite decisions should be equivalent:

$$k_{00}p + k_{01}(1-p_l) = k_{10}p + k_{11}(1-p_l) \quad (5)$$

which leads to:

$$p_l = \frac{k_{11} - k_{01}}{k_{00} - k_{01} - k_{10} + k_{11}} \quad (6)$$

Note that a continuous utility function depending on the exact value could also be considered in a similar manner. In order to take the appropriate decision, the probability $P(Z(\mathbf{s}) < \beta)$ at an unsampled location have to be estimated from the available observations. This can be done by applying appropriate interpolation procedures of the random field $Z(\mathbf{s})$ at a given set of unsampled locations denoted by $\mathbf{S}^* = (\mathbf{s}^*_1, \ldots, \mathbf{s}^*_N)$. For this purpose, the approach described in Bárdossy and Li (2008) is applied. The probability $P(Z(\mathbf{s}) < \beta)$, i.e., $p(\theta(\mathbf{s}) = \theta_0)$ required by Equation (2) at the unsampled location is calculated as the conditional copula at the unsampled locations $C_{\mathbf{sj}^*,\, n}(u_j^*|u_1, \ldots, u_n)$ for $j = 1, \ldots, N$, $u^*_j = F_Z(\beta)$ with $F_Z(\cdot)$ denoting the univariate empirical distribution of the dataset onwards, conditioned on the n observations.