Computers & Geosciences 35 (2009) 1255–1270

Contents lists available at ScienceDirect

Computers & Geosciences

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

AUTO-IK: A 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences ☆

P. Goovaerts *

BioMedware, 516 North State Street, Ann Arbor, MI 48104, USA

ARTICLE INFO

Article history:
Received 26 April 2008
Received in revised form
12 July 2008
Accepted 4 August 2008

Keywords:
Interpolation
Thresholds
Cross-validation
Jack-knife
Accuracy
Fortran

ABSTRACT

Indicator kriging (IK) provides a flexible interpolation approach that is well suited for datasets where: (1) many observations are below the detection limit, (2) the histogram is strongly skewed, or (3) specific classes of attribute values are better connected in space than others (e.g. low pollutant concentrations). To apply indicator kriging at its full potential requires, however, the tedious inference and modeling of multiple indicator semivariograms, as well as the post-processing of the results to retrieve attribute estimates and associated measures of uncertainty. This paper presents a computer code that performs automatically the following tasks: selection of thresholds for binary coding of continuous data, computation and modeling of indicator semivariograms, modeling of probability distributions at unmonitored locations (regular or irregular grids), and estimation of the mean and variance of these distributions. The program also offers tools for quantifying the goodness of the model of uncertainty within a cross-validation and jack-knife frameworks. The different functionalities are illustrated using heavy metal concentrations from the well-known soil Jura dataset. A sensitivity analysis demonstrates the benefit of using more thresholds when indicator kriging is implemented with a linear interpolation model, in particular for variables with positively skewed histograms.

1. Introduction

Two common features of environmental datasets are the occurrence of a few very large concentrations (hot-spots) and the presence of data below the detection limit (censored observations). Extreme values can strongly affect the characterization of spatial patterns, and subsequently the prediction. Several approaches exist to handle strongly positively skewed histograms (Saito and Goovaerts, 2000). One common approach is to first transform the data (e.g. normal score, Box Cox, or lognormal transform), perform the analysis in the transformed space, and back-transform the resulting estimates. Such transform, however, does not solve problems created by the presence of numerous censored data since either it yields a spike of similar transformed values or, in the case of the normal-score transform, it requires a necessarily subjective ordering of all equally valued observations. Moreover, except for the normal score transform (Deutsch and Journel, 1998), it does not guarantee the normality of the transformed histogram, which is required to compute confidence intervals for the estimates. Last, the back-transform of estimated moments is not straightforward and can introduce bias if not done properly (Saito and Goovaerts, 2000); for example, lognormal kriging estimates cannot simply be exponentiated. Another way to attenuate the impact of extreme values is to use more robust statistics and estimators. The non-parametric approach of indicator kriging (IK) falls within that category (Journel, 1983; Goovaerts, 2001). The basic idea

☆ Code available from server at http://www.iamg.org/CGEditor/index.htm.

* Tel.: +1 734 913 1098; fax: +1 734 913 2201.
E-mail address: goovaerts@biomedware.com

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.08.014

is to discretize the range of variation of the environmental attribute by a set of thresholds (e.g. deciles of sample histogram, detection limit, and regulatory threshold) and to transform each observation into a vector of indicators of non-exceedence of each threshold. Kriging is then applied to the set of indicators and estimated values are assembled to form a conditional cumulative distribution function (ccdf). The mean or median of the probability distribution can be used as an estimate of the pollutant concentration (e.g. Barabás et al., 2001; Cattle et al., 2002; Goovaerts et al., 2005).

A frequent criticism of the indicator approach is that the binary coding amounts to discarding some of the information in the data. In theory, this loss of information can be compensated by accounting for indicator values defined at different thresholds, i.e. using indicator cokriging instead of kriging. Practice has shown, however, that indicator cokriging improves little over indicator kriging (Goovaerts, 1994; Pardo-Igúzquiza and Dowd, 2005), because cumulative indicator data carry substantial information from one threshold to the next one, and all indicator values are available at each sampled location (isotopic or equally sampled case). Another way to increase the resolution of the discrete ccdf is to conduct a fine discretization of the continuous sample distribution using a large number of thresholds. For example, 15 indicator cutoffs were used by Lark and Ferguson (2004) to map the risk of soil nutrient deficiency in a field of Nebraska. Goovaerts et al. (2005) used indicator kriging with 22 thresholds to model probabilistically the spatial distribution of arsenic concentrations in groundwater of Southeast Michigan. Cattle et al. (2002) used 100 threshold values to characterize the spatial distribution of urban soil lead contamination. The extreme situation is to identify the set of thresholds with the sample dataset, i.e. to use as many thresholds as observations. In this case, typically only observations the closest to the interpolated location (e.g. located within the search window) are used as thresholds. Such tailoring of thresholds to the local information available leads to a better resolution of the discrete ccdf by selecting low thresholds in the low-valued parts of the study area and high thresholds in the high-valued parts (Saito and Goovaerts, 2000; Lloyd and Atkinson, 2001; Cattle et al., 2002).

The trade-off costs for the finer resolution of the ccdf are the tedious inference and modeling of multiple indicator semivariograms, as well as the increasing likelihood that the estimated probabilities would not honor the axioms of a cumulative distribution function: all probabilities must be valued between 0 and 1 and form a non-decreasing function of the threshold value. Failure to honor such constraints, referred to as order relation deviations, requires the a posteriori correction of the set of estimated probabilities (Deutsch and Journel, 1998). To keep these deviations within reasonable limits, Deutsch and Lewis (1992) recommend using no more than 9–15 thresholds. Several authors have proposed alternate implementations of the indicator approach that reduce the proportion and magnitude of order relation deviations, while maintaining a reasonable resolution for the ccdf. For example, Pardo-Igúzquiza and Dowd (2005) developed a procedure that requires solving a single indicator cokriging system at each location, leading to far fewer order relation problems than the traditional indicator (co)kriging. Two other implementation tips (Goovaerts, 1997) are to avoid sudden changes in indicator semivariogram parameters from one threshold to the next, and to select thresholds z_k, so that within each search neighborhood there is at least one datum from each class (z_{k-1}, z_k). This is ensured by using locally adaptive thresholds (i.e. thresholds identified with observations within the search window) and the same semivariogram model (i.e. semivariogram for the median threshold) for all thresholds (Saito and Goovaerts, 2000; Lloyd and Atkinson, 2001). For large datasets Cattle et al. (2002) developed a program where indicator semivariograms are computed and modeled locally, whereas the same 100 global thresholds are used across the entire study area.

A critical, yet often overlooked, step in the non-parametric approach is the interpolation or extrapolation of the corrected probabilities to derive a continuous ccdf model. Statistics of the local probability distribution, such as the mean or variance, may overly depend on the modeling of the upper and lower tails of the distribution (Goovaerts, 1997). Popular software, such as Gslib (Deutsch and Journel, 1998) or SGEMS (Remy et al., 2009), offer a piecewise interpolation/extrapolation of the ccdf model: a linear model is usually adopted for interpolation within each class, whereas power or hyperbolic models are used for extrapolation beyond the two extreme threshold values. The choice of these models is, however, completely arbitrary and usually poorly documented. An alternative, which is implemented in the computer code described in this paper, is to capitalize on the higher level of discretization of the cdf (i.e. the cumulative histogram) to improve the within-class resolution of the ccdf. It is noteworthy that a few authors proposed to accomplish the correction and interpolation/extrapolation of ccdf estimates in one step using logistic regression (Pardo-Igúzquiza and Dowd, 2005) or through the fitting of a continuous function (Cattle et al., 2002). In all cases, the impact of extrapolation models can be reduced by selecting more threshold values within the two tails of the distribution (Deutsch and Lewis, 1992; Chu, 1996).

This paper presents an automated implementation of non-parametric geostatistics that integrates Gslib routines for semivariogram computation and indicator kriging with a Fortran code for semivariogram modelling (Pardo-Igúzquiza, 1999). Topsoil heavy metal concentrations from the Jura dataset (Atteia et al., 1994) are used to illustrate the impact of the number of thresholds and type of interpolation model on results, such as the magnitude of prediction errors, the accuracy and precision of uncertainty models, and the frequency and magnitude of order relation deviations.

2. Methodology

Consider the problem of estimating the value of an attribute z at an unsampled location $\mathbf{u}$. The information