Published in final edited form as:
Comput Geosci. 2009 June ; 35(6): 1255–1270.

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

P. Goovaerts[1]

[1] BioMedware, 516 North State Street, Ann Arbor, MI 48104, USA. email: goovaerts@biomedware.com, phone: 734-913-1098, fax: 734-913-2201

Abstract

Indicator kriging 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.

Keywords

interpolation; thresholds; cross-validation; jack-knife; accuracy; Fortran

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

Corresponding Author: Pierre Goovaerts, Biomedware, Inc. 516 North State Street, Ann Arbor, MI 48104, USA, Email: goovaerts@biomedware.com, Tel: (734) 913-1098, Fax: (734) 913-2201.

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 is to discretize the range of variation of the environmental attribute by a set of thresholds (e.g. deciles of sample histogram, detection limit, 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. Barabas 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, that is 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 Fergusson (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, that is to use as many thresholds as observations. In this case, typically only the 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 won't 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, 2008), 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 **u**. The information available consists of a set of *n* observations $z(\mathbf{u}_\alpha)$ that display some degree of spatial correlation. In geostatistics, the unmonitored value $z(\mathbf{u})$ is interpreted as a realization of a random variable $Z(\mathbf{u})$ which is fully characterized by the probability distribution $F(\mathbf{u};z) = \text{Prob}\{Z(\mathbf{u}) \leq z\}$. Indicator kriging does not provide a direct estimate of the unknown attribute value; rather it yields a set of *K* probability estimates:

$$\begin{aligned} \widehat{i}(\mathbf{u};z_k) &= F_{IK}(\mathbf{u};z_k|(n)) \\ &= \text{Prob}\{Z(\mathbf{u}) \leq z_k|(n)\} \quad k=1,\ldots,K \end{aligned} \tag{1}$$

where z_k are *K* thresholds discretizing the range of variation of the attribute *z* (e.g. 9 deciles).

2.1 Indicator kriging

Ccdf values are estimated by applying kriging to indicator transforms of the data. Although both simple and ordinary kriging are implemented in the computer code, the following presentation is restricted to the most commonly used ordinary kriging estimator:

$$F_{IK}(\mathbf{u};z_k|(n)) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha k}\, i(\mathbf{u}_\alpha;z_k) \quad k=1,\ldots,K \tag{2}$$

which is based on a preliminary coding of each observation $z(\mathbf{u}_\alpha)$ into a vector of indicators of non-exceedence of the threshold values:

$$i(\mathbf{u}_\alpha;z_k)=\begin{cases}1 & \text{if } z(\mathbf{u}_\alpha)\le z_k\\ 0 & \text{otherwise}\end{cases}\qquad k=1,\ldots,K \tag{3}$$

The weights in equation (2) are the solution of the following system of (n(**u**)+1) linear equations, known as ordinary indicator kriging system:

$$\sum_{\beta=1}^{n(\mathbf{u})}\lambda_{\beta k}\gamma_I(\mathbf{u}_\alpha-\mathbf{u}_\beta;z_k)-\mu_k=\gamma_I(\mathbf{u}_\alpha-\mathbf{u};z_k)\quad \alpha=1,\ldots,n(\mathbf{u})$$
$$\sum_{\beta=1}^{n(\mathbf{u})}\lambda_{\beta k}=1 \tag{4}$$

where μ_k is a Lagrange multiplier accounting for the constraint on the weights. The weights $\lambda_{\alpha k}$ account for the data configuration (i.e. clustering of observations), the proximity of data to the unsampled location **u**, as well as the spatial pattern of indicator data modelled from the experimental semivariogram:

$$\widehat{\gamma}_I(\mathbf{h};z_k)=\frac{1}{2N(h)}\sum_{\alpha=1}^{N(h)}[\,i(\mathbf{u}_\alpha;z_k)-i(\mathbf{u}_\alpha+\mathbf{h};z_k)]^2 \tag{5}$$

The indicator variogram $2\hat{\gamma}_I(\mathbf{h};z_k)$ measures how often two *z*-values a vector **h** apart are on the opposite side of the threshold value z_k. Therefore, it quantifies the lack of spatial connectivity of the values exceeding z_k. In most situations, the spatial connectivity of low and high values will differ, hence the need to model a semivariogram and solve a kriging system for each threshold.

2.2 Post-processing the discrete probability distributions

Because the *K* probabilities are estimated individually (i.e. *K* indicator kriging systems are solved at each location) the following constraints, which are implicit to any probability distribution, might not be satisfied by all sets of *K* estimates:

$$0\le F_{IK}(\mathbf{u};z_k|(n))\le 1\quad \forall k \tag{6}$$

$$F_{IK}(\mathbf{u};z_{k'}|(n))\le F_{IK}(\mathbf{u};z_k|(n))\quad \text{if } z_{k'}\le z_k \tag{7}$$

All probabilities that are not between 0 and 1 are first reset to the closest bound, 0 or 1. Then, condition (7) is ensured by averaging the results of an upward and downward correction of ccdf values (Deutsch and Journel, 1998).

Common estimators of the unknown *z*-value are the mean or the median of the ccdf, whereas the uncertainty is measured by the spread of the ccdf. Here the mean (E-type estimate) and variance of the ccdf were selected as predictor and associated measure of uncertainty. In the program, these two quantities are estimated as follows:

$$\widehat{z}_{IK}(\mathbf{u})=\frac{1}{100}\sum_{j=1}^{100} z_p(\mathbf{u}) \text{ with } p=0.01\times(j-0.5) \tag{8}$$

$$\widehat{\sigma}^2_{IK}(\mathbf{u})=\frac{1}{100}\sum_{j=1}^{100}\left[z_p(\mathbf{u})-\widehat{z}_{IK}(\mathbf{u})\right]^2 \text{ with } p=0.01\times(j-0.5) \tag{9}$$

The computation of the 100 percentiles of each ccdf, $z_p(\mathbf{u})$, requires the interpolation of the set of K probabilities within each class $(z_k,z_{k+1}]$ and extrapolation beyond the smallest and the largest thresholds to build a continuous model for the conditional cdf. One popular choice is the linear interpolation within each class $(z_k,z_{k+1}]$, which amounts to assuming that all values between z_k and z_{k+1} are equally likely to be observed. On the other hand, power and hyperbolic models are typically applied to the tails of the distribution, which requires the subjective choice of a power parameter. Instead of adopting an analytical function for ccdf interpolation and extrapolation, it seems less arbitrary to borrow information from the sample histogram whenever the number of observations exceeds the number of thresholds. Therefore, the resolution of the discrete ccdf in the computer code AUTO-IK is increased by performing a linear interpolation between tabulated bounds provided by the sample histogram (Deutsch and Journel, 1998).

2.3 Validation of the prediction models

One might want to assess the sensitivity of results to implementation variants, such as the number of thresholds or the use of anisotropic versus isotropic semivariogram models. Such questions can be answered by comparing the interpolation results with observations that have been either temporarily removed one at a time (cross-validation or leave-one-out approach) or set aside for the whole analysis (jack-knife). Note that these definitions are swapped in the statistical literature. The first two performance criteria are the mean error (ME) and mean absolute error (MAE) of prediction computed as:

$$ME=\frac{1}{N}\sum_{\alpha=1}^{N}\left[\widehat{z}_{IK}(\mathbf{u}_\alpha)-z(\mathbf{u}_\alpha)\right] \tag{10}$$

$$MAE=\frac{1}{N}\sum_{\alpha=1}^{N}\left|\widehat{z}_{IK}(\mathbf{u}_\alpha)-z(\mathbf{u}_\alpha)\right| \tag{11}$$

The ability of the prediction variance to capture the actual magnitude of the prediction error is quantified using the following mean square standardized residual (MSSR):

$$MSSR=\frac{1}{N}\sum_{\alpha=1}^{N}\left[\frac{\widehat{z}_{IK}(\mathbf{u}_\alpha)-z(\mathbf{u}_\alpha)}{\widehat{\sigma}_{IK}(\mathbf{u}_\alpha)}\right]^2 \tag{12}$$

If the actual estimation error is equal, on average, to the error predicted by the model, the MSSR statistic should be about one (Wackernagel, 1998 p. 91).

The prediction variance is just a summary statistic that does not fully capture the uncertainty attached to the unknown z-value. From the ccdf one can compute a series of symmetric p-probability intervals (PI) bounded by the $(1-p)/2$ and $(1+p)/2$ quantiles of that distribution. For example, the 0.5-PI is bounded by the lower and upper quartiles (i.e. inter-quartile range). According to this model of local uncertainty, there is then a 0.5-probability that the actual attribute value falls into the 0.5-PI or, equivalently, that over the study area 50% of the 0.5-PI include the true z-value. Cross-validation or jack-knife yields a set of z-measurements and independently derived ccdfs at the N locations $\boldsymbol{u}_\alpha$, allowing the fraction of true values falling into the symmetric p-PI to be computed. Following Deutsch (1997), the agreement between observed, p_k^*, and expected fractions, p_k, is quantified using the following "goodness" statistic:

$$G=1-\frac{1}{K'}\sum_{k=1}^{K'} w_k|p_k^*-p_k| \quad \text{with } 0\le G\le 1 \tag{13}$$

where $w_k=1$ if $p_k^*>p_k$, and 2 otherwise. K' represents the discretization level of the computation. Twice more importance is given to deviations when $p_k^*<p_k$ (inaccurate case). The weights penalize less the accurate case, which is the case where the fraction of true values falling into the p-probability interval is larger than expected. The goodness statistic is completed by the so-called "accuracy plot" that allows one to visualize departures between observed and expected fractions as a function of the probability p; see Figure 1A.

Not only should the true attribute value fall into the PI according to the expected probability p, but this interval should be as narrow as possible to reduce the uncertainty about that value. The average width of these local PIs should also be smaller than the global PI inferred from the sample histogram. Following Goovaerts et al. (2008) the average width of the PIs that include the true value are computed for a series of probabilities p and plotted versus the corresponding p-PI inferred from the global histogram; see Figure 1B. The ratio of local PI width versus global PI width, named standardized width, is averaged over a series of probability p and should be as small as possible.

3. Program description

The different steps of indicator kriging were implemented in a single ANSI Fortran-77 code, named AUTO-IK, which proceeds as follows:

1. Data are imported, and the K thresholds are either specified by the user or computed as equally spaced p-quantiles of the sample histogram.
2. Each observation is coded as a vector of K indicators, and the corresponding K indicator semivariograms are estimated and modeled using weighted least-square regression; model parameters are outputted by the program and figures of the experimental semivariograms with the model fitted are automatically created.
3. Indicator kriging is performed for each threshold using four types of destination geography: 1) grid of points specified by the user, 2) rectangular grid, 3) sampled locations (cross-validation option), and 4) set of test locations (jack-knife option).
4. The set of probabilities is corrected for order relation deviations, and the resulting discrete ccdf is completed by interpolation/extrapolation.
5. The ccdf mean (E-type estimate) and variance are computed at each location.
6. For both cross-validation and jack-knife options, the true values are used to assess the goodness and precision of the model of uncertainty.

The source code was built around the Gslib programs GAMV, KB2D, IK3D, POSTIK, VARGPLT, VMODEL and the semivariogram modelling program VARFIT (Pardo-Igúzquiza, 1999). When running the executable the user needs to specify the name of a parameter file that includes all the parameters and names of input/output files required by the program. A typical parameter file, which was used to analyze cobalt concentrations in the Jura dataset (Atteia et al., 1994), is illustrated in Figure 2. The text file, called **AUTO-IK.par**, includes the following information:

① Name of the text file including the data. This dataset must be in Geo-EAS format (Englund and Sparks, 1988). An example for the file **Jura_sample.dat** is given in Figure 3. The first line is the name of the data file. The second line should be a numerical value specifying the number of variables (i.e. *nvar* columns) in the data file. The next *nvar* lines contain the name of each variable. The following lines, until the end of the file, are considered as observations and must have *nvar* numerical values per line. For example, the cobalt concentration recorded for the 1st soil sample, with spatial coordinates X= 2.386 km and Y= 3.077 km, is 9.320 ppm.

② The column numbers for the X and Y coordinates, and the variable to be kriged.

③ Code for missing value. All observations equal to that value are ignored in the analysis.

④ Four options available for indicator kriging:

1. Estimation at the nodes of a grid that has been created by the user.
2. Estimation at the nodes of a rectangular grid specified by the user later in the parameter file.
3. Estimation at each sampled location after removal of that particular observation (cross-validation approach).
4. Estimation at test locations where observations, which have not been used in the analysis, are available (jack-knife approach).

⑤ Name of the file (Geo-EAS format) with the interpolation grid (option 1) or the data used for jack-knifing (option 4).

⑥ The column numbers for the X and Y coordinates of the interpolation grid or data used for jack-knifing. The column number for the test data is used for option 4.

⑦ Definition of the rectangular grid (x axis): number of nodes, starting coordinate, and grid spacing.

⑧ Definition of the rectangular grid (y axis): number of nodes, starting coordinate, and grid spacing.

⑨ The number of thresholds used for the indicator coding of observations.

⑩ Two options available for defining threshold values: 0 = automatic computation of thresholds as equally spaced *p*-quantiles of the sample histogram, 1= threshold values are specified by the user.

⑪ Threshold values specified by the user for the indicator coding.

⑫ Options available for indicator kriging: full IK where threshold-specific semivariograms are used for the estimation, and median IK where the median semivariogram model is used for all thresholds.

⑬ Type of kriging algorithm: simple IK where the indicator means are automatically identified with the arithmetical averages of indicator transforms, and ordinary IK where the indicator means are implicitly estimated within each search window.

⑭ Number and width of classes of distance used for the computation of the semivariogram; 20 classes of 0.1 km are used in this example.

⑮ Number of directions for the computation of the semivariogram. Options are *ndir*=1 (omnidirectional) and *ndir*=4. In the later case, the semivariogram is computed in four directions (angular tolerance = ±22.5°), starting with the azimuth direction specified by the user. Using Gslib convention, angles are measured in degrees clockwise from the NS direction.

⑯ Weighting scheme used in the least-square fitting of a semivariogram model to experimental values. The program will try all possible combinations of 1 or 2 basic models among the spherical and exponential functions. The selected model is the one that minimizes the weighted sum of squares of differences between the experimental and model curves:

$$WSS=\sum_{l=1}^{L} w(\mathbf{h}_l)[\widehat{\gamma}(\mathbf{h}_l) - \gamma(\mathbf{h}_l)]^2$$

where L is the number of classes of distance. The user can choose among the five following types of weighting schemes: $w(\mathbf{h}_l)=1$, $w(\mathbf{h}_1)=\sqrt{N(\mathbf{h}_l)}/\gamma(\mathbf{h}_1)$, $w(\mathbf{h}_l)=1/\gamma(\mathbf{h}_1)^2$, $w(\mathbf{h}_l)=N(\mathbf{h}_l)$, $w(\mathbf{h}_l)=N(\mathbf{h}_l)/\log|\mathbf{h}_l|$. Except for the first option, each alternative set of weights aims to assign more importance to: semivariogram values computed from many data pairs (hence more reliable), and/or smaller semivariogram values that are typically observed for short distances since the behavior of the semivariogram at the origin has the largest impact on kriging results.

⑰ Minimum and maximum number of neighboring observation to be used in the estimation, and the radius of the circular search window. Missing estimated ccdf values (coded -9) and corresponding statistics (coded -999) are reported when the minimum number of observations is not reached.

⑱ Name of output text file reporting the experimental semivariogram values and the parameters (i.e. type of basic model, nugget effect, sill, range, anisotropy angle) of the model fitted, plus information on order relation deviations. For jack-knife and cross-validation options, this file also includes statistics on the prediction errors and results used to create the accuracy and PI-width plots. An example for cobalt data, file **Co-variog.txt**, is given in Figure 4.

⑲ Name of output text file (Geo-EAS format) that includes the X and Y coordinates of each estimated location, and the estimated ccdf values for all thresholds. An example for cobalt data, file **Co-IK.out**, is given in Figure 5.

⑳ Name of output text file (Geo-EAS format) that includes the X and Y coordinates of each estimated location, and the mean and variance of the local ccdf. For jack-knife and cross-validation options, this file also includes the test observations that were not used in the analysis, and the prediction error. An example for cobalt data, file **Co-stat.out**, is given in Figure 6.

In addition to text files with indicator kriging results, the program AUTO-IK generates graphs that display: 1) all the experimental indicator semivariogram values and the model fitted (individual plot + combined plot of up to 12 semivariograms), and 2) the accuracy and PI-width plots for jack-knife and cross-validation options. These figures are in PostScript format and can be viewed using the public-domain program GSview (http://www.cs.wisc.edu/~ghost/gsview). These graphs should help detecting any poor choice

of the number and width of classes of distance, as well as poor fits of semivariogram models. In the later case, the user should select other options for the weighting scheme. Figures 7 and 8 show the 19 indicator semivariogram plots created for cobalt using the parameter file of Figure 2. These semivariograms were rescaled by the variance of the indicator variable to facilitate comparison across thresholds. The corresponding accuracy and PI-width plots are displayed in Figure 1.

4. Case-study

The functionalities of the program are illustrated using the 259 cobalt concentrations displayed in Figure 9A. Omnidirectional indicator semivariograms were computed for 19 thresholds identified with equally-spaced *p*-quantiles of the sample histogram (e.g. p=0.05, 0.10, 0.15). Isotropic models were automatically fitted by AUTO-IK (Figures 7–8). Anisotropic modeling was also undertaken but the azimuth appeared to have only a minor impact on the spatial connectivity of cobalt indicators (Figure 10). For the first couple of thresholds the indicator semivariograms display a very large nugget effect and a short-range structure (less than 200 meters), whereas larger ranges and smaller nugget effects occur for the majority of other thresholds. This pattern reflects the existence of small clusters of low concentrations (i.e. 10% smallest values), which are mainly located on Argovian rocks (Figure 9B).

Ordinary indicator kriging with 19 thresholds is used to model the ccdf and derive the associated statistics at the nodes of a 50 m non-rectangular grid constrained to the boundaries of the study area (kriging option 1), as well as at 100 test locations (jack-knife, kriging option 4). The map of E-type estimates (Figure 9C) shows lower cobalt concentration on Argovian rocks. According to the variance map (Figure 9E), the uncertainty is smaller on Argovian rocks where concentrations are consistently small, which reflects the impact of both the proportional effect (i.e. lower variances are associated with lower means) and spatial homogeneity. Larger ccdf variances are observed in sparsely sampled areas (e.g. Western edge of the study area), as well as where high and low concentrations are intermingled, like in the central part of the study area. These zones of spatial heterogeneity can also be detected by mapping the variance of observations within moving windows (Figure 9F). The ccdfs modeled at 100 test locations (Figure 9D) were used to assess the quality of the model of uncertainty through the creation of the accuracy and PI-width plots displayed in Figure 11. Relatively to the results obtained using cross-validation (Figure 1), IK-based models appear to less precise (standardized PI-width is 50% larger) and slightly less accurate according to the goodness statistic.

A sensitivity analysis was conducted to assess the impact of the following factors on the prediction performances:

- number of thresholds. The cross-validation was conducted by increasing gradually the number of thresholds from 5 to 100.
- type of ccdf interpolation model: 1) crude linear interpolation between thresholds which is the option used in most published studies and implemented in SGEMS, and 2) linear interpolation between tabulated bounds provided by the sample histogram as available in Gslib and AUTO-IK.
- Semivariogram model. Full indicator kriging, which requires the fitting of a separate model for each threshold, is compared to median indicator kriging where the same model is used across all thresholds.

The generality of the findings was investigated by repeating the analysis for three heavy metals in the Jura dataset: cobalt and two heavy metals (Cd, Zn) with asymmetric histograms and reasonable spatial autocorrelation (Figure 12).

As expected, increasing the number of thresholds attenuates the impact of the type of ccdf interpolation model on the prediction performances. For the two heavy metals with positively skewed distributions, the largest prediction errors are observed when performing a simple linear interpolation between fewer than 20 thresholds (Figures 13A–B). There is no substantial benefit in using more thresholds for cobalt that displays smaller relative prediction errors (16%) than cadmium (40%) or zinc (20%). Full and median indicator kriging yield very similar results, which suggests that the benefit of using threshold-specific semivariogram models is canceled out by the impact of more and greater order relation deviations, in particular as the number of thresholds increases (Figure 14). Regardless the type of indicator kriging, the largest proportion of order relation deviations is observed for cobalt, the metal with the strongest spatial structure (i.e. longest semivariogram range), leading to a higher likelihood of negative kriging weights caused by screening effect (Goovaerts, 1997). For example, median indicator kriging with the semivariogram models of Figure 12 (right column) generates 35% of negative kriging weights for cobalt, whereas this percentage is 0.02% for cadmium and 7.9% for zinc.

As for the prediction errors, best results for the MSSR statistic (i.e. values closer to 1) are recorded when the number of thresholds exceeds 20, especially for Zn and Co using the crude linear interpolation model (Figures 13C–D). According to this criterion, median indicator kriging outperforms full indicator kriging for cobalt.

Unlike previous criteria, the goodness statistic does not improve as more thresholds are used (Figures 13E–F), although the decline is small in absolute terms. The slight increase in the width of the probability intervals (Figures 13G–H) is driven mainly by wider intervals for low probabilities (i.e. $p < 0.2$); see examples in Figures 1B and 11B. Overall the best results (i.e. more accurate and precise uncertainty models) are recorded for cobalt, which is expected since this metal displays the strongest spatial correlation.

5. Conclusions

This study demonstrated that nowadays indicator kriging with multiple thresholds is accessible and computationally tractable thanks to the development of automatic procedures for semivariogram fitting and the growing processing speed. In particular, the user should no longer feel restricted in using a maximum of nine thresholds for indicator coding, as typically done in many analyses. Using more thresholds attenuates the loss of information caused by the coding of continuous attributes into a set of binary indicators, which is a common criticism of the technique. Indicator kriging seems a better alternative to data transforms, since: (1) the transform procedure might be hard to automate (except for normal score transform, there is no guaranty that the transformed distribution will meet the requirements of the analysis) and (2) one-to-one transforms, such as normal score transform, require an arbitrary untie of censored data. This code should foster the application of indicator kriging in earth sciences and facilitates its optimal implementation by accounting for class-specific anisotropic patterns of spatial correlation, while allowing a quick assessment of the accuracy and precision of the non-parametric model of local uncertainty. This software should, however, not be used as a black-box and the reader is advised to analyze the semivariogram plots and cross-validation statistics to diagnose any potential problem in the application of the method.

According to the sensitivity analysis, for the dataset analyzed in this paper there appears to be little benefit in considering more than nine thresholds if the interpolation/extrapolation of discrete ccdf values is conducted using linear interpolation between tabulated bounds provided by the sample histogram. The pattern of spatial variability (i.e. range of indicator semivariogram model) seems to matter more than the use of threshold-specific semivariograms or the tenfold increase in the number of thresholds. The generalization of these preliminary conclusions will require repeating the analysis for other datasets with potentially larger

differences in the spatial connectivity of small and large values. The benefit of using locally adaptive thresholds should also be explored.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by grant R44-CA132347-01 from the National Cancer Institute. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI.

References

Atteia O, Dubois JP, Webster R. Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution 1994;86:315–327. [PubMed: 15091623]

Barabás N, Goovaerts P, Adriaens P. Geostatistical assessment and validation of uncertainty for three-dimensional dioxin data from sediments in an estuarine river. Environmental Science & Technology 2001;35(16):3294–3301. [PubMed: 11529567]

Cattle JA, McBratney AB, Minasny B. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. Journal of Environmental Quality 2002;31:1576–1588. [PubMed: 12371175]

Chu J. Fast sequential indicator simulation: beyond reproduction of indicator variograms. Mathematical Geology 1996;28(7):923–936.

Deutsch, CV. Direct assessment of local accuracy and precision. In: Baafi, EY.; Schofield, NA., editors. Geostatistics Wollongong ‘96. Kluwer Academic Publishers; Dordrecht: 1997. p. 115-125.

Deutsch, CV.; Lewis, R. Advances in the practical implementation of indicator geostatistics. Proceedings of the 23rd International APCOM Symposium; Tucson, AZ, Society of Mining Engineers. 1992. p. 169-179.

Deutsch, CV.; Journel, AG. GSLIB: Geostatistical Software Library and User’s Guide. 2. Oxford University Press; New York, NY: 1998. p. 369

Englund, E.; Sparks, A. Geo-EAS 1.2.1 User’s Guide. EPA Report # 60018-91/008. EPA-EMSL; Las Vegas, NV: 1988.

Goovaerts P. Comparative performance of indicator algorithms for modeling conditional probability distribution functions. Mathematical Geology 1994;26(3):389–411.

Goovaerts, P. Geostatistics for Natural Resources Evaluation. Oxford University Press; New York, NY: 1997. p. 483

Goovaerts P. Geostatistical modelling of uncertainty in soil science. Geoderma 2001;103:3–26.

Goovaerts P, AvRuskin G, Meliker J, Slotnick M, Jacquez GM, Nriagu J. Geostatistical modeling of the spatial variability of arsenic in groundwater of Southeast Michigan. Water Resources Research 2005;41(7):W07013 10.1029.

Goovaerts P, Trinh HT, Demond AH, Towey T, Chang S-C, Gwinn D, Hong B, Garabrant D, Adriaens P. Geostatistical modeling of the spatial distribution of soil dioxin in the vicinity of an incinerator. 2. Verification and calibration study. Environmental Science & Technology 2008;42(10):3655–3661. [PubMed: 18546704]

Journel AG. Nonparametric estimation of spatial distributions. Mathematical Geology 1983;15(3):445–468.

Lark RM, Ferguson RB. Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging. Geoderma 2004;118(1):39–53.

Lloyd CD, Atkinson PM. Assessing uncertainty in estimates with ordinary and indicator kriging. Computers and Geosciences 2001;27(8):929–937.

Pardo-Igúzquiza E. VARFIT: a Fortran-77 program for fitting variogram models by weighted least squares. Computers and Geosciences 1999;25(3):251–261.

Pardo-Igúzquiza E, Dowd PA. Multiple indicator cokriging with application to optimal sampling for environmental monitoring. Computers and Geosciences 2005;31(1):1–13.

Remy, N.; Boucher, A.; Wu, J. Applied Geostatistics with SGEMS: A User's Guide. Cambridge University Press; 2008. in press

Saito H, Goovaerts P. Geostatistical interpolation of positively skewed and censored data in a dioxin contaminated site. Environmental Science & Technology 2000;34(19):4228–4235.

Wackernagel, H. Multivariate Geostatistics, 2nd completely revised edition. Berlin: Springer; 1998. p. 2561998

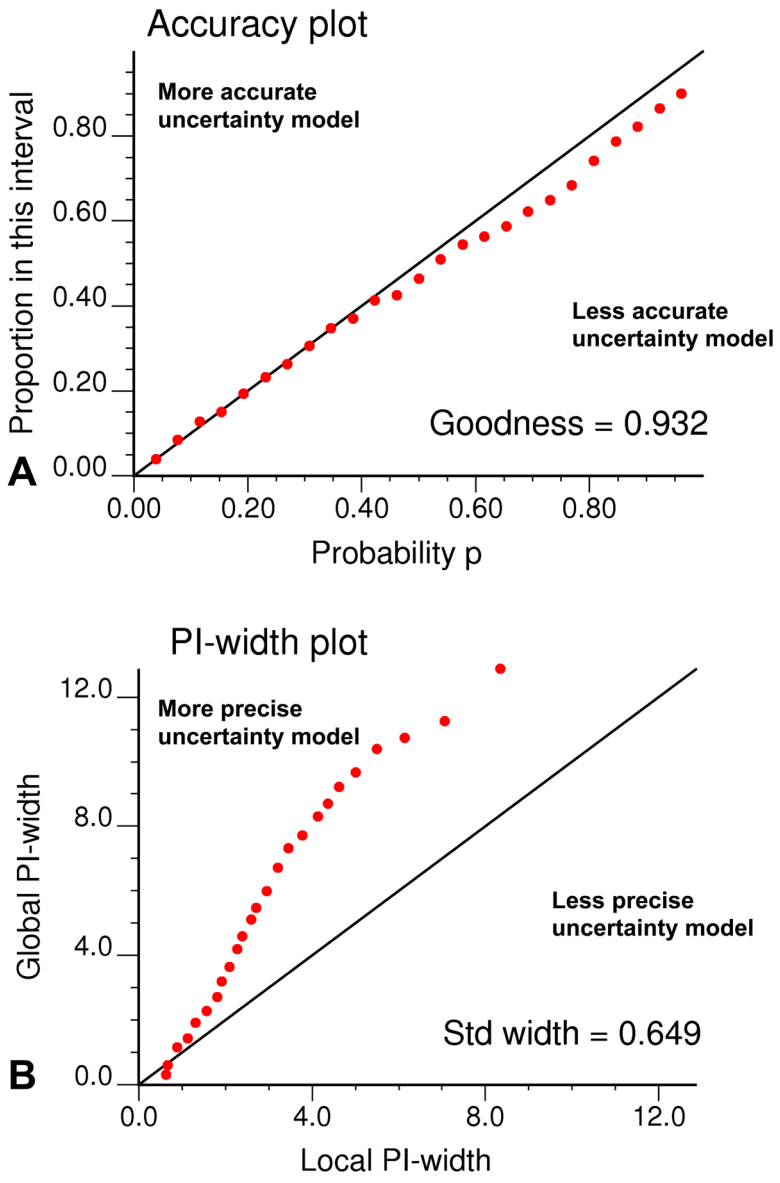

Figure 1.
Example of accuracy plot (A) and standardized PI-width plot (B) created automatically by the program AUTO-IK.exe when selecting the cross-validation option. The accuracy plot shows the good correspondence between expected and empirical proportions of Co data falling within probability intervals (PI) of increasing size, as measured by the goodness statistic (best if 1). The width of these local PIs are consistently smaller (average ratio=0.649) than the global PIs derived from the sample histogram (B).

```
                    Parameters for AUTO-IK
                    **********************

   START OF PARAMETERS:
①  Jura_sample.dat       -File with data
②  1 2 6                 -Column numbers for X & Y coord. + variable under study
③  -9999                 -Code for missing value
④  3                     -Options: 1=user grid, 2=regular grid, 3=Xvalidation, 4=jack-knife
⑤  Jura-50mgrid.dat      -File with user grid or jackknife data
⑥  1 2 6                 -Column numbers for X & Y node coord. + observations (jack-knife)
⑦  60   0    0.1         -nx,xmn,xsiz
⑧  60   0    0.1         -ny,ymn,ysiz
⑨  19                    -Number of thresholds for indicator kriging
⑩  0                     -Choice of thresholds (0=automatic computation,1=user's choice)
⑪  -9999                 -Values of thresholds if specified by the user
⑫  0                     -IK options: 0=full IK, 1=median IK
⑬  1                     -Kriging types: 0=simple kriging, 1=ordinary kriging
⑭  20 .1                 -Number of lags + lag spacing for variogram computation
⑮  1 22.5                -Number of directions (ndir=1 or 4) + 1st azimuth for ndir=4
⑯  2                     -Weights for semivariogram modeling
⑰  8 32 2.0              -Maximum number of observations & search radius
⑱  Co-variog.txt         -Output file for semivariogram values + models
⑲  Co-IK.out             -Output file for probability estimates(GEO-EAS format)
⑳  Co-stat.out           -Output file for Ccdf statistics(GEO-EAS format)

   Weights option for semivariogram modeling:
   1 => constant weight
   2 => weight = (Number of data pairs)^0.5/gamma
   3 => weight = 1/gamma^2
   4 => weight = Number of data pairs
   5 => weight = Number of data pairs/log(lag distance)
```

Figure 2.

Example of parameter file required by AUTO-IK.exe. This parameter file is used to conduct a geostatistical analysis of soil cobalt concentrations (cross-validation option). Indicator semivariograms for thresholds corresponding to 19 equally spaced *p*-quantiles of the sample histogram, are computed using 20 classes of 0.1 km. The models are fitted automatically and used to perform full ordinary indicator kriging using up to the 32 closest observations located within a radius of 2 km.

```
Jura.dat
11
Xloc
Yloc
Landuse
Rock
Cd
Co
Cr
Cu
Ni
Pb
Zn
2.386  3.077  3  3  1.740  9.320 38.320 25.720 21.320 77.360 92.560
2.544  1.972  2  2  1.335 10.000 40.200 24.760 29.720 77.880 73.560
2.807  3.347  2  3  1.610 10.600 47.000  8.880 21.400 30.800 64.800
4.308  1.933  3  2  2.150 11.920 43.520 22.700 29.720 56.400 90.000
4.383  1.081  3  5  1.565 16.320 38.520 34.320 26.200 66.400 88.400
3.244  4.519  3  5  1.145  3.508 40.400 31.280 22.040 72.400 75.200
3.925  3.785  3  5  0.894 15.080 30.520 27.440 21.760 60.000 72.400
```

Figure 3.
Example of dataset for AUTO-IK.exe. Data must be in Geo-EAS format. The first line is the name of the data file. The second line should be a numerical value specifying the number of variables (i.e. nvar columns) in the data file. The next nvar lines contain the name of each variable. The following lines, until the end of the file, are considered as observations and must have nvar numerical values per line.

```
Experimental semivariogram for threshold #:          19
 Lag #  Lag dist.  Semivariogram value  # data pairs
    1       0.02374        0.70644        193
    2       0.10444        0.60898        155
    3       0.20629        0.37908        249
    4       0.29742        0.96266        414
    5       0.39149        1.23771        644
    6       0.49575        0.96998        692
    7       0.60103        1.07219        538
    8       0.69989        0.63900        755
    9       0.80025        0.93888        916
   10       0.89570        0.78401        709
   11       1.00206        1.31803        931
   12       1.09891        1.10861       1315
   13       1.20139        0.81573        810
   14       1.29774        0.96403       1262
   15       1.39760        1.17480       1107
   16       1.49447        1.01549       1291
   17       1.59717        0.95956       1093
   18       1.70017        0.88279       1093
   19       1.79546        1.06411       1370
   20       1.89798        0.83169        971

 Semivariogram model for threshold value:   14.42596
 Nugget effect:    0.5530
 Number of basic models:   1
 Model 1:
 Type: spherical model
    Sill:    0.4448
    Max. range, Min. range:       0.4721      0.4721
    Azimuth for max. range:      90.0000

==================================

Order relation deviations:
---------------------------
Frequency over all thresholds:            0.651
Average magnitude over all thresholds:    .124E-01

Results of the accuracy plot:
---------------------------
Expected freq.   Observed freq.   PI-width   Std PI-width
    0.038           0.039           0.186       0.588
    0.077           0.085           0.240       0.401
    0.115           0.127           0.395       0.343
    0.154           0.151           1.009       0.702
    0.192           0.193           1.498       0.782
    0.231           0.232           2.384       1.050
    0.269           0.263           2.650       0.978
      :               :               :           :
      :               :               :           :
    0.846           0.788           9.416       0.906
    0.885           0.822          10.014       0.933
    0.923           0.865          10.266       0.912
    0.962           0.900          11.161       0.867

Prediction performances
-----------------------
Mean error:                          -0.054
Mean absolute error:                  1.513
Mean square standardized residual:    0.874
Goodness statistic:                   0.932
PI-width statistic:                   0.649
```

Figure 4.

Output file created by AUTO-IK.exe following the cross-validation analysis of cobalt concentrations. The text file reports for each threshold the experimental semivariogram values and the parameters (i.e. type of basic model, nugget effect, sill, range, anisotropy angle) of the model fitted, plus information on order relation deviations. For jack-knife and cross-validation options, this file also includes statistics on the prediction errors and results used to create the accuracy and PI-width plots. This file was obtained when running the code with the parameter file of Figure 2.

```
Co-IK.out
21
Xloc
Yloc
Probability_Indicator1
Probability_Indicator2
Probability_Indicator3
Probability_Indicator4
Probability_Indicator5
Probability_Indicator6
Probability_Indicator7
Probability_Indicator8
Probability_Indicator9
Probability_Indicator10
Probability_Indicator11
Probability_Indicator12
Probability_Indicator13
Probability_Indicator14
Probability_Indicator15
Probability_Indicator16
Probability_Indicator17
Probability_Indicator18
Probability_Indicator19
 2.38600 3.07700 0.0345  0.0677  0.0677  0.0677  0.0677  0.0677  0.0729  0.1909  0.3197  0.6415  …
 2.54400 1.97200 0.0162  0.0162  0.0162  0.0222  0.0448  0.0448  0.0491  0.0850  0.0867  0.1094  …
 2.80700 3.34700 0.0728  0.1684  0.2352  0.2641  0.2641  0.2809  0.2866  0.5244  0.7014  0.7014  …
 4.30800 1.93300 0.0231  0.0248  0.0248  0.0248  0.0248  0.0248  0.2077  0.3191  0.5052  0.5052  …
 4.38300 1.08100 0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0605  0.0605  0.0656  0.0656  …
 3.24400 4.51900 0.1238  0.1372  0.3900  0.4990  0.6351  0.8064  0.8064  0.9533  0.9753  0.9758  …
```

Figure 5.
Output file created by AUTO-IK.exe following the cross-validation analysis of cobalt concentrations. The text file (Geo-EAS format) includes the X and Y coordinates of each sampled location, and the estimated ccdf values for all 19 thresholds. This output file was obtained when running the code with the parameter file of Figure 2.

```
Co-stat.out
7
Xloc
Yloc
Co
Etype estimate
Conditional_variance
Prediction error
Abs(Prediction error)
2.38600     3.07700     9.32000     9.52869     4.70491     0.20869     0.20869
2.54400     1.97200    10.00000    11.88089     5.62748     1.88089     1.88089
2.80700     3.34700    10.60000     8.42308    11.87935    -2.17692     2.17692
4.30800     1.93300    11.92000    10.38675     7.58975    -1.53325     1.53325
4.38300     1.08100    16.32000    14.23361     6.30268    -2.08638     2.08638
```

Figure 6.

Output file created by AUTO-IK.exe following the cross-validation analysis of cobalt concentrations. The text file (Geo-EAS format) includes the X and Y coordinates of each sampled location, the cobalt concentration that was discarded for the analysis, the mean and variance of the local ccdf, and the prediction error. This output file was obtained when running the code with the parameter file of Figure 2.

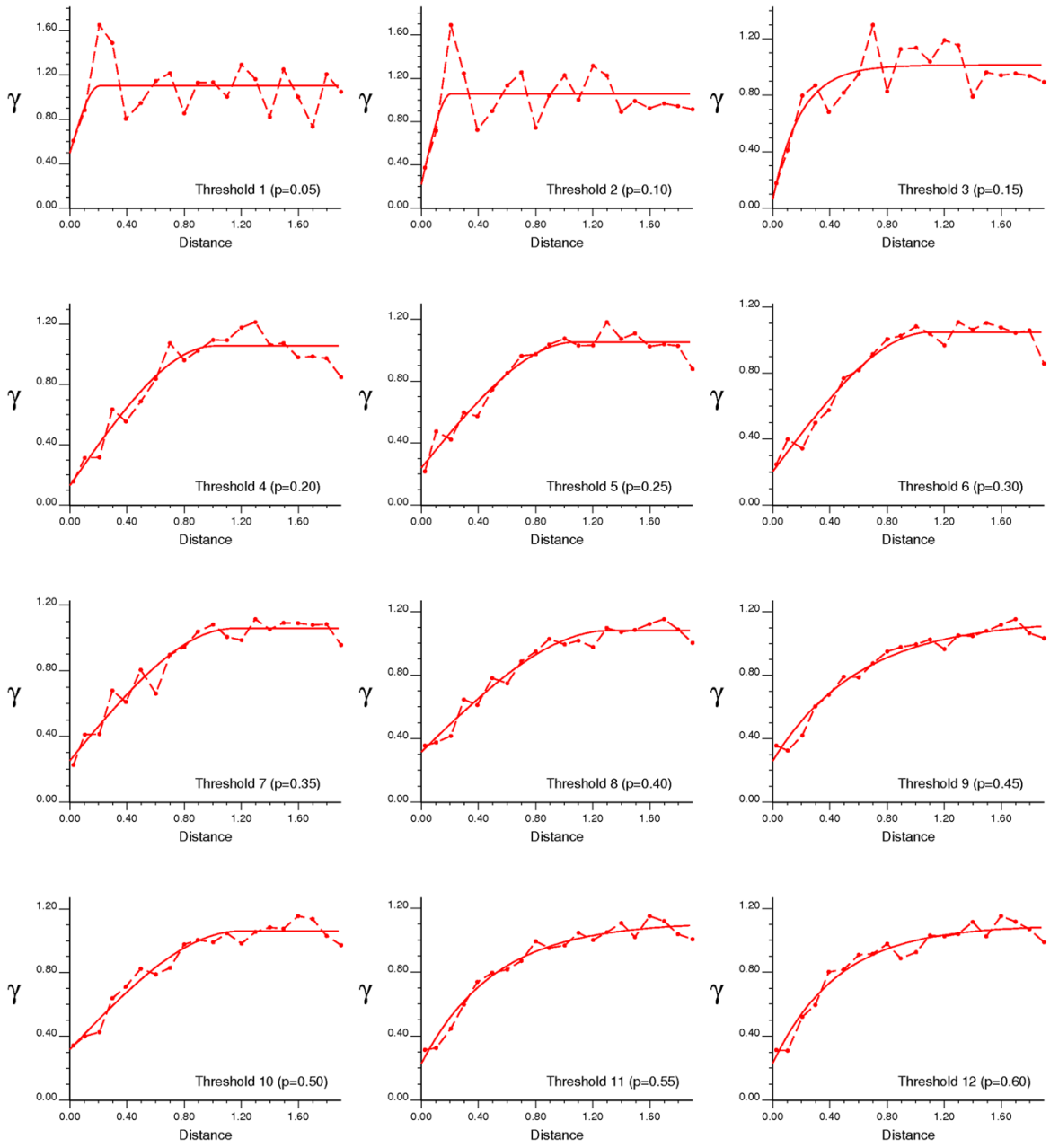

Figure 7.
Omnidirectional semivariograms computed for thresholds 1 through 12, with the model fitted using weighted-least square regression. These semivariograms were rescaled by the variance of the indicator variable. This graph is created automatically by AUTO-IK.

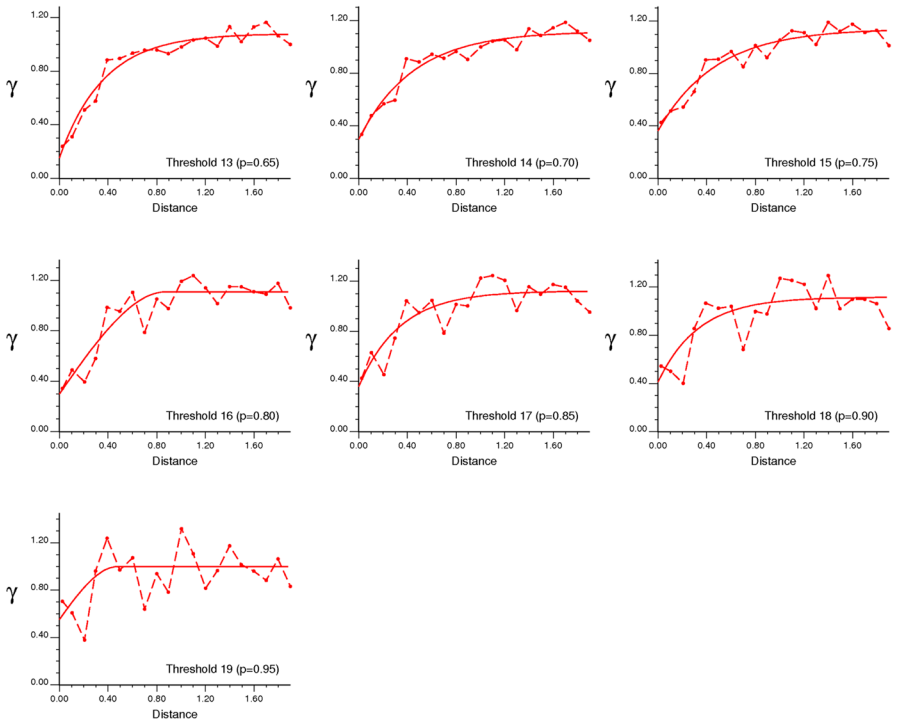

Figure 8.

Omnidirectional semivariograms computed for thresholds 13 through 19, with the model fitted using weighted-least square regression. These semivariograms were rescaled by the variance of the indicator variable. This graph is created automatically by AUTO-IK.

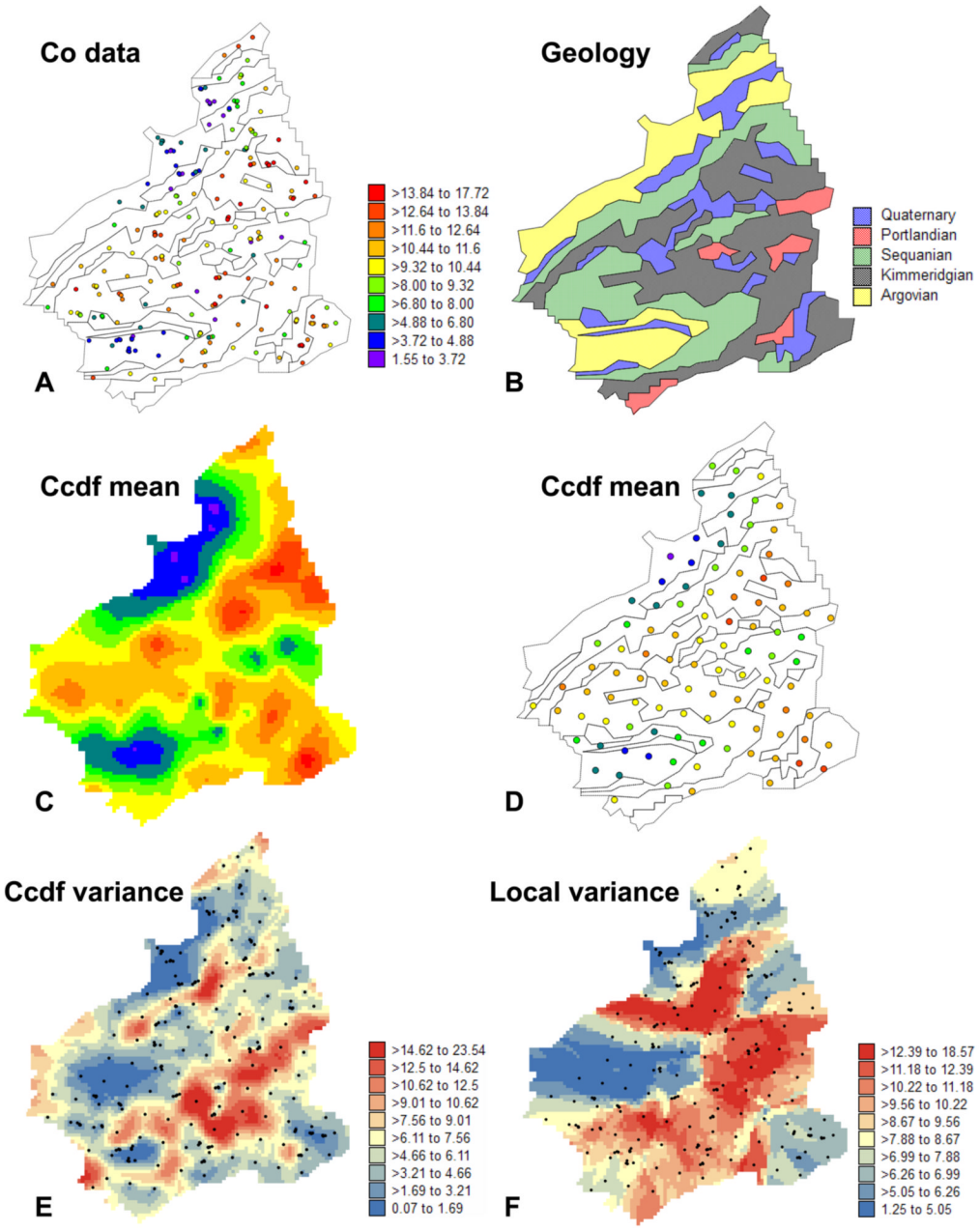

Figure 9.

(A) 259 observations available for estimating soil cobalt concentration at the nodes of a 50-m grid (C) or at 100 test locations (D) using ordinary indicator kriging. The map of the ccdf mean (C) shows lower cobalt concentrations on Argovian rocks (B). (E) The map of the ccdf variance indicates larger uncertainty where data, depicted by black dots, are scarce and/or where the local variance of the data (F) is large.

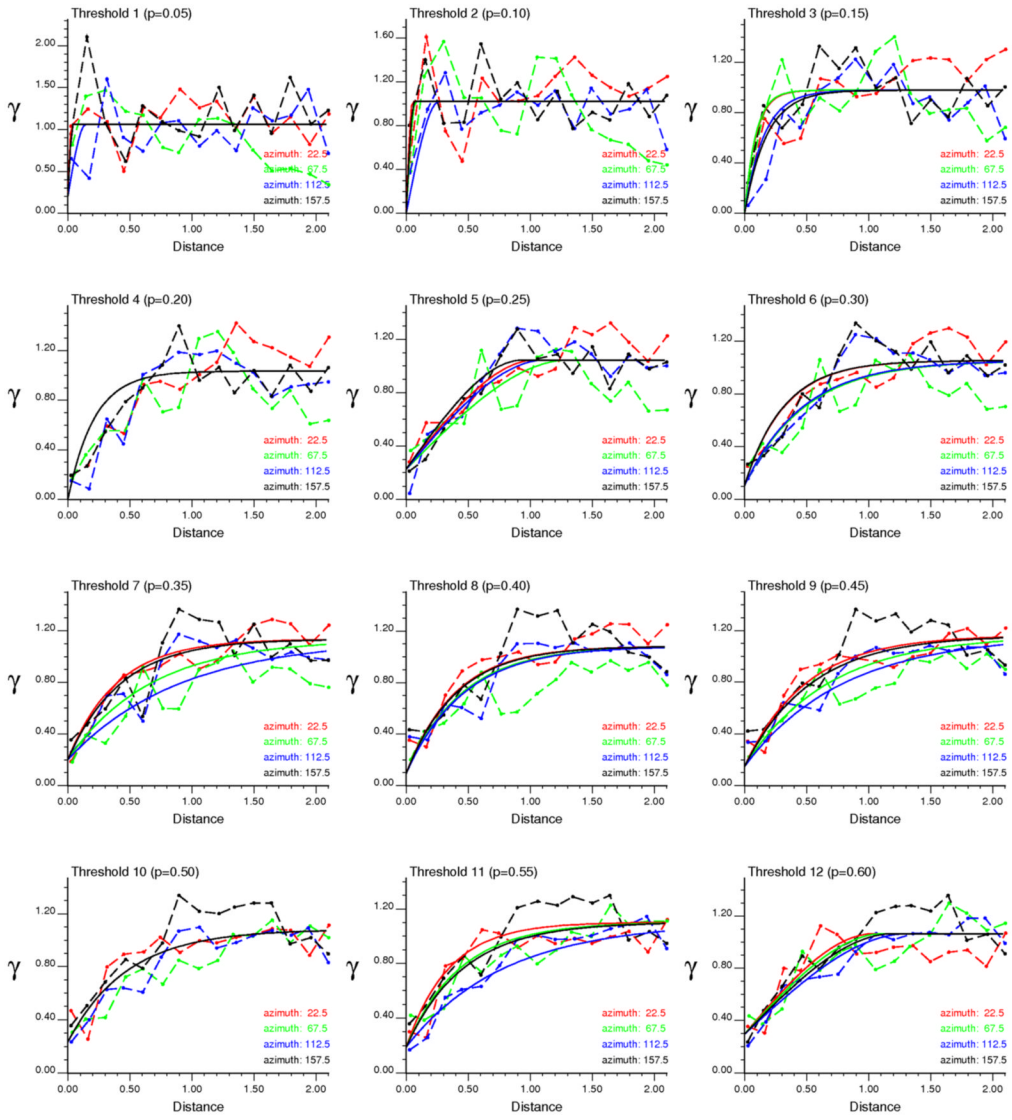

Figure 10.

Directional semivariograms computed for thresholds 1 through 12, with the anisotropic model fitted using weighted-least square regression. These semivariograms were rescaled by the variance of the indicator variable. This graph is created automatically by AUTO-IK.

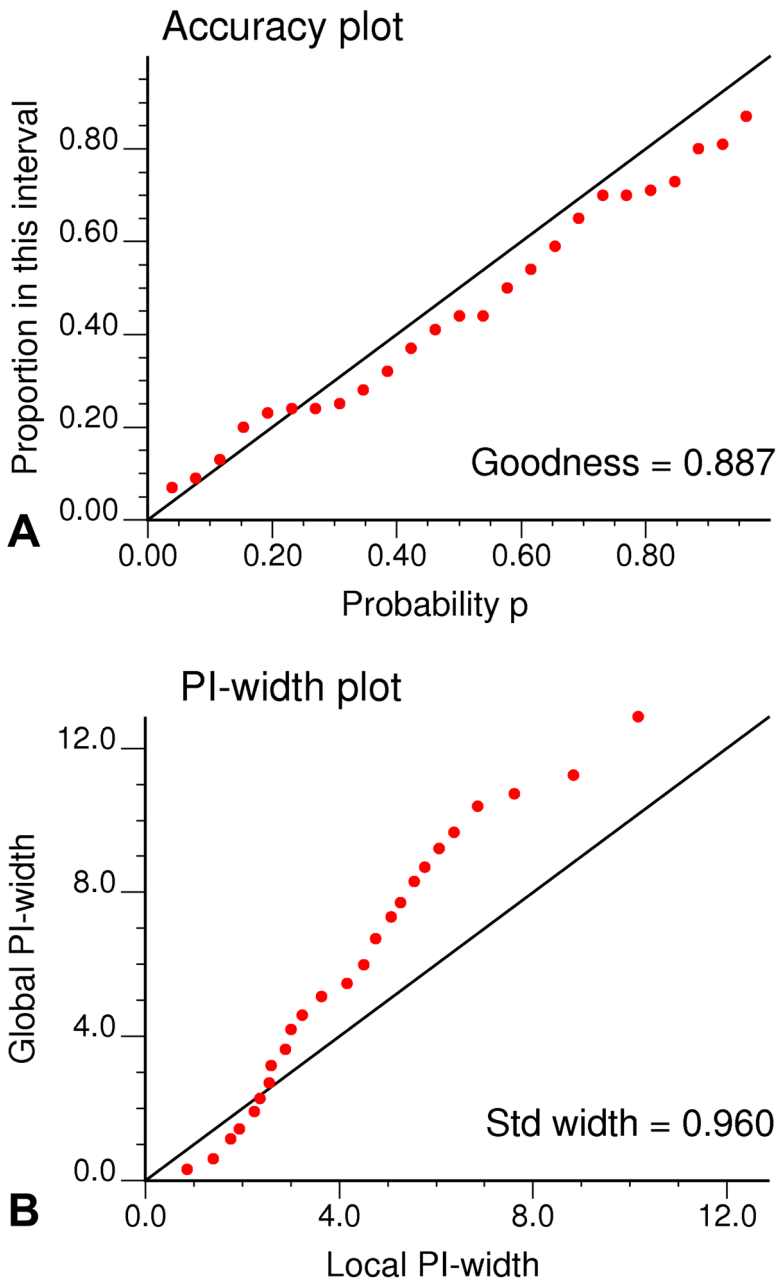

Figure 11.
Accuracy plot (A) and standardized PI-width plot (B) obtained by jack-knife using the 100 test locations displayed in Figure 9D. The accuracy plot shows the good correspondence between expected and empirical proportions of Co data falling within probability intervals (PI) of increasing size, as measured by the goodness statistic (best if 1). The width of these local PIs are on average smaller (average ratio=0.96) than the global PIs derived from the sample histogram (B).

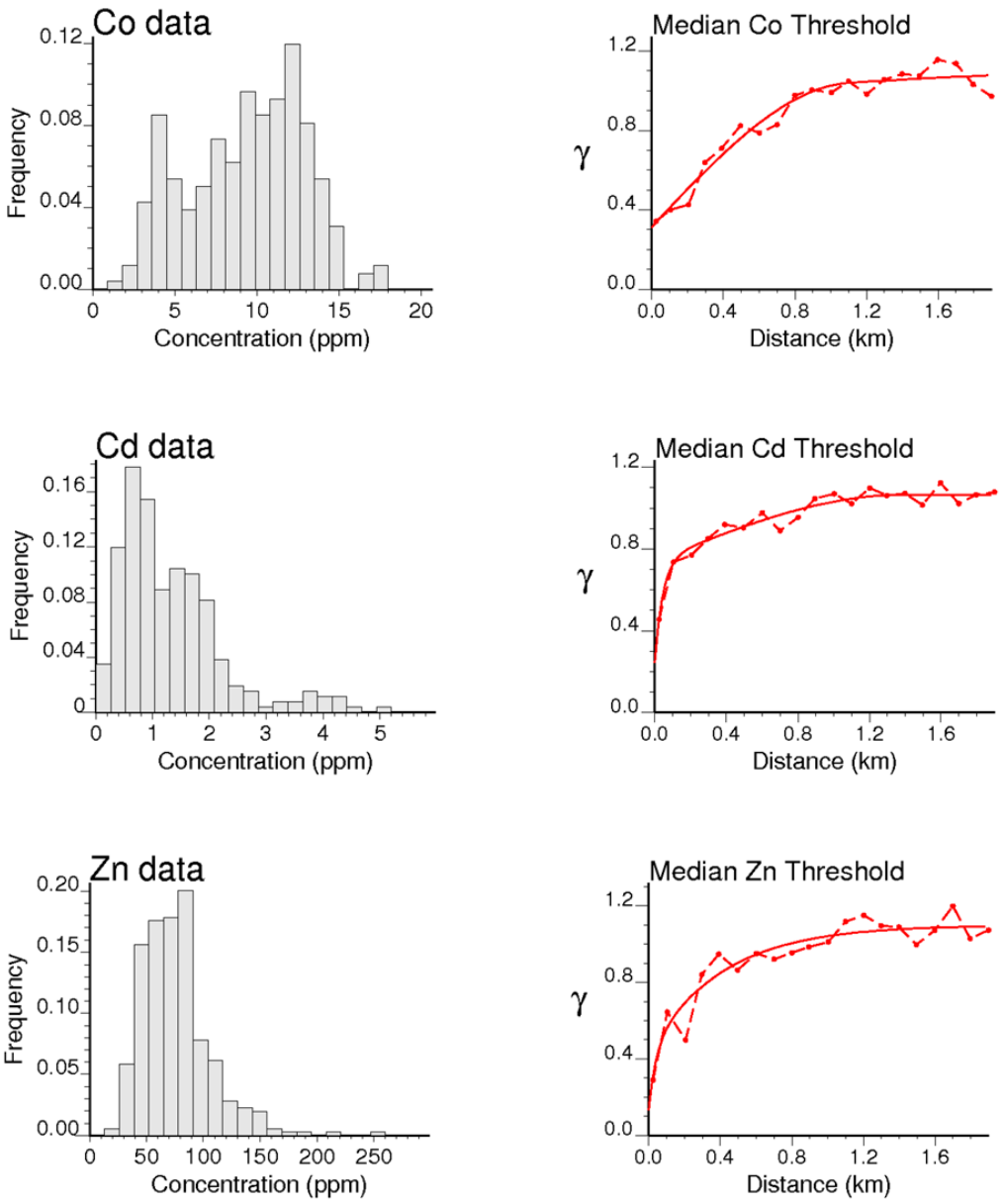

Figure 12.
Histograms and median indicator semivariograms for the three heavy metals used in the sensitivity analysis.

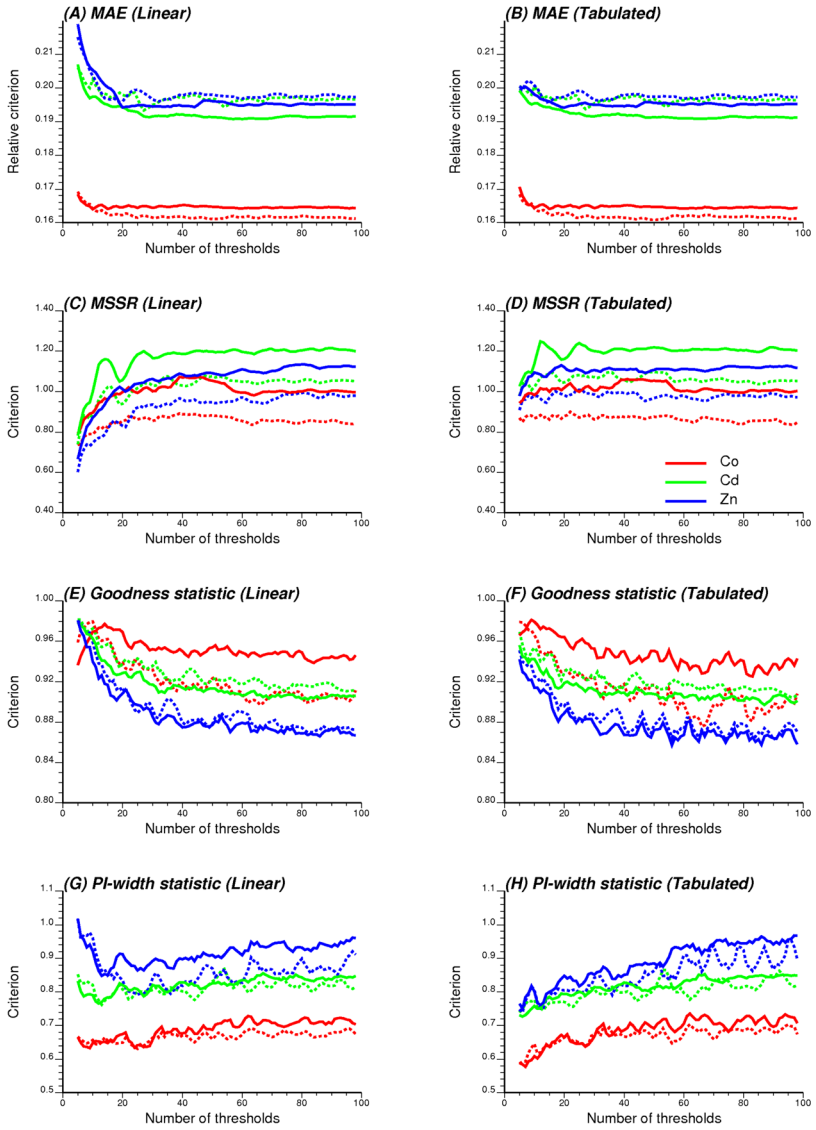

Figure 13.

Impact of the number of thresholds on the prediction accuracy and the goodness of the model of uncertainty obtained using indicator kriging and two types of ccdf interpolation models: crude linear interpolation between thresholds (left) and linear interpolation between tabulated bounds provided by the sample histogram (right). Solid lines correspond to results obtained for median indicator kriging where the same semivariogram model is used across all thresholds. Graphs (A) and (B) were rescaled by the average concentration (values for Cd were halved for graph clarity) so that they are comparable across metals and interpolation models.

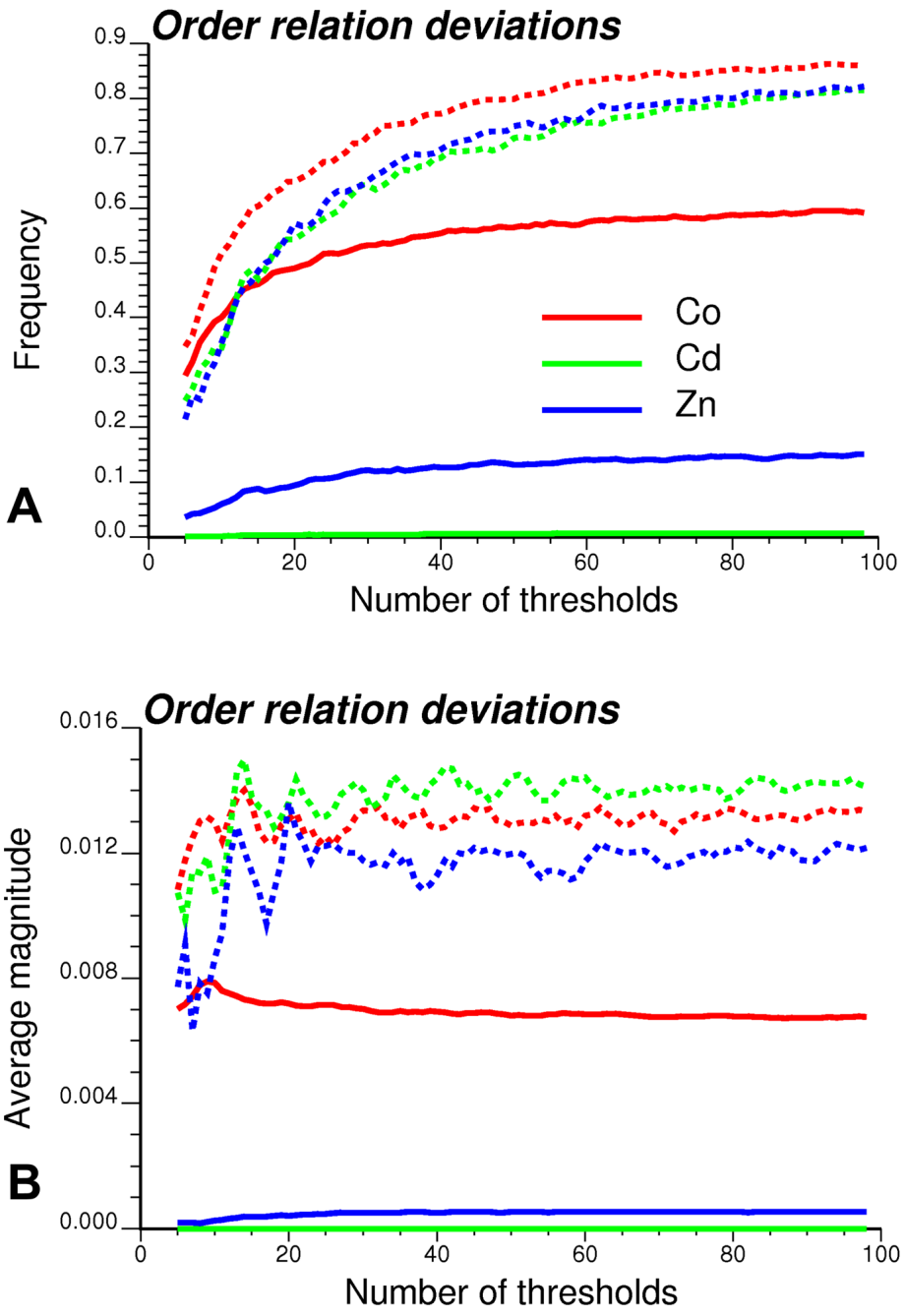

Figure 14.
Impact of the number of thresholds on the frequency (A) and magnitude (B) of order relation deviations for three different heavy metals. Solid lines correspond to results obtained for median indicator kriging where the same semivariogram model is used across all thresholds.